

Trustworthy Machine Learning: Explainability and Distribution-Free Uncertainty Quantification

Apprentissage automatique fiable: explicabilité et quantification d'incertitude sans
hypothèse de distribution

Thèse de doctorat de l'Université Paris-Saclay

Ecole doctorale n° 574 : mathématiques Hadamard (EDMH)

Spécialité de doctorat : Mathématiques appliquées

Graduate School : Mathématiques

Référent : Université d'Evry-Val d'Essonne

Thèse préparée dans l'unité de recherche Laboratoire de Mathématiques et
Modélisation d'Evry (Université Paris-Saclay, CNRS, Univ Evry) sous la direction
de Nicolas BRUNEL, Professeur


Thèse soutenue à Paris-Saclay, le 15 décembre 2023, par

Salim IBRAHIM AMOUKOU

Composition du jury

Membres du jury avec voix délibérative

Véronique MAUME-DESCHAMPS Professeure, Université Claude Bernard Lyon 1	Présidente & Examinatrice
Pierre GEURTS Professeur, Université de Liège	Rapporteur
Jean-Michel LOUBES Professeur, Université de Toulouse III	Rapporteur
Nicolas BOUSQUET Maître de conférences, Sorbonne Université	Examineur
Juhyun PARK Maître de conférences, ENSIIE	Examinatrice
Erwan SCORNET Professeur, Sorbonne Université	Examineur



Laboratoire de
Mathématiques
et Modélisation
LaMME d'Évry

Titre: Trustworthy Machine Learning: Explainability and Distribution-Free Uncertainty Quantification

Mots clés: Explicabilité, Interprétabilité, Quantification d'Incertitude, Prédiction Conforme, Forêt Aléatoire, Ensemble d'Arbres de Décision

Résumé: Le principal objectif de cette thèse est d'accroître la confiance dans les modèles de Machine Learning en développant des outils capables d'expliquer leurs prédictions et de quantifier l'incertitude qui y est associée. La première partie de cette thèse se concentre sur les méthodes d'explication locales. Nous mettons d'abord en évidence les limites des estimateurs existants des indices de Shapley pour les modèles basés sur les arbres de décision, ainsi que les problèmes liés à leur utilisation en présence de variables catégorielles. Après avoir proposé des solutions à ces problèmes, nous démontrons que les indices de Shapley et la méthode LIME ne sont pas fiables pour fournir des explications locales. Nous introduisons ensuite de nouvelles méthodes d'explication, sous forme de mesures d'importance, de sélection de sous-ensembles de variables importantes, de règles de décision locales, d'action contrefactuelles et de contrefactuels basés sur des règles de décision. Toutes les méthodes que nous proposons sont

"model-free", c'est-à-dire qu'elles n'ont pas besoin d'avoir accès au modèle pour effectuer des prédictions. De plus, elles n'impliquent pas la génération de nouvelles observations, évitant ainsi les problèmes d'extrapolation inhérents aux méthodes existantes qui se basent sur des prédictions utilisant des observations improbables ou impossibles, générées en combinant de manière aléatoire les attributs des variables provenant de multiples observations. En outre, les méthodes proposées se distinguent des différentes heuristiques que l'on trouve dans la littérature, car les quantités qui les définissent sont clairement définies et sont accompagnées de résultats de consistance. Dans la deuxième partie, nous analysons la prédiction conforme, qui permet de construire des intervalles prédictifs avec une garantie de couverture non asymptotique, en se basant uniquement sur l'hypothèse d'échangeabilité des observations. Nous proposons une méthode pour rendre ces intervalles plus adaptatifs, tout en garantissant le taux de couverture conditionnellement à un jeu de calibration donné.

Title: Trustworthy Machine Learning: Explainability and Distribution-Free Uncertainty Quantification

Keywords: Explainability, Interpretability, Uncertainty Quantification, Conformal Prediction, Random Forest, Tree-based models

Abstract: The main objective of this thesis is to increase trust in Machine Learning models by developing tools capable of explaining their predictions and quantifying the associated uncertainty. The first part of this thesis focuses on local explanation methods. We first highlight the limitations of existing estimators of Shapley Values for tree-based models and the issues related to their use with categorical variables. After proposing solutions to these problems, we demonstrate the unreliability of Shapley Values and the LIME method in providing local explanations. Subsequently, we introduce novel explanation techniques, including importance measures, selection of important variable subsets, local decision rules, counterfactual actions, and rule-based counterfactuals. All the proposed methods are "model-free," meaning they do not require access to the underlying model to make predictions. Furthermore, they do not involve generating new observations, thus avoiding the extrapolation problems inherent in most existing methods that rely on predictions using implausible or impossible observations created by randomly combining variable values from multiple instances. Moreover, the proposed methods stand out from the diverse heuristics found in the literature by offering precise definitions of the involved quantities accompanied by consistency results. In the second part of the thesis, we analyze conformal prediction, which allows for constructing predictive intervals with non-asymptotic coverage guarantees based solely on the assumption of exchangeability. We propose a method to make these intervals more adaptive and ensure coverage rates given a single calibration set.

Université Paris-Saclay

Espace Technologique / Immeuble Discovery

Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

Contents

Contents	i
Notations	iv
Introduction (français)	1
1 Contexte	1
2 IA de confiance	3
3 Contributions	8
1 State Of The Art	14
1 Explanaible AI	15
2 Distribution-Free Predictive Inference	27
2 Accurate Shapley Values for explaining tree-based models	36
1 Introduction	37
2 Coalition and Invariance for Shapley Values	38
3 Shapley Values for tree-based models	42
4 Comparison of the estimators	48
5 Discussion and Future works	50
3 Please stop using SHAP and LIME and use Regional Explanations instead	51
1 Introduction	52
2 Stop using Local Shapley Values	53
3 Stop using LIME	56
4 From Global Explanations to Regional Explanations	58
5 Experiments	61
6 Discussion	65
4 Beyond Features attributions: Sufficient Explanations and Rules	67
1 Introduction	68
2 Motivations and Related works	69
3 Probabilistic Sufficient Explanations for Regression	70
4 SDP, Sufficient Explanations and Sufficient Rules via Random Forest	72
5 Experiments	79
6 Conclusion	83
5 Rethinking Counterfactual Explanations as Local and Regional Counterfactual Policies	84

1	Introduction	85
2	Motivation and Related works	86
3	Minimal Counterfactual Rules	88
4	Estimation of the <i>CDP</i> and <i>CRP</i>	91
5	Learning the Counterfactual Rules	93
6	Sampling CE using the CR	95
7	Experiments	96
8	Conclusion	98
6	Adaptive Conformal Prediction by Reweighting Nonconformity Score	99
1	Motivations	100
2	Related works and contributions	102
3	Random Forest Localizer	104
4	Weighted Conformal Prediction	106
5	Asymptotic conditional coverage	111
6	Experiments	112
7	Conclusion	113
7	Future works	114
1	Prediction with reject option using conformal p-value	115
2	Conformal Protection Layers for Counterfactual Explanations	116
	Conclusion	117
	Bibliographie	119
	Appendix for Chapter 2	134
1	Proof of SV invariance for transformed continuous variables	134
2	Proof of SV invariance for encoded categorical variables	134
3	Proof of the limitation of SV as local explanation	136
4	Relation between the Algorithm 1 (TreeSHAP with path-dependent) and \hat{f}^{SHAP}	138
5	Additional experiments	140
6	Experiments setting	141
	Appendix for Chapter 3	142
7	Proof of Theorem 2.2	142
	Appendix for Chapter 4	145
8	Proof of the Projected CDF Forest consistency	145
9	Empirical evaluations of the estimator \hat{F}_S	154
10	Additional experiments	155
11	From Sufficient Rules to Global Interpretable model	160
12	Projected Forest CDF algorithm	161
	Appendix for Chapter 5	162
13	Regional RF detailed	162
14	Additional experiments	163
15	Parameters detailed	164
	Appendix for Chapter 6	165
16	Proof of Lemma 4.1	165

17	Proof of Theorem 4.4: Training-conditional of LCP-RF	166
18	Proof of marginal coverage of groupwise LCP-RF	167
19	Proof of Theorem 5.4: asymptotic conditional coverage	167
20	Additional experiments	175

Notations

Tout au long de cette thèse, les variables aléatoires sont définies sur l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$.

Si X est une variable aléatoire réelle positive ou intégrable¹, nous notons

$$\mathbb{E}(X) := \int_{\Omega} X(\omega) \mathbb{P}(d\omega) \in \mathbb{R},$$

son espérance mathématique.

Lorsque X est une variable aléatoire prenant ses valeurs dans un espace mesurable quelconque (E, \mathcal{E}) , nous utilisons la notation P_X pour désigner la loi de cette variable aléatoire. Cette loi est définie par $P_X(B) = \mathbb{P}(X^{-1}(B))$ pour tout $B \in \mathcal{E}$, il s'agit la mesure image de \mathbb{P} par l'application mesurable $X : (\Omega, \mathcal{F}) \rightarrow (E, \mathcal{E})$. P_X est l'unique mesure de probabilité sur (E, \mathcal{E}) qui vérifie

$$\mathbb{E}(X) := \int_{\Omega} X(\omega) \mathbb{P}(d\omega) = \int_{\mathbb{R}} x P_X(dx).$$

Ainsi, nous employons fréquemment la notation \mathbb{E}_{P_X} pour spécifier que l'espérance est calculée selon la loi de la variable aléatoire X , notamment lorsqu'il existe des risques d'ambiguïté. Plus particulièrement, si nous disposons de n variables aléatoires indépendantes (X_1, \dots, X_n) à valeurs dans (E, \mathcal{E}) et partageant la même loi P_X , nous adoptons la notation $\mathbb{E}_{P_X^n}$ pour indiquer que l'espérance est prise par rapport à la loi jointe de ces n variables aléatoires.

Nous utilisons des lettres majuscules pour représenter les variables aléatoires, des minuscules pour leurs réalisations, et des caractères gras pour désigner des vecteurs. Dans les sections à venir, nous supposons que nous avons à notre disposition un ensemble de p variables aléatoires, notées $\mathbf{X} = (X_1, \dots, X_p)$, qui représentent les entrées ou covariables. La réalisation de ces variables aléatoires est représentée par $\mathbf{x} = (x_1, \dots, x_p)$. Nous utilisons la notation $[p]$ pour désigner l'ensemble $\{1, \dots, p\}$.

¹C'est-à-dire une variable aléatoire X telle que $\mathbb{E}(|X|) < \infty$.

Introduction (français)

1 Contexte

Les modèles de Machine Learning (ML) sont omniprésents dans notre quotidien, que ce soit à travers nos smartphones où ils animent l'assistant vocal et la reconnaissance faciale, dans le secteur bancaire avec la prédiction d'octroi de crédit, ou encore dans les recommandations personnalisées de musique et de films que nous recevons. Leur fonctionnement est généralement le même: l'algorithme reçoit des données et apprend à prédire une valeur cible en minimisant une erreur. La quête incessante de modèles prédictifs de plus en plus performants a conduit à la création de systèmes de plus en plus complexes, et, par conséquent, moins transparents. Cette opacité a rendu difficile la compréhension du processus de prédiction de ces modèles et leur contrôle. Cette situation a donné lieu à de nombreux scandales, comme celui survenu chez Amazon [Dastin, 2018], où un biais de genre a été découvert dans leur système de recrutement basé sur le Machine Learning. Les modèles avaient été entraînés sur des données historiques qui présentaient un déséquilibre en faveur des hommes pour les postes techniques. En conséquence, les modèles ont favorisés les candidats masculins aux dépens des candidates féminines. Cette course effrénée vers la performance prédictive, au détriment de la transparence des processus décisionnels, pose un problème majeur. Imaginons un modèle prédictif dans le secteur médical qui diagnostique un patient comme susceptible de développer une maladie, sans toutefois expliquer pourquoi. Dans de telles circonstances, il serait difficile d'accorder notre confiance à ce modèle. Cette opacité ne permet pas de valider cliniquement ces prédictions ni d'identifier les éventuelles erreurs, compromettant ainsi la sécurité des patients. Ainsi, la transparence et l'explicabilité des modèles de Machine Learning deviennent essentielles dans tous les domaines où répondre au "*pourquoi cette prédiction ?*" est tout aussi important que connaître la prédiction elle-même. De plus, l'un des objectifs fondamentaux de la science, au-delà de prédire les effets d'une cause, est de comprendre les causes de l'effet. Il devient donc crucial que nous puissions expliquer les prédictions de ces modèles pour mieux comprendre les données qu'ils utilisent et faciliter leur déploiement dans tous les secteurs. Par ailleurs, pour renforcer la confiance entre les utilisateurs et le Machine Learning, il est aussi nécessaire de mesurer l'incertitude associée à chaque prédiction. Reprenant l'exemple médical précédent, avant même de vouloir expliquer la prédiction d'une maladie par le modèle, il est primordial de s'assurer que cette prédiction est

fiable. Cette thèse s'engage donc à renforcer la confiance envers les modèles de Machine Learning. Pour mériter cette confiance, un modèle de Machine Learning doit être capable d'exprimer ce qu'il sait et ce qu'il ne sait pas. Ainsi, cette thèse se concentrera sur deux sujets cruciaux: l'explication des modèles de Machine Learning (ce qu'ils savent) et l'estimation de l'incertitude associée aux prédictions (ce qu'ils ignorent).

Contexte de la thèse. Cette thèse a été effectuée en collaboration avec [Stellantis](#) (anciennement PSA Groupe) et le Laboratoire de Mathématiques et Modélisation d'Evry (LaMME), avec le soutien de la Convention Industrielle de Formation par la Recherche (CIFRE) de l'Association Nationale de la Recherche et de la Technologie (ANRT). Nous avons également travaillé avec les équipes de [Quantmetry](#), une société de conseil en Intelligence Artificielle (IA), qui nous a permis d'obtenir des retours terrains sur l'interaction entre les divers acteurs de IA et les méthodes d'explicabilité. Plus spécifiquement, cela nous a permis d'identifier la nécessité d'adapter les solutions d'explicabilité en fonction des différents acteurs impliqués, tels que les auditeurs, les clients, les experts métiers et les data scientists, qui ont des besoins distincts. Ce projet s'inscrit dans un appel récent de la société civile et de la communauté scientifique visant à réguler, encadrer et maîtriser les modèles d'apprentissage automatique. En France, cet intérêt se manifeste notamment par le [rapport Villani](#) qui aborde les questions éthiques posées par l'usage des modèles de machine learning. En 2021, nous assistons à l'émergence du consortium "[Confiance AI](#)" lancé par l'Etat dans le cadre du Grand Défi « Sécuriser, certifier et fiabiliser les systèmes fondés sur l'intelligence artificielle », auprès d'une quarantaine de partenaires industriels et académiques, pour concevoir et industrialiser des systèmes à base d'IA de confiance. À l'échelle internationale, ce mouvement se poursuit également avec des initiatives telles que la proposition de réglementation de la Commission Européenne appelée "[AI ACT](#)". De même, la Maison Blanche a aussi proposé une réglementation similaire intitulée "[Blueprint for an AI bill of rights](#)". Notre travail a été finalement de traduire ces questions et besoins, et de proposer des outils mathématiques pour y répondre.

Application industrielle chez Stellantis. Le cœur de l'industrie automobile réside dans la fabrication de véhicules. Une chaîne de production de véhicules est composée d'une série d'opérations complexes de transformation et d'assemblage de pièces, qui sont déterminées par un grand nombre de variables telles que la température, la durée de chaque étape d'assemblage, le type de pièce, le type de voiture, le type d'opération, etc. Afin de réduire les retours de véhicules après vente, un processus de contrôle qualité a été mis en place, consistant à tester de manière aléatoire certains véhicules à la sortie de l'usine. Cependant, ce processus, étant coûteux et peu optimal, peut être remplacé par un modèle de machine learning qui apprend le lien entre les variables d'entrée du processus et la présence ou absence de défauts sur le véhicule. Il est donc essentiel de mesurer avec précision l'incertitude associée aux prédictions de ces modèles, étant donné que le nombre de tests réalisables est limité en raison de contraintes budgétaires. De plus, nous pouvons utiliser l'explication des prédictions de ces modèles afin de mieux comprendre les conditions qui mènent aux défauts à la sortie de l'usine, et par conséquent, ajuster le processus de production pour les éviter. Une autre application de l'explicabilité chez

Stellantis concerne l'amélioration de l'expérience client. Une plateforme permet aux utilisateurs de laisser des commentaires pour partager leur avis. Cette plateforme est naturellement soumise à des règles de modération, telles que l'interdiction des injures ou des données personnelles. Cette modération a été automatisée grâce à des modèles de Machine Learning, dans le but d'optimiser les performances et de réduire les coûts de modération. Ainsi, l'explication des prédictions de ces modèles peut être intéressante pour accompagner les utilisateurs en leur fournissant une explication sur le refus de leur commentaire. De plus, cela permet de détecter les motifs ou les subtilités du langage dans lesquels le modèle se trompe fréquemment. Cette compréhension fine des erreurs de modération peut ensuite être utilisée pour améliorer le modèle en production.

2 IA de confiance

2.1 Motivations

Dans cette section, nous mettons en évidence quatre raisons qui soulignent la nécessité d'ajouter des mécanismes de protection aux modèles de Machine Learning, en développant des outils pour expliquer les prédictions et évaluer l'incertitude qui leur est associée. Tout au long de la thèse, nous considérons les prédictions du modèle comme la sortie $Y \in \mathcal{Y}$ étant donné les variables $\mathbf{X} \in \mathcal{X}$, issues d'un processus $(\mathbf{X}, Y) \sim P_{\mathbf{X}}P_{Y|\mathbf{X}}$, où $P_{Y|\mathbf{X}}$ représente un processus aléatoire ou un modèle déterministe. Ainsi, nous nous trouvons dans un contexte purement agnostique, qui consiste à expliquer les données ou les sorties d'un modèle fixe.

Gérer les risques liés aux prédictions. L'objectif principal d'un modèle de Machine Learning est de réaliser des prédictions. Les modèles sont généralement construits en minimisant une fonction de coût. Dans des situations réelles où les hypothèses classiques sur les données (normalité, homoscedasticité, etc.) ne s'appliquent pas et où les données ne sont pas infinies, nous disposons souvent uniquement que d'une mesure empirique de la performance globale de notre modèle. Cependant, dans certains contextes où une seule erreur peut avoir des conséquences graves, il est nécessaire d'estimer de manière raisonnable l'erreur individuelle. Une approche consiste à construire un intervalle de prédiction qui fournit un ensemble de valeurs ou d'intervalles susceptibles de contenir la valeur cible de chaque observation avec une probabilité contrôlée. Ainsi, en analysant la taille de l'intervalle prédictif, nous pouvons déduire l'incertitude associée aux prédictions afin d'utiliser sereinement nos modèles de Machine Learning pour prendre des décisions ou automatiser des tâches dans les domaines à haut risque.

Justifier les prédictions. Les dérives des modèles de Machine Learning ont entraîné la mise en place de nombreuses réglementations à travers le monde. Par exemple, le Règlement Général sur la Protection des Données (RGPD) comprend l'article 22 [Goodman, 2017], qui exige que toute décision automatisée soit en mesure de justifier sa prédiction. En d'autres termes, cela donne le droit à une explication à toute personne concernée par les modèles de Machine Learning. Nous pouvons également trouver des réglementations similaires dans le secteur de l'assurance ou bancaire [EBA, 2020], notamment en ce qui concerne la transparence des modèles utilisés pour tarifier les produits d'assurance ou le droit à une explication après le refus d'un prêt.

Ces réglementations soulignent l'importance de pouvoir comprendre les raisons sous-jacentes aux prédictions des modèles de Machine Learning. Il ne suffit plus de se fier uniquement aux résultats obtenus, mais il est désormais nécessaire de pouvoir expliquer pourquoi ces résultats ont été obtenus. Cela permet aux individus concernés de comprendre les critères utilisés, d'éviter des discriminations, de remettre en question les décisions prises et de vérifier si celles-ci sont conformes aux principes éthiques et légaux.

Détecter les biais. Les explications des prédictions peuvent être utilisées pour détecter les biais présents dans les modèles [Corbett-Davies, 2018; Pessach, 2022]. Vous pourriez vous demander comment cela est possible. Un algorithme peut-il être biaisé ? Ne serait-ce pas une forme de personnalisation ? Cette question est d'autant plus pertinente à l'ère de ChatGPT [OpenAI, 2023], où beaucoup considèrent l'IA comme une forme d'intelligence mystérieuse, parfois associée à une AGI (Intelligence Générale Artificielle) qui pourrait surpasser l'humain, dans la veine de ce qu'on voit dans Terminator ou Blade Runner. Cependant, ce n'est pas le cas. L'IA n'est ni consciente ni intelligente en tant qu'entité autonome. Ce sont simplement des modèles qui se sont spécialisés dans des tâches bien spécifiques en extrayant des schémas à partir de nos données. Ainsi, si les données elles-mêmes présentent des biais, tels que des biais de genre ou de race, l'algorithme peut les reproduire. Prenons l'exemple de l'algorithme COMPAS [Washington, 2018], utilisé par certains systèmes judiciaires pour évaluer le risque de récidive des individus. Des études [Washington, 2018] ont montré que ce modèle présente des biais raciaux, avec des taux de fausses prédictions plus élevés pour certaines communautés. L'analyse des explications des prédictions peut fournir des indices sur la présence de biais, en identifiant les facteurs discriminatoires pris en compte par le modèle. De telles détectons de biais peuvent aider à corriger et à améliorer les modèles, afin de garantir une prise de décision équitable et impartiale.

Détecter les facteurs de confusion. Il est crucial de pouvoir détecter les facteurs de confusion dans les modèles de Machine Learning, car un modèle peut obtenir de bonnes performances prédictives pour de mauvaises raisons. Cela signifie qu'il peut être efficace dans un contexte spécifique, mais incapable de généraliser ou de s'adapter à de nouvelles situations. De nombreux exemples mettent en évidence comment les modèles peuvent prendre des raccourcis afin de minimiser leur fonction de coût, sans réellement comprendre les véritables caractéristiques ou motifs liés aux prédictions. À titre d'exemple, [Beery, 2018] a exposé ce problème avec des modèles entraînés pour reconnaître des animaux : les vaches dans des environnements communs tels que les pâturages étaient correctement classifiées, tandis que celles dans des contextes atypiques comme la plage étaient mal identifiées. Ceci démontre que le modèle privilégie l'arrière-plan des images pour formuler ses prédictions. Ainsi, le modèle peut sembler performant, mais pas pour les bonnes raisons. Il n'a pas réellement appris à détecter les vaches, mais exploite plutôt l'arrière-plan de l'image pour minimiser sa fonction de coût. Cette situation a été récemment observée par [DeGrave, 2021] lors de la détection du COVID-19 à partir d'images de radiographies pulmonaires. Certains modèles de Machine Learning se sont basés uniquement sur des marqueurs présents sur les images qui étaient parfaitement corrélés avec la présence de la maladie, sans utiliser aucune information de la région pulmonaire elle-même. Ainsi, en analysant

les explications des prédictions, nous pouvons repérer les signaux indiquant que le modèle se base sur des facteurs de confusion plutôt que sur des caractéristiques significatives. Cela permet d'ajuster et de corriger les modèles afin qu'ils se concentrent sur les aspects pertinents pour les prédictions, favorisant ainsi une meilleure généralisation et une adaptation à différents contextes.

2.2 Explicabilité

Commençons par définir le concept d'explicabilité. Il n'existe pas de consensus absolu sur la définition de ce terme, et les chercheurs utilisent souvent leurs propres intuitions pour définir ce qui constitue une explication [Bellucci, 2021; Adadi, 2018; Doshi-Velez, 2017]. Par conséquent, il existe dans la littérature de nombreuses taxonomies souvent contradictoires et une utilisation interchangeable des termes tels que "interprétabilité" et "explicabilité". En réalité, la notion d'explication n'est pas nouvelle et a alimenté de nombreuses discussions, aussi bien en philosophie qu'en sciences cognitives. Les philosophes ont depuis longtemps cherché à comprendre ce qui constitue une explication, se demandant si toutes les explications sont de nature causale et quelle est leur structure sous-jacente [Salmon, 2006]. Au cours des deux dernières décennies, les sciences cognitives ont étudié la manière dont nous générons des explications, évaluons leur pertinence et pourquoi nous demandons des explications [Malle, 2006]. Sans nous égarer dans cette vaste littérature composée d'opinions diverses, dans cette thèse, nous nous appuyons sur les travaux de [Miller, 2019] qui a analysé plus de 300 articles en philosophie, en sciences cognitives et en sciences sociales afin de trouver une définition adéquate pour le Machine Learning. Il en ressort que l'explication est le résultat d'un processus cognitif, et la plupart des recherches s'accordent pour dire qu'une explication est une réponse à une question du type "pourquoi". Dans notre contexte, l'explication des modèles de Machine Learning consisterait à répondre à la question suivante: *Pourquoi le modèle a-t-il fait telle prédiction ?* D'autre part, l'interprétabilité fait référence à la capacité de comprendre, d'expliquer et de rendre compte des décisions ou des prédictions d'un modèle ou d'un système. Elle est directement liée à l'humain, car c'est lui qui interprète les résultats. Nous distinguons donc le Machine Learning Interprétable, qui regroupe les méthodes intrinsèquement compréhensibles pour l'humain sans nécessiter d'explications supplémentaires, telles que les modèles linéaires ou les arbres de décision, et l'IA explicable qui regroupe les méthodes qui fournissent des explications, permettant de répondre à la question *"Pourquoi le modèle a-t-il fait telle prédiction ?"*. C'est ce dernier aspect que nous étudierons dans cette thèse, en nous efforçant de répondre aux questions associées:

1. Quels sont les variables qui ont joué un rôle déterminant dans le processus de prédiction ?
2. Sous quelles conditions le modèle privilégie une certaine prédiction ?
3. Pourquoi le modèle a-t-il privilégié telle prédiction par rapport à une autre ?

Au cours de cette thèse, nous nous attacherons principalement à répondre à ces trois questions. Plusieurs formes d'explications peuvent être utilisées pour y parvenir. Par exemple, les mesures d'importance [Wei, 2015] ou l'analyse de sensibilité [Razavi, 2021; Da Veiga, 2021; Saltelli, 2008] peuvent être employées pour la question 1, fournissant une valeur représentant la contribution

d'une variable donnée à une prédiction spécifique. Il y a aussi la sélection de sous-ensemble de variables importantes, qui sont capables de maintenir la prédiction même en modifiant les autres variables. Pour la question 2, les règles de décision permettent d'identifier les zones de l'espace des variables où la prédiction du modèle est identique, ce qui permet d'expliquer les prédictions en fonction de leur zone d'appartenance. La question 3 met en lumière une approche mimant la façon dont les êtres humains fournissent habituellement des explications. Lorsqu'on nous demande une explication, nous avons tendance à adopter une perspective contrefactuelle - plutôt que de chercher à comprendre pourquoi l'événement P s'est produit, on cherche à comprendre pourquoi l'événement P s'est produit au lieu d'un événement alternatif Q [Miller, 2019]. Par conséquent, les modèles peuvent être expliqués en proposant des actions contrefactuelles, c'est-à-dire en identifiant les changements minimaux des variables qui permettraient de modifier la décision du modèle. Par exemple, si un modèle de scoring de crédit refuse un prêt et qu'il suffit de modifier le salaire du demandeur pour changer la décision, nous pouvons en déduire que le salaire est l'une des raisons possibles pour laquelle le crédit n'a pas été accordé. Dans cette thèse, nous allons étudier toutes ces formes d'explications à savoir les mesures d'importance locales, la sélection de variables importantes, les règles de décisions et enfin les contrefactuelles.

La tâche d'expliquer les modèles de Machine Learning présente plusieurs difficultés majeures. La plupart des méthodes d'explicabilité reposent essentiellement sur le principe de la modification d'un sous-ensemble de variables puis à l'observation du comportement du modèle [Covert, 2021]. Cependant, cela soulève des défis importants liés à la sélection des variables à modifier et au choix des nouvelles valeurs pour ces variables. Une approche couramment utilisée dans la littérature consiste à utiliser la loi marginale de chaque variable pour générer des observations modifiées. Par exemple, le MDA (*Mean Decrease Accuracy*) [Breiman, 2001] calcule la variation de l'erreur du modèle suite à la permutation aléatoire des valeurs d'une variable dans le jeu de données. Cependant, cette approche peut conduire à des observations impossibles dans la réalité. Par exemple, dans le cas de la prédiction du prix d'une maison de telles approches peuvent amener à appliquer le modèle à des maisons de 16 m² ayant 10 pièces. L'un des objectifs de notre travail est de nous assurer de ne pas nous retrouver dans de telles situations comme le font la majeure partie des méthodes d'explicabilité, car nous n'avons aucune garantie sur le comportement du modèle dans ces scénarios impossibles. De telles extrapolations sont au minimum pas fiables, et potentiellement dénuées de sens. Une autre source de complexité réside dans les relations entre les différentes variables, qui peuvent être non linéaires ou interdépendantes, rendant ainsi difficile l'identification des influences spécifiques de chaque variable sur les prédictions du modèle.

2.3 Quantification d'Incertitudes

Une autre façon de renforcer la confiance dans les modèles d'apprentissage automatique est de pouvoir mesurer l'incertitude associée à leurs prédictions. Lorsque l'on parle d'incertitude, on pense généralement aux intervalles de confiance. Cependant, dans cette thèse, nous nous intéressons plutôt aux intervalles prédictifs, qui diffèrent des intervalles de confiance. Les intervalles de confiance sont principalement utilisés pour trouver un intervalle qui contient un paramètre du

modèle génératif des données avec une certaine probabilité, généralement $1 - \alpha$, ou α est le risque de se tromper. Ces intervalles permettent d'estimer la précision ou la fiabilité de l'estimateur du paramètre associé. En revanche, les intervalles prédictifs sont utilisés pour trouver un intervalle ou un ensemble de valeurs qui contiennent la valeur cible en utilisant un modèle de prédiction. Plus formellement, ayant un jeu de données $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, une observation de test $(\mathbf{X}_{n+1}, Y_{n+1})$, où $(\mathbf{X}_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ sont générées selon $(\mathbf{X}, Y) \sim P_{\mathbf{X}}P_{Y|\mathbf{X}}$ et un modèle de prédiction $\hat{f} : \mathcal{X} \mapsto \mathcal{Y}$ estimé pour prédire Y à partir de \mathbf{X} . Nous souhaitons construire un ensemble $\hat{C}(\cdot)$ sachant \mathbf{X}_{n+1} qui contiendrait Y_{n+1} avec probabilité $1 - \alpha$,

$$\mathbb{P}\{Y_{n+1} \in \hat{C}(\mathbf{X}_{n+1})\} \geq 1 - \alpha.$$

$\hat{C}(\cdot)$ fournit une estimation de la plage de valeurs dans laquelle les sorties des futures observations sont susceptibles de se situer compte tenu de l'incertitude inhérente au modèle \hat{f} et aux données, appelées respectivement incertitude épistémique et incertitude aléatoire. Les intervalles prédictifs permettent d'appréhender l'incertitude associée au modèle \hat{f} , en transformant les prédictions ponctuelles de chaque observation en intervalle ou ensemble de valeur plausible, ce qui peut être particulièrement utile pour la prise de décision.

Il existe plusieurs méthodes pour construire des intervalles prédictifs, telles que les techniques de rééchantillonnage [Yu, 2002], les approches bayésiennes [Dawid, 1982; Fraser, 2011], et la régression quantile [Koenker, 2001]. Cependant, l'application de ces méthodes dans un contexte industriel présente des difficultés en raison de la nécessité de minimiser les hypothèses, compte tenu de la complexité des modèles utilisés. Notamment, il n'est pas réaliste de supposer que les résidus suivent une distribution gaussienne, comme le font certaines méthodes. De plus, le besoin de calculer l'incertitude se présente fréquemment après avoir déjà choisi et estimé le modèle. Nous avons aussi besoin de garanties non asymptotiques afin de prendre des décisions en pratique. Les techniques de rééchantillonnage sont très coûteuses en temps de calcul et ne fournissent pas de garanties de couverture. Les méthodes bayésiennes et la plupart des autres méthodes ne donnent pas non plus de garanties de couverture non asymptotique. Bien que les méthodes bayésiennes fournissent des garanties asymptotiques, nous ne savons pas ce qui se passe dans le cas fini. Cette problématique est d'autant plus complexe avec le fléau de la dimension. Idéalement, nous cherchons une méthode capable de construire des intervalles prédictifs qui contiendraient notre variable cible avec une certaine probabilité contrôlée avec des données finies, tout en ayant des hypothèses faibles, telles que l'échangeabilité des données et aucune hypothèse sur le modèle. Un cadre qui permet de répondre à ces limitations ou attentes est celui de la prédiction conforme [Vovk, 2005; Lei, 2016]. Dans la partie consacrée à l'estimation de l'incertitude, nous allons essentiellement nous focaliser sur ce dernier. En utilisant la prédiction conforme, nous construisons des intervalles prédictifs qui offrent des garanties de couverture non asymptotique pour les prédictions générées par n'importe quel modèle de machine learning. Cela permettra d'obtenir une mesure fiable de l'incertitude associée aux prédictions des modèles, offrant ainsi une base solide pour la prise de décision et l'évaluation des performances des modèles, adaptée aux exigences du contexte industriel.

3 Contributions

Cette thèse est divisée en six parties. Le chapitre 1 se compose de deux sections: la première est consacrée à la description des techniques couramment utilisées pour l’explication locale des modèles, tandis que la seconde introduit la prédiction conforme. Le chapitre 2 propose une étude approfondie d’une des méthodes les plus populaires pour expliquer les modèles, à savoir les indices de Shapley [Lundberg, 2017a]. Nous identifions quelques problèmes liés à l’estimation des indices de Shapley et à leur utilisation en présence de variables catégorielles. Le chapitre 3 poursuit l’analyse précédente en soulignant que les indices de Shapley et la méthode LIME [Ribeiro, 2016a] ne sont pas fiables comme explication locale, et propose une approche pour construire une mesure d’importance locale de façon plus rigoureuse. Dans le chapitre 4, nous proposons d’aller au-delà des mesures d’attribution et introduisons une méthode d’explication capable de capter les interactions. Cette approche repose sur la sélection de sous-ensembles minimaux de variables importantes, suffisants pour maintenir la prédiction lorsqu’on modifie les autres variables en respectant la distribution des données. En utilisant les variables sélectionnées, nous avons également proposé une méthode d’explication sous forme de règles de décision locales. Le chapitre 5, le dernier chapitre consacré à l’explicabilité, est en quelque sorte le dual du chapitre 4. Nous utilisons essentiellement la même approche, mais cette fois pour générer des exemples contrefactuels. C’est-à-dire que nous cherchons le sous-ensemble minimal de variables qui permet de changer la décision, puis les règles de décision locales permettant de modifier la décision. Le chapitre 6 porte sur l’estimation des incertitudes. Nous proposons une stratégie de pondération visant à améliorer la fidélité des intervalles de prédiction fournis par la prédiction conforme, de manière à les rendre plus adaptatifs tout en contrôlant le taux de couverture conditionnellement au jeu de calibration. Le dernier chapitre est essentiellement consacré aux travaux futurs qui sont des prolongements des travaux de la thèse. Il s’agit notamment de l’utilisation de la prédiction conforme pour effectuer des prédictions avec abstention, c’est-à-dire utiliser la prédiction conforme pour établir une stratégie permettant de s’abstenir de prédire lorsque l’incertitude est trop élevée, tout en contrôlant notre taux de faux positifs. Enfin, nous décrivons un croisement entre explicabilité et la prédiction conforme, utilisant la prédiction conforme pour obtenir des garanties non asymptotiques avec des hypothèses minimales sur les explications retournées. Ces travaux ont données lieu à quatre publications et deux packages:

- Chapitre 2: Accurate Shapley Values for explaining tree-based models [Amoukou, 2022b], publié à AISTATS 2022.
- Chapitre 4: Consistent Sufficient Explanations and Minimal Local Rules for explaining regression and classification models [Amoukou, 2021a], publié à NeurIPS 2022.
- Chapitre 5: Rethinking Counterfactual Explanations as Local and Regional Policies [Amoukou, 2022a], soumis à NeurIPS 2023. Une première version du papier a été acceptée à un workshop ([Counterfactuals in Minds and Machines](#)) à ICML 2023.
- Chapitre 6: Adaptive Conformal Prediction By Reweighting Nonconformity Scores [Amoukou, 2023], soumis à NeurIPS 2023.

Active Coalition of Variables. A Python package that provides explanations for any machine learning model or data. It gives local rule-based explanations for any model or data (regression and classification), different Shapley Values for tree-based models, and a new line of counterfactual explanations.

Adaptive Conformal Prediction Intervals. A Python package that provides Adaptive Prediction Intervals that effectively capture the uncertainty of any given model, with finite-sample marginal and PAC coverage, as well as asymptotic conditional coverage. It has been proven to significantly outperform the split-conformal approach, regardless of the nonconformity score used (e.g., mean score, quantile score).

3.1 Chapitre 2: Accurate Shapley Values for explaining tree-based models

L'une des principales motivations de ce chapitre découle du constat de la sur-représentation des indices/valeurs de Shapley dans les études d'explicabilité des modèles, en particulier pour l'explication locale. Dans la plupart des échanges avec des Data Scientists au sujet des techniques d'explicabilité qu'ils utilisent, les indices de Shapley sont mentionnés de manière quasi-systématique. Ce phénomène est illustré par le nombre impressionnant d'étoiles attribuées au package SHAP [Lundberg, 2017a], qui est de loin le package d'explicabilité le plus populaire avec plus de 19 000 étoiles. La figure 1 compare le nombre d'étoiles attribuées au package SHAP à celles attribuées à d'autres méthodes populaires telles que LIME [Ribeiro, 2016b], imodels [Singh, 2021], et interpretML [Nori, 2019]. Pour une liste exhaustive des différentes méthodes d'explicabilité disponibles, vous pouvez consulter le package [awesome-explainable-ai](#).

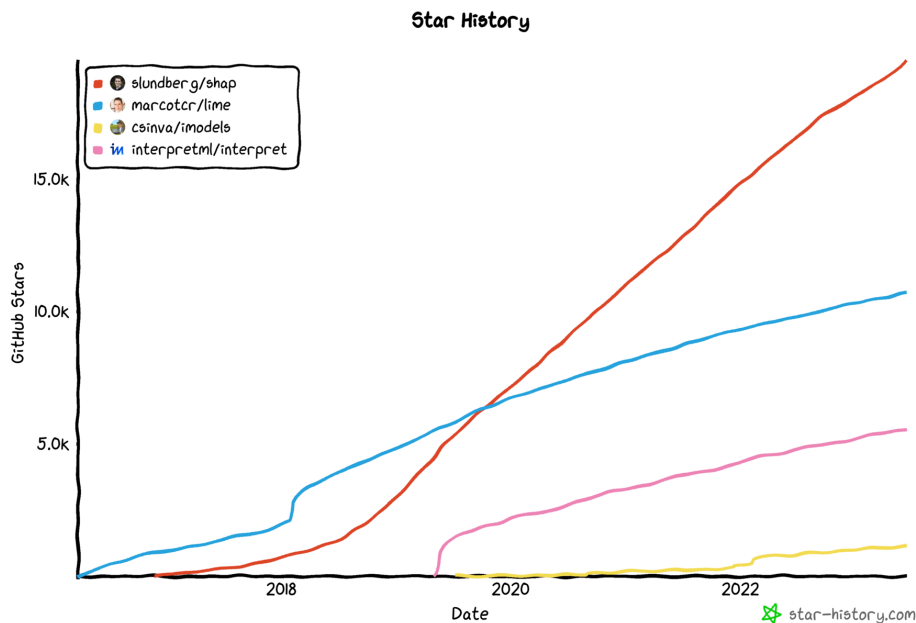


Figure 1: Nombre d'étoiles du package SHAP [Lundberg, 2017a] par rapport à d'autres méthodes populaires telles que LIME [Ribeiro, 2016b], imodels [Singh, 2021], interpretML [Nori, 2019].

Cette tendance a également été confirmée lors de notre collaboration avec le cabinet de conseil Quantmetry, qui a observé le même phénomène au cours de leurs missions. Cette prédominance des indices de Shapley a été aussi mise en évidence lors de notre participation au [tech-sprint/hackathon](#) organisé par l’Autorité de Contrôle Prudentiel et de Résolution [ACPR, 2022] et la Banque de France, un événement où nous avons [remporté la première place](#). L’objectif de cette compétition était d’expliquer des modèles de risque de crédit. Les participants étaient au nombre de 52, répartis en 12 équipes: Crédit Mutuel, BNP paribas, Crédit Agricole, Tinubu square, MAIF, DataRobot, La Banque Postale, Zelros, NukkAI, Stellantis-Quantmetry pour n’en citer que quelques uns. Le [rapport final](#) de la compétition montre que la grande majorité des équipes participantes ont utilisé les indices de Shapley dans leurs solutions. Cet engouement de l’industrie pour les indices de Shapley et leur utilisation comme une réponse systématique, a été l’un des principaux facteurs qui nous ont incités à approfondir notre analyse de ces indices afin de savoir si leurs utilisation était justifiée en terme de pouvoir explicatif.

Nous nous sommes concentrés sur l’analyse de l’algorithme TreeSHAP [Lundberg, 2020b] - dédié au calcul des indices de Shapley pour les méthodes à base d’arbre de décision telles que XGBoost [Chen, 2016] et Random Forest [Breiman, 2001]. Cet intérêt tient au fait que les modèles d’ensemble d’arbre sont les modèles de référence pour l’analyse des données tabulaires [Grinsztajn, 2022] et qu’ils sont massivement utilisés en pratique. Le papier original propose l’algorithme sans toutefois détailler la quantité exacte qui est calculée, ni analyser le comportement de l’algorithme d’un point de vue théorique. Nous avons donc explicité l’estimateur exact de cet algorithme, ce qui nous a permis de constater qu’il introduisait un fort biais lorsque les variables présentaient une dépendance. Pour remédier à cette limitation, nous avons proposé deux estimateurs moins biaisés, dont l’un permet de réduire considérablement la complexité exponentielle liée au calcul des indices de Shapley en fonction du nombre de variables. Par ailleurs, nous avons identifié un autre problème fréquemment rencontré dans la pratique, lié au calcul des indices de Shapley pour les variables catégorielles. Généralement, les variables catégorielles sont encodées, ce qui introduit de nouvelles variables. Il était courant de prendre la somme des indices de Shapley associées à ces nouvelles variables comme approximation de l’indice de Shapley de la variable catégorielle. Cependant, nous avons démontré que cette pratique était erronée et nous avons montré comment calculer correctement les indices de Shapley des variables catégorielles après encodage. Enfin, nous avons soulevé une problématique plus générale qui se cache derrière tous ces problèmes: la pertinence des indices de Shapley en tant qu’explications locales.

3.2 Chapitre 3: Please stop using SHAP and LIME as Local Explanations and use Regional Explanations instead

Ce chapitre constitue une extension directe du précédent, approfondissant notre analyse des méthodes d’attribution locales, notamment les plus populaires telles que les indices de Shapley et la méthode LIME [Ribeiro, 2016a]. À travers plusieurs exemples, nous mettons en évidence les limitations de ces techniques pour donner une explication locale, même dans le cas de fonc-

tions de régression simples et lorsque les variables sont indépendantes. Face à ces constats, nous proposons une nouvelle approche pour le problème d’attribution locale. Notre approche consiste tout d’abord à trouver une partition pertinente de l’espace des entrées, afin de délimiter des régions spécifiques. Ensuite, nous appliquons des techniques d’attribution globale à chaque région ainsi définie. Cette approche nous permet de bénéficier de toutes les garanties statistiques offertes par les méthodes d’attribution globale. En exploitant les avantages des méthodes d’attribution globale tout en conservant une granularité locale, nous sommes en mesure d’obtenir des résultats plus fiables et plus robustes lors de l’explication des prédictions spécifiques des modèles.

3.3 Chapitre 4: Consistent Sufficient Explanations and Minimal Local Rules for explaining regression and classification models

Dans ce chapitre, nous nous efforçons d’aller au-delà des méthodes d’attribution locale, dont la nature additive limite leur capacité à saisir les interactions. Pour illustrer ce problème, prenons l’exemple du processus de croissance d’une plante. Pour pousser, une plante a besoin à la fois d’eau et de soleil. Si elle n’a que l’un des deux éléments, elle ne pourra pas se développer. Il n’est pas très pertinent d’expliquer ce phénomène en disant que la contribution de l’eau et du soleil est de moitié chacun, car il y a un effet d’interaction difficile à appréhender avec les attributions. Ainsi, notre objectif est d’aller au-delà de l’importance d’une variable individuelle et d’être capable de mesurer simultanément l’importance d’un groupe de variables localement. De plus, nous souhaitons disposer d’une méthode d’explication basée sur des quantités théoriques bien définies, contrairement aux approches précédentes telles que les indices de Shapley et LIME. Dans ce chapitre, nous développons une méthode permettant de trouver et de sélectionner un sous-ensemble minimal de variables qui maintiennent la décision quelle que soit la variation des autres variables selon la distribution des données. En guise d’illustration, considérons une observation (\mathbf{x}, y) avec $\mathbf{x} \in \mathbb{R}^p, y = 1$ issue d’un processus génératif $(\mathbf{X}, Y) \sim P_X P_{Y|\mathbf{X}}$, nous cherchons le sous-ensemble minimal de variables $\mathbf{x}_S, S \subseteq \{1, \dots, p\}$ tel que

$$\mathbb{P}(Y = 1 \mid \mathbf{X}_S = \mathbf{x}_S) \geq \pi,$$

avec π une probabilité relativement grande. Cette approche permet l’identification des interactions entre les variables, en déterminant le sous-ensemble minimal de variables capable de maintenir la prédiction. L’algorithme de Random Forest [Breiman, 2001] est au cœur de cette méthode, car il nous permet d’estimer efficacement les différentes espérances conditionnelles, tout en nous aidant à trouver plus rapidement ce sous-ensemble de variables importantes. De plus, nous avons étendu cette approche pour construire des règles explicatives qui décrivent le comportement du modèle localement. Ces règles sont également déduites à partir du partitionnement appris par la Random Forest. Ces approches sont applicables aux problèmes de régression et de classification.

3.4 Chapitre 5. Rethinking Counterfactual Explanations as Regional and Local Policies

Ce chapitre est le dual du chapitre précédent où nous utilisons essentiellement la même approche pour générer des actions contrefactuelles, qui consiste à identifier les changements minimaux des variables qui permettraient de modifier la prédiction du modèle. Nous débutons en recherchant le sous-ensemble minimal de variables qui permet de modifier la prédiction, puis nous identifions les règles de décision locales qui permettent également de modifier la décision. Une fois de plus, l'algorithme de Random Forest joue un rôle central dans cette méthode, car il permet d'identifier le sous-ensemble de variables qui influence la décision, ainsi que les règles pertinentes. De plus, il nous aide à trouver les zones à haute densité, ce qui permet de générer des actions contrefactuelles plausibles et stables. Notre approche permet aussi de proposer des actions pour changer directement la vraie sortie Y des observations \mathbf{X} .

3.5 Chapitre 6. Adaptive Conformal Prediction By Reweighting Nonconformity Score

Dans ce chapitre, nous proposons d'améliorer les intervalles prédictifs retournés par la Prédiction Conforme (PC). Pour un jeu de calibration donné $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ et une observation de test $(\mathbf{X}_{n+1}, Y_{n+1})$, la PC permet de construire un ensemble $\hat{C}(\mathbf{X}_{n+1})$ à partir de l'observation \mathbf{X}_{n+1} , un modèle prédictif \hat{f} et du jeu de calibration \mathcal{D}_n qui a un taux de couverture marginal,

$$\mathbb{P}_{P^{n+1}} \left\{ Y_{n+1} \in \hat{C}(\mathbf{X}_{n+1}) \right\} \geq 1 - \alpha,$$

où $\mathbb{P}_{P^{n+1}}$ est la probabilité jointe des $n+1$ observations $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_{n+1}, Y_{n+1})$. Pour garantir le taux de couverture marginal, la prédiction conforme calcul un terme de correction \hat{Q} qu'elle ajoute à toutes les prédictions pour construire l'intervalle $\hat{C}(\mathbf{X}_{n+1}) = [\hat{f}(\mathbf{X}_{n+1}) \pm \hat{Q}]$. Pour améliorer les intervalles prédictifs, nous proposons une technique de pondération qui permet d'avoir un terme de correction adaptatif pour chaque instance $\hat{C}(\mathbf{X}_{n+1}) = [\hat{f}(\mathbf{X}_{n+1}) \pm \hat{Q}(\mathbf{X}_{n+1})]$, rendant ainsi les intervalles plus adaptatifs tout en gardant les garanties de couverture. Nous nous sommes aussi intéressés à la garantie du taux de couverture conditionnel au jeu de calibration. En effet, le contrôle du taux de couverture marginal ne correspond pas à la couverture que nous souhaitons contrôler en pratique, car elle garantit la couverture en utilisant un jeu de calibration différent pour chaque observation de test, ce qui n'est pas réaliste. En pratique, il est plus intéressant de pouvoir contrôler le taux de couverture en utilisant le même jeu de calibration pour l'ensemble des observations de test. Nous souhaitons un taux de couverture qui vérifie la propriété de PAC (Probably Approximately Correct) [Valiant, 1984] suivante

$$\forall \alpha, \delta \in (0, 1), \quad \mathbb{P}_{P^n} \left\{ \mathbb{P}_P \left\{ Y_{n+1} \in \hat{C}(\mathbf{X}_{n+1}) \mid \mathcal{D}_n \right\} \geq 1 - \alpha \right\} \geq 1 - \delta.$$

Nous proposons une étape supplémentaire de calibration afin que les intervalles adaptatifs que nous proposons vérifient aussi le taux de couverture PAC. La Random Forest est au coeur de la pondération que nous introduisons. Nous utilisons également les poids de la Random

Forest pour regrouper les instances similaires, ce qui permet d’accélérer le calcul de la calibration et d’améliorer l’adaptabilité des intervalles prédictifs. Pour terminer, nous montrons qu’asymptotiquement les intervalles de prediction de notre approche vérifient la propriété de couverture conditionnel suivante

$$\mathbb{P}(Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1}) \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1}) = 1 - \alpha.$$

3.6 Chapitre 7. Future works

Dans ce chapitre, nous présentons deux travaux en cours. Le premier est une extension du chapitre 6, dans le but d’apporter davantage de confiance aux modèles de machine learning en gérant l’incertitude associée aux prédictions. Nous proposons une méthode de prédiction avec abstention, où le modèle s’abstient de faire une prédiction si l’incertitude est trop élevée. Nous avons formalisé ce problème comme un problème de tests multiples [Jin, 2022; Bates, 2023], où nous testons si l’erreur du modèle dépasse un certain seuil. Nous utilisons la prédiction conforme pour contrôler le taux de faux positifs de manière non asymptotique. Cette approche vise à améliorer la fiabilité des prédictions en évitant les décisions lorsque l’incertitude du modèle est trop élevée. La deuxième partie du chapitre consiste à réaliser un croisement entre l’explicabilité et la prédiction conforme, dans le but d’apporter des garanties statistiques non asymptotiques aux méthodes d’explicabilité. Nous explorons les possibilités d’intégrer des mesures d’incertitude basées sur la prédiction conforme dans les techniques d’explicabilité existantes. Un exemple d’application consiste à utiliser la prédiction conforme pour tester si les contrefactuels générés sont des outliers.

Chapter 1

State Of The Art

Abstract

In this chapter, we present a comprehensive review of two pillars of trustworthy machine learning: Explainable AI and Distribution-Free Uncertainty Quantification. The first part discusses inherently interpretable methods and post-hoc techniques, with emphasis on those offering local explanations. We delve into the functionalities of these models and methods, highlighting their strengths and weaknesses. In the second part, we give an introduction to conformal prediction, presenting its foundations, limitations, and the latest research trends, offering readers an up-to-date perspective on the advancements in this field.

Contents

1	Explainable AI	15
2	Distribution-Free Predictive Inference	27

1 Explanaible AI

In this thesis, our primary focus is on local explanations or uncovering the rationale behind individual predictions. We define explanations as answers to "why-question" and, in particular, for machine learning models to be able to answer this question:

"Why does the model predict Y given \mathbf{X} ?"

To provide an answer, we aim to address the following related questions:

1. Which features are important for this specific prediction?
2. Under what conditions does the model favor a certain prediction?
3. Why did the model choose one prediction over another?

Hence, our review is mainly directed toward methods that are appropriate for this purpose. A more comprehensive review of the literature addressing global explanations can be found in [Molnar, 2022]. This section begins by exploring interpretable or glass-box models, which offer direct insights into the prediction process for each instance, then, we present various explanation methods that can be employed to shed light on predictions generated by black-box models.

1.1 Interpretable models

Let us consider an observation (\mathbf{X}, Y) , where $\mathbf{X} = (X_1, \dots, X_p) \in \mathcal{X}$ represents the input variables and $Y \in \mathcal{Y}$ represents the output variable. These variables are generated from a process $(\mathbf{X}, Y) \sim P_{\mathbf{X}}P_{Y|\mathbf{X}}$. The conditional distribution $P_{Y|\mathbf{X}}$ can represent either a random process or a deterministic model f that we aim to explain. In the deterministic scenario, we have $Y = f(\mathbf{X})$.

In interpretable machine learning, the main idea is to approximate f using a simpler model that the target audience can easily understand and interpret. This process can be perceived as distilling [Hinton, 2015; Zhou, 2023; Gou, 2021] a complex model into a simpler one, enabling us to derive insights into the relationship between input variables and output. [Bénard, 2021a; Murdoch, 2019] emphasize that interpretable models must satisfy three essential properties: Simplicity, Stability, and Accuracy. Simplicity can be measured in terms of the number of operations, variables used, or the simulatability of the interpretable models. Simulatability refers to the ability of the target audience to internally simulate and reason about the complete prediction process of the interpretable model. This requirement poses a stringent limitation on the model and is only feasible when the important features are few and the model's underlying relationship is simple. Decision trees are a well-known example of simulatable models. Stability is another crucial criterion for interpretable methods. Instability often indicates arbitrary inferences and reduces confidence in the predictions. It is imperative to ensure the model's stability to draw reliable statistical conclusions [Yu, 2013]. This can involve assessing whether the prediction remains consistent under minor input perturbations [Alvarez-Melis, 2018] or slight variations in the training set [Bousquet, 2002]. Lastly, to ensure the validity of insights derived from the

interpretable model, it is essential that it is accurate and faithfully represents the model it seeks to elucidate. Otherwise, the extracted information might be inaccurate.

1.1.1 Linear model

The simplest interpretable model is the linear model, which predicts the output of a given instance \mathbf{x} as the sum of the products between the feature values and corresponding coefficients:

$$\hat{f}(\mathbf{x}) = \sum_{j=1}^p \hat{\beta}_j x_j. \quad (1.1)$$

This additive combination allows for easy separation of individual effects. Nonetheless, to maintain interpretability, the linear model should have a limited number of nonzero coefficients to ensure simplicity. The Lasso penalty [Tibshirani, 1996] can be employed to discover such sparse linear models. However, sparse linear models are known to be unstable when features have dependencies [Meinshausen, 2010b; Hebiri, 2012], leading to unreliable interpretations. Although there exist strategies to stabilize sparse linear model [Zou, 2005; Bach, 2008; Meinshausen, 2010b; Lim, 2016; Hastie, 2015], if the target function f cannot be adequately captured by a linear model, especially in the presence of interactions, then the practical applicability of the linear model as an interpretable model may be circumscribed.

1.1.2 Generalized Additive Model

The linear model belongs to a more general class of functions called Generalized Additive Models (GAM) [Hastie, 1987; Stone, 1985a]. GAMs express functions as a sum of univariate functions:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^p \hat{f}_i(x_i),$$

where originally the \hat{f}_j are spline functions [Wahba, 1990], but in a broader sense, they can be any functions. [Lou, 2013] have improved the predictive power of this model by using tree-based methods, such as Random Forest or Boosted Trees, as the univariate functions \hat{f}_j and including pairwise interactions:

$$\hat{f}(x) = \sum_{i=1}^p \hat{f}_i(x_i) + \sum_{i,j=1}^p \hat{f}_{i,j}(x_i, x_j).$$

This model, known as the Explainable Boosting Machine (EBM) [Nori, 2019], has demonstrated performance comparable to state-of-the-art methods for tabular data such as Random Forest and XGBoost. While this model may initially appear to be a black-box due to the complexity of the f_j functions, its modularity allows for the extraction of local explanations and the ability to answer question 1, *which features are important for this specific decision?*, assuming the model is accurate and stable. For instance, the contribution of x_i to the prediction $\hat{f}(\mathbf{x}) = \hat{f}(x_1, \dots, x_p)$

can be quantified as:

$$\phi_{x_i} = \hat{f}_i(x_i) + \sum_{j=1}^p \hat{f}_{i,j}(x_i, x_j).$$

However, like multicollinearity in linear models, GAMs can suffer from feature dependencies known as Concurvity [Buja, 1989; Ramsay, 2003], which is a nonparametric counterpart to multicollinearity. In essence, if each function f_j belongs to a class of functions \mathcal{F}_j , such as splines, boosted trees, or high-order polynomials, Concurvity arises when one variable, say X_p , can be approximated as a linear combination of the others, as follows:

$$f_p(X_p) \approx \sum_{i=1}^{p-1} f_i(X_i). \quad (1.2)$$

Note that the model becomes nonidentifiable if the equality in Equation (1.2) holds exactly. When Concurvity is present, GAM estimates tend to be highly unstable, resulting in disparate interpretations depending on the initialization. Some recent works propose variable selection [Kovács, 2022] or regularization techniques [Siems, 2023] to mitigate these challenges.

1.1.3 Decision Tree

A decision tree is a machine learning algorithm that recursively partitions data into cells with similar output. The algorithm starts at the root node, considering the entire data set. It identifies the best feature to split on, according to a specified criterion, such as minimizing entropy for classification or variance for regression tasks. The data are then divided into two child nodes. This process is then recursively applied to each child node, resulting in a binary tree structure until a termination condition is met. The terminal condition can be: all instances at a node having the same value for the target variable, a node containing fewer instances than a predetermined threshold, or the tree reaching a predetermined maximum depth. An illustration of a decision tree is shown in Figure 1.1. The most popular decision trees are CART [Breiman, 1984], ID3 [Quinlan, 1986], C4.5 [Quinlan, 2014], and RIPPER [Cohen, 1995].

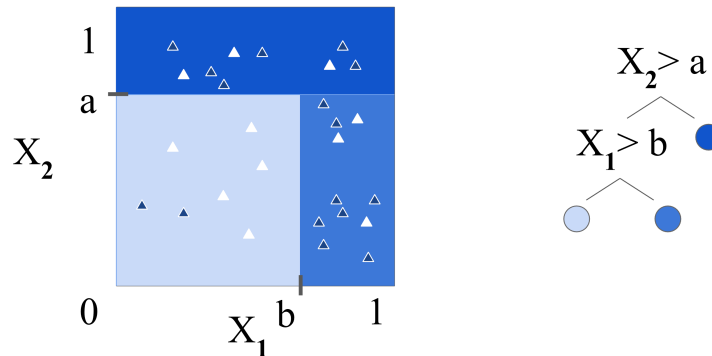


Figure 1.1: Illustration of a decision tree as a piece-wise constant model and binary tree [Singh, 2021].

A decision tree offers a highly interpretable model, as every prediction of the tree corresponds to a specific condition on certain features. This provides a transparent and logical rationale for each prediction, effectively addressing question 2: *under what conditions does the model favor a particular decision?* However, most decision tree algorithms utilize a greedy local optimization strategy, which tends to result in a suboptimal tree. This process often generates trees with excess depth, compromising the inherent simplicity that is one of the key advantages of the method.

Full optimization of decision trees is NP-hard, and a polynomial-time approximation is impossible [Laurent, 1976]. However, some algorithms have been developed recently [Hu, 2019; Lin, 2020] that can discover optimal (or provably near-optimal) decision trees for classification tasks within a reasonable time. Similar progress has also been made for regression tasks [Zhang, 2022].

1.1.4 Decision Rule

A rule is a simple IF-THEN statement that defines a condition on the input variables and specifies the corresponding predicted output. For a given output y , and intervals $R_i \subseteq \mathbb{R}, i = 1, \dots, p$, a rule can be defined as

$$\begin{aligned} &\text{IF } X_1 \in R_1 \text{ AND } X_2 \in R_2 \dots \text{ AND } X_p \in R_p \\ &\text{THEN predict } y. \end{aligned}$$

In high-dimensional spaces, to ensure interpretability or simplicity, it is desirable to have a limited number of active conditions (i.e., small number of intervals R_i that are not equal to \mathbb{R}). Capturing diverse data patterns often requires multiple rules. This raises the question of how to effectively combine these rules. There are two common approaches: rule lists [Rivest, 1987] and rule sets [Cohen, 1999]. In a rule list, the rules are ordered. When applying the rules to an instance, we start with the first rule. If the condition of the first rule holds true, we use its associated prediction. If not, we move on to the next rule and check if it applies. This sequential evaluation continues until a matching rule is found or all rules have been exhausted. On the other hand, a rule set combines multiple rules without a specific order. Each rule in the set can contribute to the final prediction independently, and the prediction can be determined by aggregating the individual rule predictions in some manner, such as voting or averaging. As decision tree, rule-based models address question 2: *under what conditions does the model favor a particular decision?* Figure 1.2 illustrates the difference between a rule set, rule list, and decision tree (rule tree).

Historically, each rule is learned using a greedy heuristic, where elementary constraints are added one by one to maximize a given loss function on the training set. A comprehensive review of rule learning algorithms can be found in [Fürnkranz, 2015]. Another strategy is to extract rules from decision trees or tree-based ensemble models. One of the most popular rule-based algorithms are RuleFit [Friedman, 2008] and Node Harvest [Meinshausen, 2010a]. These algorithms extract rules from tree-based ensembles, treating the activations of these rules as binary variables. Each

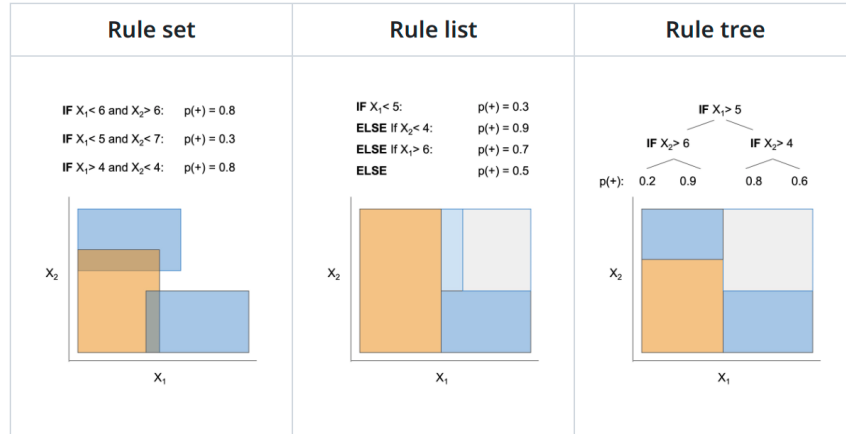


Figure 1.2: Difference between rule set, rule list and decision tree (rule tree) [Singh, 2021]

variable is assigned a value of 1 if the rule is active and 0 otherwise. These binary variables are then combined linearly within a sparse linear model using Lasso penalty [Tibshirani, 1996] for RuleFit and constraint quadratic linear program for NodeHarvest.

However, RuleFit and Node Harvest have displayed certain limitations, such as instability and a tendency to generate longer rules, which can impede interpretability. To overcome these challenges, recent advancements have introduced novel approaches. One such approach is SIRUS [Bénard, 2021c], which extracts rules from small trees utilizing empirical quantiles as split values to improve stability. Another approach is FIGS [Tan, 2022], which is a generalized version of the CART algorithm designed to handle additive functions, improving the effectiveness of rule set methods. In a different direction, [Agarwal, 2022] proposes a post-hoc algorithm known as "Hierarchical Shrinkage" that improves the predictive performance of any tree-based models without changing the structure of the trees, instead, it regularizes each leaf's prediction. The algorithm replaces the average response over a leaf in the tree with a weighted average of the mean responses over the leaf and each of its ancestors. Similarly, [Breiman, 1976; Bloniarz, 2016; Friedberg, 2020; Künzel, 2022] propose fitting a linear model in each leaf to adapt to smooth signal. A comprehensive list and implementation of rule-based methods can be found in the package imodels [Singh, 2021].

1.2 Post-hoc explanations

In contrast to the interpretable machine learning framework, which aims to construct a simple surrogate model of f , post-hoc explanation methods directly operate on the model f itself to explain specific predictions $f(\mathbf{x})$.

1.2.1 Explaining by visualisation

The most natural way to analyze the effect of a given feature X_j on the model $f(X_1, \dots, X_p)$ is to plot the function, which allows us to visualize the variation or behavior of the function. However, visualizing a function becomes challenging when it involves more than three variables.

See this interesting video ¹ showing how to represent the fourth dimension. Partial Dependence Plot (PDP) [Friedman, 2001a] offers a methodology to plot and understand complex functions with multiple variables. The idea is to plot the marginal expectation of f given $X_j = x_j$, which can be expressed as:

$$\hat{f}(x_j) = \mathbb{E}[f(\mathbf{X}_{-j}, x_j)].$$

However, this method does not account for the dependencies among features. Consequently, this approach may evaluate the model at potentially unrealistic or impossible points, thus limiting its reliability. Marginal plots (M plots) [Apley, 2020] are alternatives to PDP that avoid such extrapolation by using the conditional expectation in place of the marginal expectation. Hence, plotting the function

$$\hat{f}(x_j) = \mathbb{E}[f(\mathbf{X}_{-j}, x_j) | X_j = x_j].$$

Although M plots avoid the extrapolation problem, it faces a limitation in distinguishing the primary effect of X_j from the effect of variables that depend on X_j . For instance, if f relies on both X_1 and X_2 , and X_1 and X_2 are dependent, modifying X_1 would influence X_2 , thereby the plot of $\hat{f}(x_j) = \mathbb{E}[f(\mathbf{X}_{-j}, x_j) | X_j = x_j]$ against x_j will reflect both of their effects. Accumulated local effects (ALE) [Apley, 2020] aims to mitigate the influence of interdependent variables by averaging the variations of the function f and accumulating them over a grid. By doing so, ALE allows us to focus on the isolated effect of a particular variable while minimizing the confounding effects caused by interdependencies. The ALE method utilizes the following function

$$\hat{f}(x_j) = \int_{x_{\min}}^{x_j} \mathbb{E} \left[\frac{\partial f}{\partial X_j}(\mathbf{X}_{-j}, z_j) | X_j = z_j \right] dz_j.$$

These methods answer question 1: *which features are important for this specific prediction?* Figure 1.3 illustrates the ALE and PD main-effect plots for the variable "feeling temperature" in a neural network model fitted to predict bike-sharing rental counts [Kaggle, 2015]. The PD plots shown in Figure 1.3b indicates that the number of bike rentals monotonically increases as the feeling temperature rises, even at feeling temperatures exceeding 40 degrees Celsius. However, the ALE plot for feeling temperature shown in Figure 1.3a aligns much better with common sense, indicating that bike rentals will decrease as the feeling temperature increases beyond the comfortable range. However, care must be taken when interpreting the curve. It may be tempting to perceive the curve as representing how the function varies when the other variables are fixed, and we gradually modify the "feeling temperature" variable. In reality, the ALE approach first decomposes the support of the variable into bins and calculates the variations of the function within each bin. It then accumulates the interval-wise effects to create a smooth curve.

¹<https://www.youtube.com/watch?v=IbV0UoXXcOY>

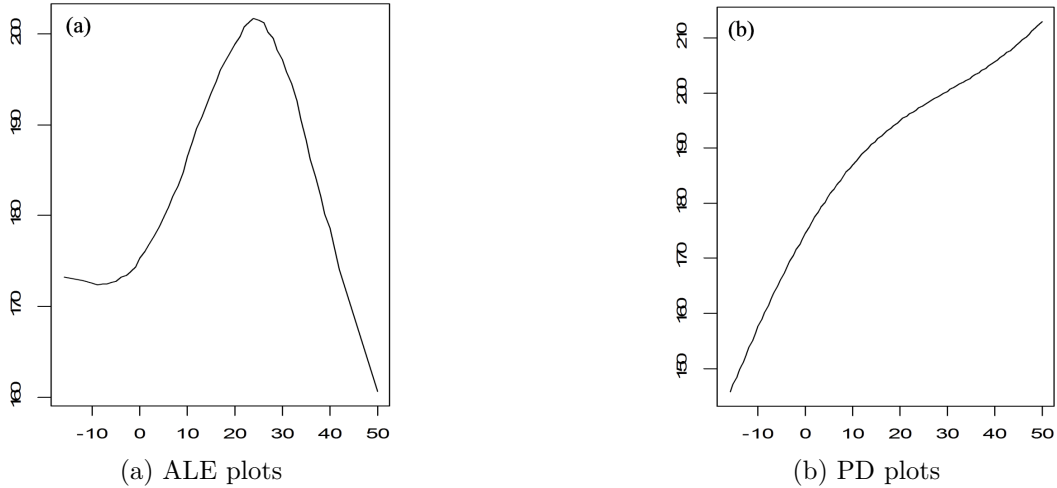


Figure 1.3: For the bike-sharing data example with neural network predicted counts for $f(x)$, ALE main-effect plot (left panel) and PD main-effect plot (right panel) for the variable feeling temperature. The two plots differ substantially, and the ALE plot seems to agree more with intuition [Apley, 2020].

[Apley, 2020] has demonstrated that ALE correctly recovers individual feature contributions for additive functions, regardless of correlation, and for multiplicative functions without uncorrelated features. [Grömping, 2020] analyzed the properties of ALE in a simple linear model with main effects and second-order interactions ($f(\mathbf{X}) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$). He showed that ALE main-effects plots are unaffected by noninteracting correlated features and interactions with uncorrelated features. However, the main effect can be influenced by interacting correlated features. [Apley, 2020] considers this as normal, while [Grömping, 2020] sees it as flawed. Despite this, the main difficulty of this approach lies in the estimation of conditional estimands.

1.2.2 Local Shapley Values

Local Shapley Values [Lundberg, 2017a] is widely recognized as one of the most used local explanation methods, mostly due to its nice implementation [Lundberg, 2017a] and game-theoretical foundation. It drew inspiration from cooperative game theory [Osborne, 1994], which primarily deals with the redistribution of contribution within a group. The original idea consists of a set of players $D = \{1, \dots, d\}$, and a value function $v : \mathcal{P}(D) \rightarrow \mathbb{R}$, where $\mathcal{P}(D)$ is the set of all subsets of D , that represents the value of each coalition of players $S \in \mathcal{P}(D)$. The primary objective is to determine an allocation strategy for redistributing the total contribution $v(D)$ among players. A popular allocation method is Shapley Values [Shapley, 1953] that defined the attribution of each player $i \in D$ as

$$\phi_i = \frac{1}{d} \sum_{S \subseteq D \setminus \{i\}} \binom{p-1}{|S|}^{-1} [v(S \cup i) - v(S)]. \quad (1.3)$$

Intuitively, the SV is the weighted average of the marginal contribution of player i , $v(S \cup i) - v(S)$, across all subsets $S \subseteq D \setminus \{i\}$. Four interesting properties arise from this allocation:

-
- Efficiency: $\sum_{i \in D} \phi_i = v(D)$.
 - Dummy: if $v(S \cup i) = v(S)$ for all $S \in \mathcal{P}(D)$, then $\phi_i = 0$.
 - Symmetric: if $v(S \cup i) = v(S \cup j)$ for all $S \in \mathcal{P}(D)$, then $\phi_i = \phi_j$.
 - Linearity: if two value function v and v' yields to Shapley Values ϕ_i and ϕ'_i , respectively, then the value function $v + v'$ yields Shapley Values $\phi_i + \phi'_i$.

[Shapley, 1953] proved Equation 1.3 is the only allocation method that satisfies these properties.

The framework of Shapley Values has been adapted for global sensitivity analysis, known as Shapley Effects [Owen, 2014; Song, 2016; Owen, 2017]. In this adaptation, the players are reinterpreted as the variables $\mathbf{X} = (X_1, \dots, X_p)$ of a given model f , and $v(S)$ measures the part of the variance of $Y = f(\mathbf{X})$ caused by the variables \mathbf{X}_S as $v(S) = \mathbb{V}(\mathbb{E}[Y|\mathbf{X}_S])/\mathbb{V}(Y)$. A variant of the Shapley Effect, known as SAGE (Shapley Additive Global importance) [Covert, 2020d], consists of defining $v(S) = \mathbb{E}[(Y - \mathbb{E}[Y|\mathbf{X}_S])^2] - \mathbb{E}[(Y - \mathbb{E}[Y|\mathbf{X}])^2] = \mathbb{E}[\mathbb{V}(Y|\mathbf{X}_S)]$, which can be interpreted as the drop in prediction accuracy when \mathbf{X}_S is omitted. By construction, Shapley Effects sum to one and are positive since the value function is increasing, i.e., $A \subseteq B, v(A) \leq v(B)$. Consequently, Shapley Effects provides a valuable alternative to traditional functional ANOVA decomposition [Hoeffding, 1948] or Sobol’s indices [Sobol, 1990; Chastaing, 2012], particularly when features are dependent [Owen, 2017]. It equally distributes the mutual contribution of each subset, accounting for dependencies and interactions, to each individual variable within the subset. Shapley Effects answer to question 1: *which features are important for this specific prediction?*

However, Shapley Effects can suffer from one main drawback: if exogenous inputs (i.e., not used by the model f) are sufficiently correlated with endogenous inputs, their SV can be non-zero [Herin, 2022; Verdinelli, 2023]. To address this issue, recent allocation methods have been developed to resolve the non-zero Shapley Effects problem caused by correlated exogenous inputs. This approach, known as Proportional Marginal Effects, was first introduced by [Feldman, 2005], and has been further developed by [Herin, 2022] for global sensitivity analysis.

In this thesis, we are particularly interested in local explanations. To cater to this need, [Lundberg, 2017a] proposed an adaptation of Shapley Effects for local explanations by considering the conditional expectation as value function $v(S) = \mathbb{E}[f(\mathbf{X})|\mathbf{X}_S = \mathbf{x}_S]$. We called the corresponding attribution as Local Shapley Values. These Shapley Values encounter the same issue as the Shapley Effects - they cannot discern exogenous variables to the models. Consequently, some studies [Heskes, 2020; Janzing, 2020; Chen, 2020] advocate the use of marginal expectation as a value function, expressed as $v(S) = \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}})]$. This variant of Shapley Values assigns zero attribution to the exogenous variables but at the risk of depending on unreliable predictions as the model’s predictions are based on improbable or impossible observations. Another significant challenge posed by Shapley Values is its computational complexity, which increases exponentially with the number of variables (2^d), as well as the approximation of the conditional expectation. The issue of computational complexity is commonly addressed through sampling among the pow-

erset $\mathcal{P}(D)$ [Covert, 2020c; Williamson, 2020; Song, 2016]. Recently, an importance-sampling approach leveraging Random Forest to identify the most significant subsets has been proposed [Bénard, 2021b]. The approximation of the conditional expectation is tackled by employing machine learning models to learn the conditional expectation [Covert, 2020c; Williamson, 2020], or by using parametric distribution such as Gaussian distribution [Aas, 2020] or vine-copula [Aas, 2021] to approximate the features’ distribution. Notably, for tree-based models, polynomial algorithms with simplified expressions of the conditional expectation have been introduced [Amoukou, 2021b; Lundberg, 2020b]. Despite the increasing popularity of Local Shapley Values, as depicted in Figure 1, we have demonstrated in Chapters 2 and 3 that they suffer from several limitations. These include estimation issues and reliability in identifying local important variables.

1.2.3 LIME

Another popular method is Local Interpretable Model-Agnostic Explanations (LIME) [Ribeiro, 2016a]. The core idea is to construct a linear model in the vicinity of each instance and utilize the corresponding linear model coefficients as local explanations. Given an instance \mathbf{x} and model f , the local explanation $\xi(\mathbf{x}^*)$ is a model $g \in G$ where G is the set of linear models such that

$$\xi(\mathbf{x}^*) = \arg \min_{g \in G} \mathcal{L}(f, g, \pi_{\mathbf{x}^*}^h) + \Omega(g), \quad (1.4)$$

where $\mathcal{L}(f, g, \pi_{\mathbf{x}^*}^h)$ measure of how unfaithful g is in approximating f over $\pi_{\mathbf{x}^*}^h$, a measure of locality around x^* with width h , and $\Omega(g)$ is a measure of complexity of the local model g . The loss \mathcal{L} is defined as

$$\mathcal{L}(f, g, \pi_{\mathbf{x}^*}^h) = \sum_{x' \sim P'} (f(x') - g(x'))^2 \pi_{\mathbf{x}^*}^h(\mathbf{x}'),$$

In the original implementation, $\pi_{\mathbf{x}^*}^h$ is a Gaussian kernel, and the sum is done over samples \mathbf{x}' drawn from the marginal distribution $P' = \prod_{i=1}^d P_{X_i}$, where P_{X_i} denotes the marginal distribution of the individual feature X_i . LIME answer to question 1: *which features are important for this specific prediction?*

While the idea behind LIME is appealing, the method faces several technical difficulties. One primary challenge is the lack of a clear guideline for defining the neighborhood $\pi_{\mathbf{x}^*}^h$ or tuning the kernel width h . Each choice of kernel width may yield different explanations, and we have no idea about how to optimize it as we don’t know the ground truth local important variables. Moreover, [Garreau, 2020] have highlighted the impact of poor parametrization on the effectiveness of LIME. Their theoretical analysis on a linear model revealed that LIME’s coefficient is directly proportional to the partial derivatives. Intriguingly, by changing a parameter of the method, it is possible to cause the coefficient of important features to vanish. This emphasizes the vulnerability of LIME to parameter settings, particularly given that we lack systematic strategies for their selection. In addition to the aforementioned challenges, LIME exhibits sta-

bility issues as highlighted by [Alvarez-Melis, 2018]. Their work demonstrates that even very close observations can have completely different explanations, which raises concerns about the reliability and consistency of the method. In Chapter 3, we delve deeper into these limitations and explore the technical difficulties associated with LIME.

1.2.4 Anchors

Similar to LIME, Anchors [Ribeiro, 2018] aims to construct an interpretable model that is locally valid around an instance, specifically in the form of a rule. Let’s define a rule as a pair (A_S, y) where $A_S = [a_1, b_1] \times \dots \times [a_{|S|}, b_{|S|}]$ is a hyperrectangle that represents the conditions of the rule on variables $\mathbf{X}_S, S \subseteq \{1, \dots, p\}, a_i, b_i \in \bar{\mathbb{R}}$ for all $i \in \{1, \dots, p\}$ and y is the output of the rule. Given an instance \mathbf{x} , the explanation $\xi(\mathbf{x})$ of the prediction $f(\mathbf{x})$ is a rule $(A_S, f(\mathbf{x}))$ such that $\mathbf{x}_S \in A_S$ and

$$\mathbb{E}_{Q_S} \left[\mathbb{1}_{f(\mathbf{x})=f(\mathbf{X})} \right] \geq \tau.$$

where $\tau \in (0, 1)$ and the probability is taken under the distribution Q_S , which typically represents the marginal distribution of the features \mathbf{X}_S such that $\mathbf{X}_S \in A_S$ or $Q_S \propto \prod_{i=1}^p P_{\mathbf{X}_i} \mathbb{1}_{\mathbf{X}_S \in A_S}$. The intuition behind is to find a minimal set of conditions or rules satisfied by the instance \mathbf{x} as well as by its neighboring observations having the same output as $f(\mathbf{x})$. Anchors aims to answer question 2: *under what conditions does the model favor a certain prediction?*. The biggest challenge here is the learning of the rule $(A_S, f(\mathbf{x}))$ while minimizing $|S|$ for simplicity and ensuring that the rule contains enough observation to be meaningful. The exact solution to this problem is intractable. Anchors propose a heuristic strategy using beam search to construct the rules and a multi-armed bandit approach for exploration. However, this method has several limitations. For instance, it requires discretizing variables, which can lead to poor results depending on the number of bins. It evaluates the model on impossible data to find the rule. Additionally, as LIME, the method is quite unstable due to the numerous hyperparameters involved in the approach. Nonetheless, [Lopardo, 2023] provide a theoretical analysis of Anchors for text classification. Their analysis assumes an exhaustive search for the best subset S and a linear or rule-based predictive model f . Anchors demonstrates meaningful results under these assumptions. Notably, they show that if a word is not utilized by the classifier, it will not be included in the subset S .

1.2.5 Counterfactual Explanations

Counterfactual Explanations (CE), also known as Algorithmic Recourse, have emerged as a concept within the explainable AI community, drawing inspiration from the notion of adversarial examples [Goodfellow, 2014]. A counterfactual explanation of a prediction describes the smallest change to the feature values that changes the prediction to a predefined output [Wachter, 2017]. This approach is motivated by the observation that human explanations are often contrastive in nature. In many cases, what we seek to understand is not the complete set of causes or explanations for a given event, but rather why that event occurred instead of an alternative

event [Miller, 2019]. For instance, if someone wants to understand why his loan application was rejected and how he can improve his chances of getting a loan, the question of "why" can be formulated as a counterfactual scenario: what is the minimum change to his income, number of credit cards, and other relevant features that would change the prediction from rejection to approval. The counterfactual action or recourse could be that the loan would have been accepted if his annual income were 10,000 higher.

This notion of counterfactual reasoning is also relevant in the field of causality. Lewis [Lewis, 1973] argues that causation can be understood in relation to an imagined counterfactual scenario. Event C is considered to have caused event E, if, in a hypothetical counterfactual case where event C did not occur, event E would not have occurred either. This counterfactual reasoning provides insights into causal relationships and helps us understand the influence of specific factors on the outcome. In summary, counterfactual explanations capture the idea of minimal feature changes required to achieve a different prediction outcome.

For instance, consider a binary classifier $f : \mathcal{X} \rightarrow \{0, 1\}$ and an instance $\mathbf{x} \in \mathcal{X}$ where $f(\mathbf{x}) = 0$. The objective of counterfactual explanation is to find an action \mathbf{a} often called recourse that alters the prediction to $f(\mathbf{x} + \mathbf{a}) = 1$, while minimizing a cost function $c(\mathbf{a} | \mathbf{x})$. This can be formulated as follows:

$$\min_{\mathbf{a}} c(\mathbf{a} | \mathbf{x}) \quad \text{subject to} \quad f(\mathbf{x} + \mathbf{a}) = 1$$

The cost function c may include terms such as $\|\mathbf{a}\|$, ensuring the changes are minimal, and a function that ensures the resulting observation $\mathbf{x} + \mathbf{a}$ remains plausible or satisfies domain-specific constraints. Most CE methods depend on gradient-based algorithms or heuristic approaches [Karimi, 2020b] and are only available for classification. However, there have been recent attempts to extend counterfactual explanations to regression problems [Spooner, 2021]. Counterfactual explanations aims to answer question 3: *why did the model choose one decision over another?*.

This method faces several challenges. One of the challenges is the high number of possible actions that can change the prediction that compromises the intelligibility or simplicity of the resulting explanations, and the synthesis of various recourses or local explanations, in general, remains an unsolved challenge [Lakkaraju, 2022]. Ensuring the plausibility of counterfactual samples $\mathbf{x} + \mathbf{a}$ is another active area of research. One approach to encourage plausibility is by incorporating constraints based on outlier scores into the optimization process. This can be done using techniques such as Local Outlier Factor [Kanamori, 2020], Isolation Forest [Parmentier, 2021], or density-weighted metrics [Poyiadzi, 2019]. Another line of research involves leveraging causal knowledge about the variables and directly intervening on the variables that influence the output Y to generate realistic samples [Kusner, 2017; Joshi, 2019; Mahajan, 2019; Karimi, 2021]. This approach utilizes Pearl's causal modeling [Pearl, 2009], employing a structural causal model (SCM) $\mathcal{M} = \langle \mathbb{F}, \mathbf{X}, \mathbf{U} \rangle$. The SCM consists of endogenous variables $\mathbf{X} \in \mathcal{X}$, exogenous variables $\mathbf{U} \in \mathcal{U}$, and a sequence of structural equations $\mathbb{F} : \mathcal{U} \rightarrow \mathcal{X}$ that specify how \mathbf{X} is determined

from \mathbf{U} . The SCM can be represented as a directed graphical model \mathcal{G} , as shown in Figure 1.4

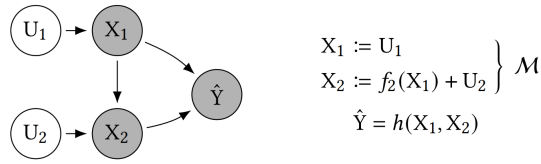


Figure 1.4: Illustration of an example of causal generative process governing the data, showing both the graphical model \mathcal{G} and the structural causal model \mathcal{M} [Karimi, 2021].

From a causal perspective, actions can be carried out using Pearl’s do-operator [Pearl, 1994; Pearl, 2000], which enforces a change in a set of variables while keeping the rest of the causal mechanism untouched, resulting in realistic counterfactuals. However, this approach is often impractical as it relies on knowing the causal graph, and structural equations, which is not feasible in many real-world applications. To address this challenge, [Black, 2020; De Lara, 2021] propose substituting causality-based counterfactual reasoning with optimal transport. In this approach, the counterfactual action is characterized as a coupling between two observable distributions. To better illustrate this concept, let’s consider a binary classification problem. Suppose we aim to find an action that changes the prediction $Y = 0$ of a given observation to $Y = 1$. This can be framed as finding a mapping from the original observation to the most similar observation in the set of observations where $Y = 1$.

Recently, [Ustun, 2019b; Barocas, 2020; König, 2021; König, 2023] have underscored the potential risks associated with the traditional approach to generating counterfactual actions in real-world contexts. The classic approach, introduced by [Wachter, 2017], is framed as an optimization problem. It seeks an action a capable of changing the prediction for a specific observation. Consider a binary classifier $\hat{f}(\mathbf{x}) \in \{0, 1\}$ developed to predict whether a candidate will default on their loan payment, using historical data $\{(X_i, Y_i)\}_{i=1}^n$. Here, \mathbf{X}_i denotes the characteristics of a candidate, and $Y_i \in \{0, 1\}$ represents the true label indicating whether the observation defaults or not. The action a proposed by the traditional counterfactual explanation approach could potentially modify the model’s prediction without actually affecting the true label Y of the resulting observation $\mathbf{x} + a$. Indeed, the fitted model \hat{f} may exploit not only the direct cause of Y , but also the associated variable. Consequently, algorithmic recourse actions suggested by the model may encourage the gaming of the predictor by intervening on non-causal variables, thereby altering the prediction without genuinely changing the true label Y . In response to this issue, [König, 2021; König, 2023] leverages on causal knowledge of variables to suggest actions that can concurrently alter both the model’s prediction and the true label.

[Pawelczyk, 2022] highlights a new problem of CE called: *noisy responses to prescribed recourses*. In real-world scenarios, some individuals may not be able to implement exactly the prescribed recourses, and they show that most CE methods fail in this noisy environment. In Chapter 5, we propose a solution to these problems.

2 Distribution-Free Predictive Inference

In statistics and machine learning, we typically rely on models and algorithms that are valid under certain assumptions, such as parametric model, smoothness, sparsity, or gaussian residuals of the data generating process. However, these assumptions may not always hold in real-world scenarios. In this case, can we trust the output of these algorithms? Moreover, when we attempt to check the assumptions through statistical tests, very often these tests only detect violations under other assumptions. To address this concern, distribution-free inference aims to provide guarantees that are valid universally over all data distributions.

Distribution-free inference encompasses various inference questions. This includes prediction, where we seek to determine the range within which the unobserved response variable Y is expected to lie. Other questions involve testing for independence between \mathbf{X} and Y given specific confounders \mathbf{Z} or constructing a confidence interval for the conditional distribution of $\mathbb{E}[Y|\mathbf{X}]$. The last two tasks are known to be impossible if \mathbf{X} is continuous [Barber, 2020; Shah, 2018] without imposing assumptions on the underlying distribution or straightforward if \mathbf{X} is discrete with bounded values [Lee, 2021]. In this thesis, we are interested in prediction inference.

The setting is the following:

- We start with a training data $\mathcal{D}_m := \{(\mathbf{X}_i, Y_i)\}_{i=1}^m$ drawn from a distribution $P = P_{\mathbf{X}}P_{Y|\mathbf{X}}$.
- An algorithm \mathcal{A} that takes the training data and return a model $\mathcal{A}(\mathcal{D}_m) = \hat{\mu}$, defined as

$$\mathcal{A} : \cup_{m \geq 0} (\mathcal{X} \times \mathcal{Y})^m \mapsto \{\text{measurable functions } \hat{\mu} : \mathcal{X} \rightarrow \mathcal{Y}\},$$

where $\hat{\mu}(\mathbf{x})$ can be an estimator of the conditional expectation $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ or any quantities related to $Y|\mathbf{X} = \mathbf{x}$ including the conditional quantile or conditional density (in the latter, \mathcal{A} maps the data to real-valued function). The algorithm \mathcal{A} is assumed to treat training data points symmetrically, i.e.,

$$\mathcal{A}\left((\mathbf{X}_{\pi(1)}, Y_{\pi(1)}), \dots, (\mathbf{X}_{\pi(m)}, Y_{\pi(m)})\right) = \mathcal{A}\left((\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_m, Y_m)\right),$$

for all $m \geq 1$ and any permutation $\pi : \{1, \dots, m\} \mapsto \{1, \dots, m\}$.

- We are given a test point with an unseen output (\mathbf{X}, Y) also drawn from P
- We aim at constructing a Predictive Interval $\hat{C}(\mathbf{X})$ such that

$$\mathbb{P}\left\{Y \in \hat{C}(\mathbf{X})\right\} \geq 1 - \alpha \text{ for a given } \alpha.$$

- This guarantee should hold for any distribution P and any predictive algorithm \mathcal{A} or model $\hat{\mu}$ with finite-sample or non-asymptotically.

Conformal Prediction (CP) offers a principled approach to address the prediction problem, allowing the construction of predictive intervals, which have finite-sample valid coverage guarantees regardless of the underlying data distribution P and choice of the predictive algorithm \mathcal{A} , under the mild assumptions of exchangeability. Recent developments have extended the scope of CP beyond exchangeability [Tibshirani, 2019; Barber, 2022; Gibbs, 2021; Zaffran, 2022b], and also for non-symmetric algorithm \mathcal{A} [Barber, 2022]. Conformal Prediction was initially introduced by Vladimir Vovk and his collaborators Alexander Gammerman, Vladimir Vapnik, and others between 1996-1999 [Gammerman, 1998; Saunders, 1999; Vovk, 1999]. It gained significant attention and popularity due to the pioneering work by Jing Lei, Larry Wasserman, and their colleagues [Lei, 2011; Lei, 2013; Póczos, 2013; Lei, 2016]. Since then, it has experienced a surge in popularity, thanks to the remarkable contributions of prominent researchers such as Rina Barber, Emmanuel Candès, Aaditya Ramdas, Ryan Tibshirani, Anastasios Angelopoulos, Stephen Bates, Michael I. Jordan, Yaniv Romano, and numerous others [Angelopoulos, 2021a].

2.1 Foundations of Conformal Prediction

Conformal Prediction methods can be broadly divided into two categories: those that involve retraining the model multiple times, such as full conformal [Vovk, 2005] or jackknife methods [Barber, 2021], and those that use sample splitting, known as split conformal methods [Papadopoulos, 2002; Lei, 2016]. The latter is more computationally feasible at the cost of splitting the data. As a result, split conformal approaches are the most used due to their practicality. In order to provide a simple introduction, we only present the split-conformal approach, as it is easier to understand.

Let's assume we have a calibration data set $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, and a training set $\mathcal{D}_m = \{(\mathbf{X}_i, Y_i)\}_{i=1}^m$ used by algorithm \mathcal{A} to generate a model $\hat{\mu} = \mathcal{A}(\mathcal{D}_m)$. The primary goal of CP is to construct a predictive set $\hat{C}(\cdot)$ that covers the unseen output Y_{n+1} given a new test input \mathbf{X}_{n+1} satisfying the marginal coverage guarantees:

$$\mathbb{P}_{P^{n+1}} \left\{ \mathbf{X}_{n+1} \in \hat{C}(\mathbf{X}_{n+1}) \right\} \geq 1 - \alpha, \quad (1.5)$$

where the probability $\mathbb{P}_{P^{n+1}}$ is taken under the joint distribution of the $n+1$ observations, which corresponds to the calibration data and the test observation. α is a predefined miscoverage rate.

In this section, we first define and discuss some properties of exchangeable random variables, which form the foundation of the Conformal Prediction framework. Next, we introduce the Quantile Lemma [Vovk, 2005; Lei, 2016; Tibshirani, 2019], a crucial component that enables the finite-sample marginal coverage guarantees of the Prediction Intervals (PI). Subsequently, we introduce the split-conformal approach. We provide proofs for each result presented in this section, to provide readers with a comprehensive presentation of Conformal Prediction. These proofs are detailed versions of the succinct proofs found in [Tibshirani, 2019; Kuchibhotla, 2020]. Lastly, we delve into strategies for adapting the PI given by CP to address various contexts, such as handling heteroscedastic data or classification problems.

Definition 2.1 (Exchangeability). A sequence of random variables $Z_1, \dots, Z_n \in \mathcal{Z}$ are exchangeable, if and only if, for any permutation $\pi : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$, and every measurable set $E \subseteq \mathcal{Z}^n$, we have

$$\mathbb{P}_{P^n} \{(Z_1, \dots, Z_n) \in E\} = \mathbb{P}_{P^n} \{(Z_{\pi(1)}, \dots, Z_{\pi(n)}) \in E\}. \quad (1.6)$$

In other words, it means that (Z_1, \dots, Z_n) and $(Z_{\pi(1)}, \dots, Z_{\pi(n)})$ have the same joint distribution for any permutation π . It's worth noting that if the variables are independent and identically distributed (*i.i.d.*), then they are necessarily exchangeable, which further implies that they are identically distributed.

A key consequence of exchangeability for real-valued random variables is that the ranks of Z_1, \dots, Z_n are uniformly distributed on $\{1, 2, \dots, n\}$. This is crucial in proving the validity of the PI given by CP.

Lemma 2.2. *Let $Z_1, \dots, Z_n \in \mathcal{Z}$ be exchangeable random variables with no ties almost surely, then their rank are uniformly distributed on $\{1, \dots, n\}$. The law of the ranks $\text{Rank}(Z_i) = R_i$ are*

$$\mathbb{P}_{P^n}(R_i = 1) = \dots = \mathbb{P}_{P^n}(R_i = n) = \frac{1}{n}, \quad \text{for all } i \in \{1, \dots, n\}. \quad (1.7)$$

Proof. We denote S_n as the set of possible permutations on $\{1, \dots, n\}$ and consider the set $E_\pi = \{(z_1, \dots, z_n) \in \mathcal{Z}^n : z_{\pi(1)} \leq \dots \leq z_{\pi(n)}\}$ for $\pi \in S_n$. By construction, the sets E_π are disjoint and $\cup_{\pi \in S_n} E_\pi = \mathcal{Z}^n$, then we have

$$\begin{aligned} 1 &= \mathbb{P}_{P^n} \{\cup_{\pi \in S_n} \{(Z_1, \dots, Z_n) \in E_\pi\}\} = \sum_{\pi \in S_n} \mathbb{P}_{P^n} \{(Z_1, \dots, Z_n) \in E_\pi\} \\ &= \sum_{\pi \in S_n} \mathbb{P}_{P^n} \{Z_{\pi(1)} \leq \dots \leq Z_{\pi(n)}\} \\ &= \sum_{\pi \in S_n} \mathbb{P}_{P^n} \{Z_1 \leq \dots \leq Z_n\}. \end{aligned}$$

Hence, for any permutation $\pi : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ we have

$$\mathbb{P}_{P^n} \{Z_{\pi(1)} \leq \dots \leq Z_{\pi(n)}\} = \mathbb{P}_{P^n} \{Z_1 \leq \dots \leq Z_n\} = \frac{1}{n!}.$$

For all $i, j \in \{1, \dots, n\}$, the event $\{R_i = j\}$ is equal to $\cup_{\pi: \pi(j)=i} \{Z_{\pi(1)} \leq \dots \leq Z_{\pi(n)}\}$, thereby we have

$$\begin{aligned} \mathbb{P}_{P^n}(R_i = j) &= \sum_{\pi: \pi(j)=i} \mathbb{P}_{P^n} \{Z_{\pi(1)} \leq \dots \leq Z_{\pi(n)}\} \\ &= \frac{(n-1)!}{n!} \\ &= \frac{1}{n} \end{aligned}$$

□

The basic idea behind the theory of conformal prediction is the following simple result about sample quantiles, often called quantile lemma. This lemma is the cornerstone of the proof of the marginal coverage (Eq. 1.5) of the CP prediction intervals.

Lemma 2.3 (Quantile lemma [Vovk, 2005; Lei, 2016; Tibshirani, 2019]). *If $Z_1, \dots, Z_{n+1} \in \mathcal{Z}$ are exchangeable random variables with no ties almost surely, then for all $\alpha \in (0, 1)$*

$$\alpha \leq \mathbb{P}_{P^{n+1}} \left\{ Z_{n+1} \leq \mathcal{Q} \left(\alpha; \frac{1}{n+1} \sum_{i=1}^n \delta_{Z_i} + \frac{1}{n+1} \delta_{+\infty} \right) \right\} \leq \alpha + \frac{1}{n+1}, \quad (1.8)$$

where $\mathcal{Q}(\alpha; F)$ represents the α -quantile of the distribution F .

Proof. Given a discrete distribution $F = \frac{1}{n} \sum_{i=1}^n \delta_{z_i}$, with $z_i \in \mathcal{Z}, i = 1, \dots, n$, let us define $q = \mathcal{Q}(\alpha; F) = z_{(\lceil \alpha n \rceil)}$. Notably, even if we change all values $z_i > q$ to arbitrary values strictly larger than q , yielding a new distribution \tilde{F} , we still have $\mathcal{Q}(\alpha; F) = \mathcal{Q}(\alpha; \tilde{F})$. Using this fact, we have

$$Z_{n+1} \leq \mathcal{Q} \left(\alpha; \frac{1}{n+1} \sum_{i=1}^n \delta_{Z_i} + \frac{1}{n+1} \delta_{Z_{n+1}} \right) \iff Z_{n+1} \leq \mathcal{Q} \left(\alpha; \frac{1}{n+1} \sum_{i=1}^n \delta_{Z_i} + \frac{1}{n+1} \delta_{+\infty} \right).$$

In addition, we have

$$\begin{aligned} Z_{n+1} \leq \mathcal{Q}(\alpha; \frac{1}{n+1} \sum_{i=1}^n \delta_{Z_i} + \frac{1}{n+1} \delta_{Z_{n+1}}) \\ \iff Z_{n+1} \text{ is among the } \lceil \alpha(n+1) \rceil\text{-smallest of } Z_1, \dots, Z_{n+1}. \end{aligned}$$

By exchangeability of (Z_1, \dots, Z_{n+1}) , Lemma 2.2 gives that $\text{Rank}(Z_{n+1}) = R_{n+1} \sim \mathcal{U}\{1, n+1\}$,

$$\begin{aligned} & \mathbb{P}_{P^{n+1}} \left\{ Z_{n+1} \leq \mathcal{Q}(\alpha; \frac{1}{n+1} \sum_{i=1}^n \delta_{Z_i} + \frac{1}{n+1} \delta_{+\infty}) \right\} \\ &= \mathbb{P}_{P^{n+1}} \left\{ Z_{n+1} \leq \mathcal{Q}(\alpha; \frac{1}{n+1} \sum_{i=1}^n \delta_{Z_i} + \frac{1}{n+1} \delta_{Z_{n+1}}) \right\} \\ &= \sum_{i=1}^{\lceil \alpha(n+1) \rceil} \mathbb{P}_{P^{n+1}} \{R_{n+1} = i\} \\ &= \frac{\lceil \alpha(n+1) \rceil}{n+1} \end{aligned}$$

We conclude using a simple inequality of the ceiling function: $\alpha \leq \frac{\lceil \alpha(n+1) \rceil}{n+1} \leq \alpha + \frac{1}{n+1}$. \square

Using this lemma as a foundation, we can now introduce one of the most popular methods in conformal prediction, known as the split-conformal. Given a trained model $\hat{\mu} = \mathcal{A}(\mathcal{D}_m)$ using algorithm \mathcal{A} and training data \mathcal{D}_m , we assume the availability of a calibration dataset $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ drawn exchangeably from P . We then compute the residuals of the model on the calibration set, denoted as $V_1 = |Y_1 - \hat{\mu}(\mathbf{X}_1)|, \dots, V_n = |Y_n - \hat{\mu}(\mathbf{X}_n)|$.

The split-conformal method defines the Predictive Interval (PI) for the test point $(\mathbf{X}_{n+1}, Y_{n+1})$ at level $1 - \alpha$ as follows:

$$\widehat{C}(\mathbf{X}_{n+1}) = \left\{ y : |y - \hat{\mu}(\mathbf{X}_{n+1})| \leq \mathcal{Q} \left(1 - \alpha; \frac{1}{n+1} \sum_{i=1}^n \delta_{V_i} + \frac{1}{n+1} \delta_{+\infty} \right) \right\}. \quad (1.9)$$

This predictive interval guarantees exact coverage in finite-sample, meaning that:

$$1 - \alpha \leq \mathbb{P}_{P^{n+1}} \left\{ Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1}) \right\} \leq 1 - \alpha + \frac{1}{n+1},$$

where $\mathbb{P}_{P^{n+1}}$ denotes that the probability is taken over the $n+1$ observations $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_{n+1}, Y_{n+1})$. Indeed, we have:

$$Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1}) \iff V_{n+1} \leq \mathcal{Q} \left(1 - \alpha; \frac{1}{n+1} \sum_{i=1}^n \delta_{V_i} + \frac{1}{n+1} \delta_{+\infty} \right) \quad (1.10)$$

As $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_{n+1}, Y_{n+1})$ are exchangeable, so are V_1, \dots, V_{n+1} and Lemma 2.3 gives the result.

Above, we use the absolute residuals $V(\mathbf{X}, Y) = |Y - \hat{\mu}(\mathbf{X})|$ in the construction of the PI. However, we can generalize to any score function V , where larger scores $V(\mathbf{X}, Y)$ encode worse agreement between $\hat{\mu}(\mathbf{X})$ and Y . V is usually called nonconformity or conformity score and represents the error of the model on (\mathbf{X}, Y) . Consequently, \widehat{C} can be defined as follows:

$$\widehat{C}(\mathbf{X}_{n+1}) = \left\{ y : V(\mathbf{X}_{n+1}, y) \leq \mathcal{Q} \left(1 - \alpha; \sum_{i=1}^n \frac{1}{n+1} \delta_{V(\mathbf{X}_i, Y_i)} + \frac{1}{n+1} \delta_{+\infty} \right) \right\}. \quad (1.11)$$

Theorem 2.4 ([Vovk, 2005; Lei, 2016]). *Assume that $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_{n+1}, Y_{n+1})$ are exchangeable with no ties. For any nonconformity score V , and any $\alpha \in (0, 1)$, the PI defined in (1.11) satisfies*

$$1 - \alpha \leq \mathbb{P}_{P^{n+1}} \left\{ Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1}) \right\} \leq 1 - \alpha + \frac{1}{n+1}. \quad (1.12)$$

The strength of Theorem 2.4 lies in its guarantees of the finite-sample coverage $1 - \alpha$ for any nonconformity score function V , any model $\hat{\mu}$, any miscoverage rate $\alpha \in (0, 1)$, and any dataset under the mild assumption of exchangeability. While being true for any nonconformity score V , it is crucial to carefully select the nonconformity score, as it directly impacts the quality of the predictive intervals. PIs that are overly large or vary weakly with the input, particularly in the presence of heteroscedasticity, are not practically useful. To address heteroscedastic data, we present two approaches: Conformalized Quantile Regression (CQR) [Romano, 2019] and Locally-Weighted CP [Lei, 2016; Papadopoulos, 2008]. These approaches utilize different nonconformity scores to handle heteroscedasticity. Additionally, we introduce an approach that applies the split-conformal technique to classification problems. Instead of predictive intervals, this approach generates predictive sets that consist of the most likely labels.

1. If the noise varies with \mathbf{X} , e.g., $Y = \mu(\mathbf{X}) + \epsilon(\mathbf{X})$, where $\epsilon(\mathbf{X})$ represents heteroscedastic noise and μ is the regression function, employing the absolute residuals $V(\mathbf{X}_i, Y_i) = |\hat{\mu}(\mathbf{X}_i) - Y_i|$ might not be optimal as it yields constant interval widths. To have varying interval widths, we can scale the residuals inversely by an estimated error [Lei, 2016; Papadopoulos, 2008] as:

$$V(x, y) = \frac{|y - \hat{\mu}(x)|}{\hat{\sigma}(x)} \quad \text{where } \hat{\sigma} \text{ is an estimator of the error of the model } \hat{\mu}$$

$$\implies \hat{C}(\mathbf{X}_{n+1}) = [\hat{\mu}(\mathbf{X}_{n+1}) - \hat{\sigma}(\mathbf{X}_{n+1}) \cdot \mathcal{Q}(1 - \alpha; F_n); \hat{\mu}(\mathbf{X}_{n+1}) + \hat{\sigma}(\mathbf{X}_{n+1}) \cdot \mathcal{Q}(1 - \alpha; F_n)].$$

To simplify the notation, we represent the empirical distribution of the nonconformity scores of the calibration data as $F_n = \sum_{i=1}^n \frac{1}{n+1} \delta_{V(\mathbf{X}_i, Y_i)} + \frac{1}{n+1} \delta_{+\infty}$. In Figure 1.5, we compare the intervals return by the absolute and rescaled residuals.



Figure 1.5: Comparison of interval lengths for absolute residuals $V(x, y) = |y - \hat{\mu}(x)|$ (left panel) and rescaled residuals $V(x, y) = \frac{|y - \hat{\mu}(x)|}{\hat{\sigma}(x)}$ (right panel).

2. If the shape of the distribution of $Y|\mathbf{X}$ changes with \mathbf{X} , then centering $\hat{C}(\mathbf{X}_{n+1})$ at $\hat{\mu}(\mathbf{X}_{n+1})$ might not be the optimal strategy. Rather than approximating the conditional mean with $\hat{\mu}(\mathbf{x}) \approx \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$, we can employ quantile regression to estimate conditional quantiles $\hat{q}_{\alpha/2}(\mathbf{x}) \approx \mathcal{Q}(\frac{\alpha}{2}; F_{Y|\mathbf{X}=\mathbf{x}})$ and $\hat{q}_{1-\alpha/2}(\mathbf{x}) \approx \mathcal{Q}(1 - \frac{\alpha}{2}; F_{Y|\mathbf{X}=\mathbf{x}})$ and use the following nonconformity score:

$$V(x, y) = \max\{\hat{q}_{\alpha/2}(x) - y; y - \hat{q}_{1-\alpha/2}(x)\}$$

$$\implies \hat{C}(\mathbf{X}_{n+1}) = [\hat{q}_{\alpha/2}(\mathbf{X}_{n+1}) - \mathcal{Q}(1 - \alpha; F_n); \hat{q}_{1-\alpha/2}(\mathbf{X}_{n+1}) + \mathcal{Q}(1 - \alpha; F_n)].$$

This approach is called Conformalized Quantile Regression (CQR) [Romano, 2019]. Figure 1.6 compares the interval of CQR with the absolute residuals.



Figure 1.6: Comparison of interval lengths for absolute residuals $V(x, y) = |y - \hat{\mu}(x)|$ (left panel) and quantile residuals $V(x, y) = \max\{\hat{q}_{\alpha/2}(x) - y; y - \hat{q}_{1-\alpha/2}(x)\}$ (right panel) [Romano, 2019]

3. In the case where Y is discrete, consider a classification problem with $\mathbf{X} \in \mathcal{X}$, $Y \in \{1, \dots, k\}$. Let $\hat{\mu} : \mathcal{X} \rightarrow [0, 1]^k$ be a probabilistic predictive model, where $\hat{\mu}(x)_k \approx P(Y = k | \mathbf{X} = x)$. A simple nonconformity score is

$$V(\mathbf{X}, Y) = 1 - \mu(\mathbf{X})_Y \implies \hat{C}(\mathbf{X}_{n+1}) = \{y : V(\mathbf{X}_{n+1}, y) \geq \mathcal{Q}(1 - \alpha; F_n)\}. \quad (1.13)$$

This method is called LABEL [Sadinle, 2019], and Figure 1.7 shows an example of the predictive sets on the ImageNet dataset [Deng, 2009]. This method generally results in small prediction sets, but it tends to produce empty ones when the model is uncertain. Alternative nonconformity scores have been proposed for binary and multilabel classification problems, including Top-K [Angelopoulos, 2020], and Adaptive Predictions Sets [Romano, 2020b; Angelopoulos, 2020].



Figure 1.7: Predictive sets’s size reflecting the level of uncertainty [Angelopoulos, 2021a]

Besides the choice of the nonconformity score to enhance the predictive intervals, a natural question is whether it is possible to enhance the statistical efficiency by using a single dataset for both training and calibrating the model, instead of using two separate datasets as with split-conformal. In fact, the first conformal prediction method introduced by [Vovk, 2005] uses a single dataset. While this method guarantees finite-sample marginal coverage, it has a high computational cost. Specifically, a different model needs to be trained for each possible value of $Y_{n+1} = y$. This computational burden limits the scalability and practicality of the method. However, intermediate solutions have been proposed to mitigate this issue by employing a leave-one-out approach, such as jackknife+ and K-fold cross-validation [Barber, 2019a]. These approaches strike a balance between computational and statistical efficiency.

To summarize, given any trained model $\hat{\mu} = \mathcal{A}(\mathcal{D}_m)$ and calibration data $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, the outline of the split-conformal approach is

1. Define a nonconformity score $V(\mathbf{X}, Y)$, where larger scores encode worse agreement between $\hat{\mu}(\mathbf{X})$ and Y .
2. Set \hat{q} as the $[(1 - \alpha)(n + 1)]$ -largest calibration score among $V_1 = V(\mathbf{X}_1, Y_1), \dots, V_n = V(\mathbf{X}_n, Y_n)$
3. Use the quantile \hat{q} to form the prediction sets given \mathbf{X}_{n+1} as

$$\hat{C}(\mathbf{X}_{n+1}) = \{y : V(\mathbf{X}_{n+1}, y) \leq \hat{q}\}. \quad (1.14)$$

2.2 Limitations of Conformal Prediction

While conformal prediction offers marginal coverage guarantees, it is important to acknowledge three main limitations of the current CP framework:

1. The guarantee is on average over the calibration and test point.
2. The guarantee is on average over the test point \mathbf{X}_{n+1} .
3. The assumption of exchangeability.

The guarantee is on average over the calibration and test point. In practical applications, what is of interest is the coverage rate on future test points based on a given calibration set. The marginal coverage guarantee does not directly address this concern. Instead, it bounds the coverage rate on average over all possible sets of calibration and test observations.

To address this limitation, the concept of training-conditional coverage has been introduced [Vovk, 2012]. It ensures that with probability $1 - \delta$ over the calibration samples \mathcal{D}_n , the resulting coverage on future test observation is still above $1 - \alpha$. Formally,

$$\mathbb{P}_{P^n} \left\{ \mathbb{P}_P \left\{ Y_{n+1} \in \widehat{C}(X_{n+1}) \mid \mathcal{D}_n \right\} \geq 1 - \alpha \right\} \geq 1 - \delta.$$

This style of guarantee is also known as ‘‘Probably Approximately Correct’’ (PAC) predictive interval [Valiant, 1984]. The roots of this type of guarantee can be traced back to the earlier works of [Wilks, 1941; Wald, 1943]. Despite the importance of training-conditional coverage in practice, only a few methods have been proven to achieve it. [Vovk, 2012] was the first to establish this result for split conformal methods, and recently [Bian, 2022] has shown that the K-fold CV+ method also achieves it. However, no analogous results are currently known for other CP methods, such as jackknife+ [Barber, 2021] and full-conformal [Vovk, 2005].

The guarantee is on average over the test point \mathbf{X}_{n+1} . The marginal guarantee of CP ensures coverage on average over the test points, but it does not guarantee coverage for each specific observation. Formally, we say that \widehat{C} satisfies distribution-free conditional coverage at level $1 - \alpha$, if

$$\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1}) \mid \mathbf{X}_{n+1} = \mathbf{x}) \geq 1 - \alpha, \text{ for all } P = P_{\mathbf{X}}P_{Y|\mathbf{X}} \text{ and almost all } \mathbf{x}, \quad (1.15)$$

where for a given distribution $P = P_{\mathbf{X}}P_{Y|\mathbf{X}}$, we write ‘‘almost all \mathbf{x} ’’ to mean that the set of point $x \in \mathcal{X}$ where the bound is not true have a measure of zero under $P_{\mathbf{X}}$.

This can lead to unequal coverage rate across different communities of the data, raising fairness issues [Romano, 2020a]. However, obtaining nontrivial distribution-free conditional coverage (Equation 1.15) is proven to be impossible with a finite sample [Lei, 2014b; Vovk, 2012; Barber, 2019b].

Theorem 2.5 ([Lei, 2014b; Vovk, 2012; Barber, 2019b]). *Let’s denote λ the Lebesgue measure, and suppose \widehat{C} satisfies Equation (1.15), then for all distribution P ,*

$$\mathbb{E}[\lambda(\widehat{C}(\mathbf{x}))] = +\infty$$

at almost all nonatomic points \mathbf{x} of $P_{\mathbf{X}}$.

Theorem 2.5 demonstrates that unless the predictive intervals produced by \widehat{C} have infinite expected length under all non-discrete distributions P , it is not possible for \widehat{C} to satisfy Eq. (1.15).

Despite this impossibility result, some approximate solutions have been proposed [Lei, 2014a; Vovk, 2012; Barber, 2019b; Guan, 2022; Amoukou, 2023]. For example, one approach involves finding a relevant partition $\mathcal{X} = \cup_{i=1}^K \mathcal{X}_i$. For each group k , data points $\{(\mathbf{X}_i, Y_i) : \mathbf{X}_i \in \mathcal{X}_k\}$ are exchangeable, allowing the conformal prediction method to be applied separately to each group.

The assumption of exchangeability. In real-world scenarios, the assumption of exchangeability may not hold. Violation of these assumptions includes covariate shift, label shift, and time series. Covariate shift refers to situations where the marginal distribution of the input variable \mathbf{X} differs between the calibration P^{cali} and test data P^{test} , while the conditional distribution $Y|\mathbf{X}$ remains the same:

$$P^{cali} = P_X^{cali} \times P_{Y|\mathbf{X}}, \quad P^{test} = P_X^{test} \times P_{Y|\mathbf{X}}. \quad (1.16)$$

Covariate shift can occur when certain subpopulations are over or underrepresented in the calibration data. If the shift $w(x)$ such that $dP_X^{test} \propto w(x) \times dP_X^{cali}$ is known or approximated, we can still have marginal guarantee using a weighted version of the quantile lemma [Tibshirani, 2019; Hu, 2020]. A similar approach can be applied to address label shift [Podkopaev, 2021], where the marginal distribution of the output variable Y differs between calibration and test data, i.e., $P^{cali} = P_Y^{cali} \times P_{\mathbf{X}|Y}$, $P^{test} = P_Y^{test} \times P_{\mathbf{X}|Y}$.

The case of distribution shift, as encountered in time series data, is still an active area of research. Most existing results focus on achieving asymptotic coverage guarantees under arbitrary distribution drift [Gibbs, 2021; Zaffran, 2022a]. However, recent developments have started addressing finite-sample results beyond the exchangeability assumption [Barber, 2022].

New research directions. Conformal prediction concepts have garnered interest in other fields, extending beyond traditional applications. These areas include missing values [Zaffran, 2023; Gui, 2023], outlier detection [Bates, 2023], survival analysis [Candes, 2023], selective inference [Jin, 2022], risk controls [Bates, 2021; Angelopoulos, 2021b], federated learning [Plassier, 2023], fairness [Gibbs, 2023]. These endeavors demonstrate the versatility and potential of CP in addressing a wide range of statistical and machine learning challenges across various domains.

Chapter 2

Accurate Shapley Values for explaining tree-based models

Abstract

Shapley Values (SV) are widely used in explainable AI, but their estimation and interpretation can be challenging, leading to inaccurate inferences and explanations. As a starting point, we recall an invariance principle for SV and derive the correct approach for computing the SV of categorical variables that are particularly sensitive to the encoding used. In the case of tree-based models, we introduce two estimators of Shapley Values that exploit the tree structure efficiently and are more accurate than state-of-the-art methods. Simulations and comparisons are performed with state-of-the-art algorithms and show the practical gain of our approach. Finally, we discuss the limitations of Shapley Values as a local explanation. These methods are available as a [Python package](#).

Contents

1	Introduction	37
2	Coalition and Invariance for Shapley Values	38
3	Shapley Values for tree-based models	42
4	Comparison of the estimators	48
5	Discussion and Future works	50

1 Introduction

The increasing use of Machine Learning (ML) models in industry, business, and society has brought the explanations of the predictions of these models to the forefront of ML research. As ML models are often considered as black-box models, there is a growing demand from scientists, practitioners, and citizens for tools that can provide insights into important variables in predictions or identify biases for specific individuals or subgroups. Standard global importance measures, such as permutation importance measures [Breiman, 2001] are insufficient to explain individual or local predictions, and new methodologies are being developed in the active field of explainable AI.

In this context, various local explanations have been proposed, focusing on model-agnostic methods that can be applied to the most successful ML models, such as ensemble methods like Random Forests and gradient boosted trees, as well as deep learning models. Some of the most widely used methods are the Partial Dependence Plot [Friedman, 2001b], Individual Conditional Expectation [Goldstein, 2015], and local feature attributions such as Local Surrogate (LIME) [Ribeiro, 2016a]. These techniques aim to better understand the predictions made by a model for individual cases, providing transparency and trust in the ML models' decision-making processes. To achieve the same objective, Shapley Values [Shapley, 1953], a concept developed primarily in Cooperative Game Theory, has been adapted to evaluate the "fair" contribution of a variable-value $X_i = x_i$ to a prediction $f(x_1, \dots, x_p)$ [Strumbelj, 2010; Lundberg, 2017a]. Shapley Values (SV) are widely used to identify important variables at both local and global levels. As remarked by [Lundberg, 2020b; Covert, 2020b], many importance measures aim to analyze the behavior of a prediction model f based on p features X_1, \dots, X_p by removing variables and considering reduced predictors. Typically, for any group of variables $\mathbf{X}_S = (X_i)_{i \in S}$, with any subset $S \subseteq [p]$ and reference distribution $Q_{S,\mathbf{x}}$, the reduced predictor is defined as:

$$f_S(\mathbf{x}_S) \triangleq \mathbb{E}_{Q_{S,\mathbf{x}}} [f(\mathbf{x}_S, \mathbf{X}_{\bar{S}})], \quad (2.1)$$

where $Q_{S,\mathbf{x}}$ represents the conditional distribution $\mathbf{X}_{\bar{S}} | \mathbf{X}_S = \mathbf{x}_S$. Other SV can be defined with marginal probabilities, but their interpretation is different [Heskes, 2020; Janzing, 2020; Chen, 2020]. There are still active debates on using or not conditional probabilities [Frye, 2020]. This work focuses only on conditional SV, as estimating them poses significant challenges. The SV for explaining the prediction $f(\mathbf{x})$ have been introduced in [Lundberg, 2017a] and are based on a cooperative game with value function $v(S) \triangleq f_S(\mathbf{x}_S)$. For any group of variables $C \subseteq [p]$ and $k \in \llbracket 1, p - |C| \rrbracket$, we denote the set $\mathcal{S}_k(C) = \{S \subseteq [p] \setminus C : |S| = k\}$ and we introduce a straightforward generalization of the SV for coalition C as

$$\phi_{\mathbf{x}_C}(f) \triangleq \frac{1}{p - |C| + 1} \sum_{k=0}^{p-|C|} \frac{1}{\binom{p-|C|}{k}} \sum_{S \in \mathcal{S}_k(C)} [f_{S \cup C}(\mathbf{x}_{S \cup C}) - f_S(\mathbf{x}_S)]. \quad (2.2)$$

This definition of the Shapley Value serves as a generalization of the classical SV for a single variable. By considering the singleton $C = \{i\}$ for $i \in [p]$, we retrieve the standard definition for

feature-value $X_i = x_i$. In the following Section, we show how this definition arises naturally when measuring the impact of a group of variables $\mathbf{X}_C = \mathbf{x}_C$, particularly in the case of categorical variables.

We propose solutions for computing and estimating the Shapley Values (SV). We focus solely on tree-based models due to their reduced computational cost and easier statistical handling. We demonstrate that the current state-of-the-art algorithm for tree-based models, TreeSHAP [Lundberg, 2020b], is highly biased when features are dependent. Consequently, we introduce statistically principled estimators to improve the estimation of the SV. Furthermore, we address the theoretical computation of SV for categorical variables when using standard encodings, which motivates the use of Equation (2.2). Specifically, we show that the true SV of a categorical variable is different from the sum of SVs of encoded variables, as generally used. Moreover, using the sum of encoded variables as the SV of a categorical variable provides incorrect estimates of all SVs in the model and leads to spurious interpretations. This is currently the only way to handle categorical variables with TreeSHAP. Therefore, we highlight the correct approach for computing the SV of encoded categorical variables and implement it using our estimators. Our contributions, which reduce bias in the estimation of SV, are implemented in a [Python package](#).

The chapter is structured as follows. In the next Section, we derive invariance principles for SV under reparametrization or encoding, which is particularly useful for dealing with categorical variables. In Section 3, we introduce two estimators of reduced predictors and SV. In Section 4, we highlight the improvement over dependent TreeSHAP. Finally, we discuss the reliability of SV in providing local explanations.

2 Coalition and Invariance for Shapley Values

In this Section, we present a unifying property of invariance for the Shapley Values of continuous and categorical variables. The property states that the explanation provided by a variable should not depend on the way it is encoded in a model. This invariance property provides a natural way to calculate the SV of categorical variables based on the notion of coalition and the general definition given in Equation (2.2). This is especially useful in our case, as we are also interested in the discretization of continuous variables to facilitate the estimation of Shapley Values and enhance their stability, which we will discuss in Section 3.

2.1 Invariance under reparametrization for continuous variables

In Equation (2.2), there is no restriction on the dimension of X_i . We assume that the p variables are vector-valued, i.e., $X_i \in \mathbb{R}^{p_i}$ where $p_i \geq 1$. We further assume that each variable X_i is transformed with a diffeomorphism $\varphi_i : \mathbb{R}^{p_i} \rightarrow \mathbb{R}^{p_i}$. We introduce the transformed variables $U_i \triangleq \varphi_i(X_i)$ and the reparametrized model \tilde{f} defined by $\tilde{f}(U_1, \dots, U_p) = f(X_1, \dots, X_p)$, i.e., $\tilde{f}(U_1, \dots, U_p) = f \circ \varphi^{(-1)}(\mathbf{U})$ where $\varphi = (\varphi_1, \dots, \varphi_p)$. Generally, we cannot relate the predictor learned from the real dataset $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ to the predictor learned from $\{(\mathbf{U}_i, Y_i)\}_{i=1}^n$ where Y is the target to predict. Estimation procedures are not invariant with respect to

reparametrization, which means we obtain different predictors after "diffeomorphic feature engineering". Consequently, we focus only on the impact of reparametrization on explanations, and we show below that the Shapley Values are invariant under reparametrization.

Proposition 2.1. *Let f and $\tilde{f} = f \circ \varphi^{(-1)}$ its reparametrization, then we have for all $i \in [p]$, and $\mathbf{u} = \varphi(\mathbf{x})$:*

$$\phi_{x_i}(f) = \phi_{u_i}(\tilde{f}).$$

See the proof in the Appendix (1.1). This identity indicates that the information provided by each feature X_i in the explanation is independent of any encoding, as mentioned by [Owen, 2017; Covert, 2020c]. The SV depends primarily on the dependence structure of the features. Therefore, the Shapley Value of a feature X_i remains the same after diffeomorphic transformation φ , we have $\phi_{x_i}(f) = \phi_{u_i}(\tilde{f})$. Suppose a variable X_i is separated into C correlated variables $\tilde{X}_C = (\tilde{X}_i)_{i \in C}$, for instance, by discretizing the variable. As X_i and \tilde{X}_C carry the same information, we may ask whether the SV of the group of features \tilde{X}_C is equal to the SV of X_i . In the next Section, we provide an affirmative answer to this question, which allows for correct computation of the SV of categorical variables after encoding.

2.2 Invariance for encoded categorical variable

In the remainder of the chapter, we use X to denote continuous predictive variables, Z to denote categorical variables and Y to denote the output of interest. While there are numerous encodings for a categorical variable Z with modalities $\{1, \dots, K\}$, we focus on two popular methods: One-Hot Encoding (OHE) and Dummy Encoding (DE). These methods introduce indicator variables Z_k such that $Z_k = 1$ if $Z = k$, 0 otherwise. In contrast to the continuous case, introducing indicator variables changes the number of "players" in the game defined for computing the Shapley Value. This change has significant consequences for calculating the SV for all variables in the model. Hence, the widely adopted practice of summing the SV of indicator variables Z_k to compute the SV of Z is generally not justified and false. To benefit from a similar invariance result as Proposition 2.1, we need to deal with the coalition of indicators and use the general expression of SV introduced in Equation (2.2). For simplicity, we assume that the model f uses only the two variables (X, Z) , where $X \in \mathbb{R}$ and $Z \in \{1, \dots, K\}$ is a categorical variable. The efficiency property of SV for a given prediction $f(x, z)$ gives the decomposition

$$f(x, z) - \mathbb{E}_P[f(X, Z)] = \phi_x(f) + \phi_z(f), \quad (2.3)$$

where P denotes the law of (X, Z) . To establish the link between the SV of the indicator variables Z_k and the SV of the variable Z , we introduce additional notations. We focus on the Dummy Encoding (DE) $\varphi : z \mapsto (z_1, \dots, z_{K-1})$ without loss of generality. The variables $(X, Z_{1:K-1})$ are defined on $\mathbb{R} \times \{0, 1\}^{K-1}$, and their distribution \tilde{P} is the image probability of P induced by transformation φ . The initial predictor $f : \mathbb{R} \times \{1, \dots, K\} \rightarrow \mathbb{R}$ is reparametrized as

a function $\tilde{f} : \mathbb{R} \times \{0, 1\}^{K-1} \rightarrow \mathbb{R}$ such that $f(X, Z) \triangleq \tilde{f}(X, Z_1, \dots, Z_{K-1})$. The function \tilde{f} is not completely defined for all $(z_1, \dots, z_{K-1}) \in \{0, 1\}^{K-1}$ and is only defined \tilde{P} -almost everywhere due to the deterministic dependence $\sum_{k=1}^{K-1} Z_k \leq 1$. Consequently, we need to extend \tilde{f} to the whole space $\mathcal{X} \times \{0, 1\}^{K-1}$ by setting $\tilde{f}(X, Z_1, \dots, Z_{K-1}) = 0$ as soon as $\sum_{k=1}^{K-1} Z_k > 1$. For the prediction $\tilde{f}(x, z_1, \dots, z_{K-1})$, we can compute the SV of x, z_1, \dots, z_{K-1} and obtain the following decomposition thank to the efficiency property

$$\tilde{f}(x, z_1, \dots, z_{K-1}) - \mathbb{E}_{\tilde{P}} \left[\tilde{f}(X, Z_1, \dots, Z_{K-1}) \right] = \phi_x(\tilde{f}) + \sum_{k=1}^{K-1} \phi_{z_k}(\tilde{f}) \quad (2.4)$$

where $\phi_{z_k}(\tilde{f})$ are the SV of the variable z_k computed with distribution \tilde{P} and model \tilde{f} . As $f(x, z) - \mathbb{E}_P[f(X, Z)] = \tilde{f}(x, z_1, \dots, z_{K-1}) - \mathbb{E}_{\tilde{P}}[\tilde{f}(X, Z_1, \dots, Z_{K-1})]$, we have

$$\phi_x(f) + \phi_z(f) = \phi_x(\tilde{f}) + \sum_{k=1}^{K-1} \phi_{z_k}(\tilde{f}). \quad (2.5)$$

In general, we have $\phi_z(f) \neq \sum_{k=1}^{K-1} \phi_{z_k}(\tilde{f})$, because the SV depends on the number of variables and they are not calculated using the same quantities. In the next proposition we show that $\phi_z(f) = \phi_{z_{1:K-1}}(\tilde{f})$, where $\phi_{z_{1:K-1}}(\tilde{f})$ is computed with Equation (2.2) and corresponds to the Shapley Values of the coalition of variables (z_1, \dots, z_{K-1}) .

Proposition 2.2. *Given a predictor $f : \mathbb{R} \times \{1, \dots, K\} \rightarrow \mathbb{R}$ and its reparametrization \tilde{f} using Dummy Encoding $\tilde{f} : \mathbb{R} \times \{0, 1\}^{K-1} \rightarrow \mathbb{R}$ such that $f(x, z) = \tilde{f}(x, z_1, \dots, z_{K-1})$, we have*

$$\begin{cases} \phi_{z_{1:K-1}}(\tilde{f}) & = \phi_z(f) \\ \phi_x(\tilde{f}; z_{1:K-1}) & = \phi_x(f), \end{cases} \quad (2.6)$$

where $\phi_x(\tilde{f}; z_{1:K-1})$ is the SV of x when the variables (z_1, \dots, z_{K-1}) are considered as a single variable. We refer to Supplementary Material (2.1) for detailed derivations. In general, for cooperative games, the SV of a coalition $\phi_{z_C}(\tilde{f})$ with $C \subseteq \{1, \dots, K-1\}$ is different from the sum of individual SV $\sum_{k \in C} \phi_{z_k}(\tilde{f})$. We note that we can compute two different SV for X when we use the reparametrized predictor \tilde{f} : $\phi_x(\tilde{f})$ and $\phi_x(\tilde{f}; z_{1:K-1})$. These two SV are different in general, as they involve different numbers of variables and different conditional expectations. Proposition 2.2 shows that we should prefer $\phi_x(\tilde{f}; z_{1:k})$ as it is equal to the SV of x in the original model $\phi_x(f)$.

2.3 Coalition or Sum: numerical comparisons

We give numerical examples that illustrate the differences between the use of coalition or sum to calculate the SV for categorical variables. We consider a linear predictor f , with categorical Z and 3 continuous variables $\mathbf{X} = (X_1, X_2, X_3)$, defined as $f(\mathbf{X}, Z) = B_Z \mathbf{X}$ with $B_Z \in \mathbb{R}^3$, $\mathbf{X}|Z = z \sim \mathcal{N}(\mu_z, \Sigma_z)$ and $\mathbb{P}(Z = z) = \pi_z$, $Z \in \{a, b, c\}$. The values of the parameters used can be found in the Supplementary Material (6.1). In Figure 2.1, we remark that the SV change

dramatically with respect to the encoding when we use the sum of the indicator variables as the SV of the categorical variable. The sign changes given the encoding (DE or OHE) and is often different from the sign of the true SV of Z without encoding. We also note important differences in the SV of the quantitative variables \mathbf{X} .

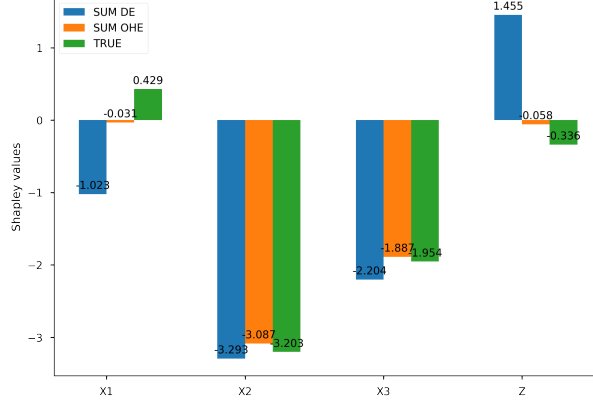


Figure 2.1: SV with and without encoding (OHE - DE) for a given observation $(X_1, X_2, X_3) = [0.35, -1.61, -0.11]$, $Z = a$

To quantify the global difference of the different methods, we compute the relative absolute error (R-AE) of the SV of each observation using DE or OHE encoding in comparison to the true SV without encoding, which is defined as:

$$\text{R-AE}(f, \tilde{f}) = \sum_{i=1}^p \frac{|\phi_{x_i}(f) - \phi_{x_i}(\tilde{f})|}{|\phi_{x_i}(f)|}. \quad (2.7)$$

We compute the SV of 1000 new observations. We observe in Figure 2.2 that the differences can be huge for almost all samples (DE is much worse than OHE in this example). Thus, we highly recommend using the coalition as it is consistent with the true SV contrary to the sum. More examples with real datasets can be found in the Supplementary Material (10).

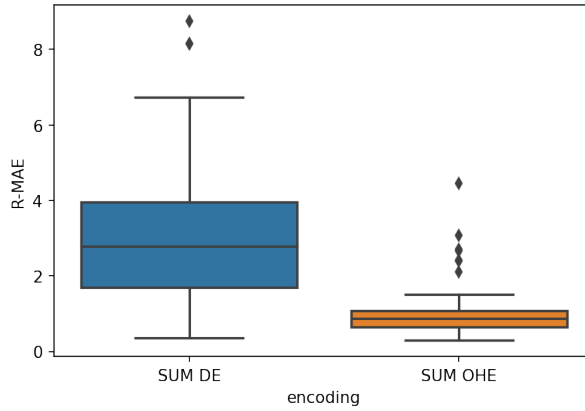


Figure 2.2: R-AE distribution between the SV with and without encoding (OHE - DE).

3 Shapley Values for tree-based models

The computation of Shapley Values (SV) faces two main challenges: the combinatorial explosion with 2^p coalitions to consider and the estimation of the conditional expectation $f_S(\mathbf{x}_S) = \mathbb{E}[f(\mathbf{X})|\mathbf{X}_S = \mathbf{x}_S]$ for any $S \subseteq [p]$. Current approaches rely on several approximations and sampling procedures that assume independence of the features [Lundberg, 2017a; Covert, 2020a]. More recently, some methods propose to model the joint distribution of features with a gaussian distribution or vine copula to draw samples from the conditional distributions [Aas, 2020; Aas, 2021]. Other methods, such as [Williamson, 2020], train one model for each selected subset S of variables, which is accurate but computationally costly. However, their final objective differs from ours since we are interested in local importances and exact computations, i.e., no sampling of the subsets. To achieve this, we focus on tree-based models, as exploited by [Lundberg, 2020b] for deriving an algorithm (TreeSHAP) for exact computation of SV, where we can compute all the terms (no sampling of the subsets $S \subseteq [p]$) and the estimation of the conditional expectations is simplified. After briefly presenting the limitations of TreeSHAP, we introduce two new estimators that use the tree structure. For simplicity, we consider a single tree and not an ensemble of trees (Random Forests, Gradient Tree Boosting, etc.), as extending our estimators to these more complex models is straightforward through linearity.

3.1 Algorithms for computing Conditional Expectations and the Tree SHAP algorithm

We consider a tree-based model f defined on \mathbb{R}^p (categorical variables are one-hot encoded). We have $f(\mathbf{x}) = \sum_{m=1}^M f_m \mathbb{1}_{L_m}(\mathbf{x})$ where L_m represents a leaf. The leaves form a partition of the input space, and each leaf can be written as $L_m = \prod_{i=1}^p [a_i^m, b_i^m]$ with $-\infty \leq a_i^m < b_i^m \leq +\infty$. Alternatively, we write the leaf with the decision path perspective: a leaf L_m is defined by a sequence of decision based on d_m variables $X_{N_k^m}$, $k = 1, \dots, d_m$, $N_k^m \in \{1, \dots, p\}$. For each node k in the path of the leaf L_m , $X_{N_k^m}$ is the variable used to split, and the region I_k^m defined by the split value t_k^m is either $] -\infty, t_k^m]$ or $]t_k^m, +\infty[$. The leaf can be rewritten as

$$L_m = \left\{ \mathbf{x} \in \mathbb{R}^p : x_{N_1^m} \in I_1^m, \dots, x_{N_{d_m}^m} \in I_{d_m}^m \right\}. \quad (2.8)$$

The crucial point is to identify the set of leaves compatible with the condition $\mathbf{X}_S = \mathbf{x}_S$. We can partition the leaf according to a coalition S as $L_m = L_m^S \times L_m^{\bar{S}}$ with $L_m^S = \prod_{i \in S} [a_i^m, b_i^m]$ and $L_m^{\bar{S}} = \prod_{i \in \bar{S}} [a_i^m, b_i^m]$. Thus, for each condition $\mathbf{X}_S = \mathbf{x}_S$ the set of compatible leaves of $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_{\bar{S}})$ is

$$C(S, \mathbf{x}) = \left\{ m \in [1 \dots M] : \mathbf{x}_S \in L_m^S \right\} = \left\{ m \in [1 \dots M] : x_{N_i^m} \in I_i^m, N_i^m \in S \right\}$$

and the reduced predictor $f_S(\mathbf{x}_S)$ has the simple expression

$$f_S(\mathbf{x}_S) = \sum_{m \in C(S, \mathbf{x})} f_m \mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S = \mathbf{x}_S)$$

where the probability is computed under the law of the features $P_{\mathbf{X}}$. If we have a model for $P_{\mathbf{X}}$, we can derive the conditional law and directly evaluate the conditional probabilities. For instance, when $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, we can exactly compute the conditional probabilities $\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S = \mathbf{x}_S) = \mathbb{P}_{P_{\mathbf{X}}}(\prod_{k=1}^{d_m} I_k^m | \mathbf{X}_S = \mathbf{x}_S)$. In general, deriving conditional probabilities can be challenging, but assumptions about the factorization of the distribution can accelerate the computation. In [Lundberg, 2018; Lundberg, 2020b], the authors introduce a recursive algorithm (TreeSHAP with path-dependent feature perturbation, Algorithm 1) that assumes that the probabilities for every compatible leaf L_m can be factored with the decision tree, which simplifies the computation as

$$\mathbb{P}_{P_{\mathbf{X}}^{SHAP}} \left(\prod_{k=1}^{d_m} I_k^m \mid \mathbf{X}_S = \mathbf{x}_S \right) = \delta_S(N_1^m) \times \prod_{i=2: N_i^m \notin S}^{d_m} \mathbb{P} \left(X_{N_i^m} \in I_i^m \mid \prod_{k=1}^i X_{N_{k-1}^m} \in I_{k-1}^m \right), \quad (2.9)$$

with $\delta_S(N_1) = \mathbb{P}(X_{N_1^m} \in I_1^m)$ if $N_1^m \notin S$, and 1 otherwise. The underlying assumption in Eq. (2.9) is that we have a Markov property defined by the path of the tree, see the algorithm description in the Supplementary Material (4). However, as we will demonstrate in our simulations, this assumption is too strong and leads to a high estimation bias. We denote \hat{f}_S^{SHAP} and $\phi_{x_i}(\hat{f}_S^{SHAP})$ as the estimators of the reduced predictor and Shapley Values using the TreeSHAP factorization. Therefore, we propose two estimators that do not rely on the factorization of $P_{\mathbf{X}}$.

3.2 Statistical Estimation of Conditional Expectations

Discrete case. To address the statistical problem of estimating the probabilities $\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S = \mathbf{x}_S)$ from a given dataset $\mathcal{D} = \{\mathbf{X}_i\}_{i=1}^n$, where $\mathbf{X}_i \sim P_{\mathbf{X}}$, without assuming any density or prior knowledge about $P_{\mathbf{X}}$ as in [Aas, 2020; Aas, 2021], we first consider the case where all variables are discrete. This allows us to estimate $\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S = \mathbf{x}_S)$ directly. A straightforward estimation is based on $N(\mathbf{x}_S)$, which is the number of observations in \mathcal{D} such that $\mathbf{X}_S = \mathbf{x}_S$, and $N(L_m, \mathbf{x}_S)$, which is the number of observations in leaf L_m of \mathcal{D} that satisfy the condition $\mathbf{X}_S = \mathbf{x}_S$. A consistent estimation of the conditional probability $\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S = \mathbf{x}_S)$ can be obtained by computing the ratio of these two terms as

$$\mathbb{P}_{\hat{P}_{\mathbf{X}}^{(D)}}(L_m | \mathbf{X}_S = \mathbf{x}_S) = \frac{N(L_m, \mathbf{x}_S)}{N(\mathbf{x}_S)}. \quad (2.10)$$

Estimating the conditional probabilities $\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S = \mathbf{x}_S)$ becomes more challenging when the variables \mathbf{X}_S are continuous. A common approach is to use kernel smoothing estimators [Nadaraya, 1964]. However, this method has several drawbacks, such as a low convergence rate in high dimensions and the need to derive and select appropriate bandwidths, which can add complexity and instability to the estimation procedure. To address these issues, we propose a simple approach based on quantile-discretization of the continuous variables. This technique is commonly used to facilitate model explainability, particularly in tree-based models, as shown in [Bénard, 2021d]. Binning observations can also help stabilize the reduced predictors and Shapley Values, thus improving the robustness of the explanation [Alvarez-Melis, 2018].

In our experiments, we use a simple approach to discretize continuous variables into q quantiles, where each feature X_i is encoded with indicator variables $X_i^{(r)}$, $r \in \{1, \dots, q\}$. Let $\hat{q}_i^{(r)}$ denote the empirical $(\frac{r}{q})$ -th quantile of feature X_i using dataset \mathcal{D} , and let $\hat{q}_i^{(0)} = -\infty$ and $\hat{q}_i^{(q)} = +\infty$. We define $X_i^{(r)} = 1$ if X_i falls in the interval $[\hat{q}_i^{(r-1)}, \hat{q}_i^{(r)})$. To compute the Shapley Values of a given feature X_i , we use the coalition of its indicator variables $(X_i^{(1)}, \dots, X_i^{(q)})$ as defined in Proposition 2.2. Then, we define the Discrete reduced predictor denoted by $\hat{f}_S^D(\mathbf{x}_S)$ as

$$\hat{f}_S^D(\mathbf{x}_S) = \sum_{m \in C(S, \mathbf{x})} f_m \mathbb{P}_{\hat{P}_X^{(D)}}(L_m | \mathbf{X}_S = \mathbf{x}_S), \quad (2.11)$$

and the corresponding estimator of the SV is denoted $\phi_{x_i}(\hat{f}^D)$. Although the discretization of continuous variables leads to some loss of information, it is often negligible in terms of performance when using tree-based models, as shown in the Supplementary Material (5.1). With only $q = 10$ quantiles, the input space is divided into a fine grid of p^{10} cells, which provides a rich representation of the data. However, the computation of Shapley Values (SV) remains exponential with respect to the number of variables using this estimator. Therefore, we propose an alternative estimator that leverages the information from the decision tree's leaves, allowing for faster SV computation.

Continuous and mixed-case. Instead of discretizing the variables, we use the leaves of the estimated trees. Essentially, we replace the condition $\{\mathbf{X}_S = \mathbf{x}_S\}$ by $\{\mathbf{X}_S \in L_m^S\}$. This change introduces bias, but aims to improve the variance during estimation. We introduce the Leaf-based estimator as

$$\hat{f}_S^{(Leaf)}(\mathbf{x}_S) = \frac{1}{Z(S, \mathbf{x})} \sum_{m \in C(S, \mathbf{x})} f_m \mathbb{P}_{\hat{P}_X^{(Leaf)}}(L_m | \mathbf{X}_S \in L_m^S) \quad (2.12)$$

where $\mathbb{P}_{\hat{P}_X^{(Leaf)}}(L_m | \mathbf{X}_S \in L_m^S)$ is an estimate of the conditional probability, and $Z(S, \mathbf{x})$ is a normalizing constant. The definition of every probability estimate is

$$\mathbb{P}_{\hat{P}_X^{(Leaf)}}(L_m | \mathbf{X}_S \in L_m^S) = \frac{N(L_m)}{N(L_m^S)}$$

where $N(L_m)$ is the number of observations of \mathcal{D} in the leaf L_m , and $N(L_m^S)$ is the number of observations of \mathcal{D} satisfying the conditions $\mathbf{x}_S \in L_m^S$. Another interpretation of this estimator is that it projects the partition of the tree along the direction defined by the variables \mathbf{X}_S . This results in a projected tree that only considers variables \mathbf{X}_S , which is then used to estimate the conditional probability $\mathbb{E}[f(\mathbf{X}) | \mathbf{X}_S = \mathbf{x}_S]$. It is important to note that the probability estimates do not necessarily sum up to one as we are not conditioning on the same event, i.e.,

$$\sum_{m \in C(S, \mathbf{x})} \mathbb{P}_{\hat{P}_X^{(Leaf)}}(L_m | \mathbf{X}_S \in L_m^S) \neq 1.$$

Therefore, we introduce a normalizing constant to ensure that the probabilities are correctly normalized. This normalizing constant is defined as $Z(S, \mathbf{x}) = \sum_{m \in C(S, \mathbf{x})} \frac{N(L_m)}{N(L_m^S)}$. The Leaf-

based reduced predictor (2.12) can be computed for continuous and categorical variables, and hence we can compare it with $\hat{f}_S^{(D)}$ in order to evaluate its bias. In both cases, the main challenge is to compute $C(S, \mathbf{x})$, for every coalition S . We show in the next Section how the computational complexity of the SV $\phi_{x_i}(\hat{f}^{(Leaf)})$ is drastically reduced using the Leaf estimator. Indeed, when we consider the leaf L_m , we only have to compute the SV for d_m variables and not for p variables.

Bias analysis. Before employing the two proposed estimators to calculate the Shapley Values, we first analyze the bias of these estimators. When the variables are discrete, it is obvious that the discrete estimator $\hat{f}_S^{(D)}$ is consistent. However, in the case of the Leaf estimator $\hat{f}_S^{(Leaf)}$, we analyze its bias with respect to the true reduced predictor $f_S(\mathbf{x}_S)$ below:

$$\begin{aligned} \hat{f}_S^{(Leaf)}(\mathbf{x}_S) - f_S(\mathbf{x}_S) &= \sum_{m \in C(S, \mathbf{x})} f_m \mathbb{P}_{\hat{P}_{\mathbf{X}}^{(Leaf)}}(L_m | \mathbf{X}_S \in L_m^S) - \sum_{m \in C(S, \mathbf{x})} f_m \mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S = \mathbf{x}_S) \\ &= \sum_{m \in C(S, \mathbf{x})} f_m \left[\mathbb{P}_{\hat{P}_{\mathbf{X}}^{(Leaf)}}(L_m | \mathbf{X}_S \in L_m^S) - \mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S \in L_m^S) \right] \\ &\quad + \sum_{m \in C(S, \mathbf{x})} f_m \left[\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S \in L_m^S) - \mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S = \mathbf{x}_S) \right] \end{aligned}$$

The control of the blue term is well established, and its rate of convergence is known. Recently, [Margot, 2021] (Proposition 3.2) inspired by [Grunewalder, 2018] (Proposition 3.2) shows that if $\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_{i,S} \in L_m^S} \geq n^{-\alpha}$, with $\alpha \in [0, 1/2)$, then $|\mathbb{P}_{\hat{P}_{\mathbf{X}}^{(Leaf)}}(L_m | \mathbf{X}_S \in L_m^S) - \mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S \in L_m^S)| = \mathcal{O}_P(n^{\alpha-1/2})$.

The second term depends on the quality of the partition obtained from the tree. The intuition behind the effectiveness of tree-based models is that they group observations with similar conditional laws in each cell. Indeed, one of the assumptions to prove the consistency of tree-based models is that the variation of the conditional law is zero in each leaf, i.e., for all $\mathbf{x} \in L_m$ and $r \in \mathbb{R}$, we have $\sup_{\mathbf{z} \in L_m} |F(r|\mathbf{z}) - F(r|\mathbf{x})| \xrightarrow{a.s} 0$ [Scornet, 2015; Meinshausen, 2006; Elie-Dit-Cosaque, 2022], or alternatively, prove that the diameter of the leaves tends to 0 [Györfi, 2002]. The latter ensures that the probability $\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S = \mathbf{x}_S)$ varies slightly as we move within a given cell if \mathbf{X} admits a continuous density. Therefore, if the diameter of the leaves tends to 0 as generally assumed for partition-based estimator [Györfi, 2002], the leaf estimator is consistent.

3.3 Fast Algorithm for Shapley Values with the Leaf estimator

Here, we focus on the computational efficiency offered by the Leaf estimator. It is well-known that the computation of the Shapley Values has exponential complexity, as we need to compute 2^p different coalitions for each observation. However, with the Leaf estimator $\hat{f}_S^{(Leaf)}$, we can reduce the complexity to being exponential in the depth of the tree D in the worst case, instead of being exponential in the total number of variables p . This is very interesting, as the depth of the tree is rarely above 10 in practice, while p can be very large, spanning different orders of magnitude. The idea is to split the original game into the sum of smaller games, as described by the following proposition.

Proposition 3.1. Consider a tree-based model $f(\mathbf{x}) = \sum_{m=1}^M f_m \mathbb{1}_{L_m}(\mathbf{x})$, and let S_m be the set of variables used along the path of leaf L_m . For any observation \mathbf{x} and variable-value $X_i = x_i$, we can decompose its Shapley value $\phi_{x_i}(f^{(Leaf)})$ into the sum of cooperative games defined on each compatible leaf of \mathbf{x} , $C(\mathbf{x}) = \{m \in \{1, \dots, M\} : \exists i \in [p], x_i \in [a_i^m, b_i^m]\}$ as follows

$$\phi_{x_i}(f^{(Leaf)}) = \sum_{m \in C(\mathbf{x})} \phi_{x_i}^m(f^{(Leaf)}) \quad (2.13)$$

where $\phi_{x_i}^m(f^{(Leaf)})$ is a reweighted version of the Shapley Value of a cooperative game with players S_m and value function $v(f^{(Leaf)}, S) = \mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S \in L_m^S) \mathbb{1}_{L_m^S}(\mathbf{x}_S)$.

Proof. In the proposition, we consider the asymptotic version $\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S \in L_m^S)$ of the leaf estimator $\mathbb{P}_{\hat{P}_{\mathbf{X}}^{(Leaf)}}(L_m | \mathbf{X}_S \in L_m^S)$, but the result is also true for the leaf estimator. By definition, we have for the variable-value $X_i = x_i$,

$$\begin{aligned} & \phi_{x_i}(f^{(Leaf)}) \\ &= \frac{1}{p} \sum_{S \subseteq [p] \setminus \{i\}} \binom{p-1}{|S|}^{-1} \left(f_{S \cup i}^{(Leaf)}(\mathbf{x}_{S \cup i}) - f_S^{(Leaf)}(\mathbf{x}_S) \right) \\ &= \frac{1}{p} \sum_{S \subseteq [p] \setminus \{i\}} \binom{p-1}{|S|}^{-1} \left(\sum_{m \in C(\mathbf{x})} f_m \left[\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_{S \cup i} \in L_m^{S \cup i}) \mathbb{1}_{L_m^{S \cup i}}(\mathbf{x}_{S \cup i}) - \mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S \in L_m^S) \mathbb{1}_{L_m^S}(\mathbf{x}_S) \right] \right) \\ &= \frac{1}{p} \sum_{m \in C(\mathbf{x})} \sum_{S' \subseteq S_m \setminus \{i\}} \left[\binom{p-1}{|S'|}^{-1} f_m \left[\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_{S' \cup i} \in L_m^{S' \cup i}) \mathbb{1}_{L_m^{S' \cup i}}(\mathbf{x}_{S' \cup i}) - \mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_{S'} \in L_m^{S'}) \mathbb{1}_{L_m^{S'}}(\mathbf{x}_{S'}) \right] \right. \\ & \quad + \sum_{Z \neq \emptyset, Z \subseteq \bar{S}_m \cup i} \binom{p-1}{|Z| + |S'|}^{-1} f_m \left[\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_{S' \cup Z \cup i} \in L_m^{S' \cup Z \cup i}) \mathbb{1}_{L_m^{S' \cup Z \cup i}}(\mathbf{x}_{S' \cup Z \cup i}) \right. \\ & \quad \left. \left. - \mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_{S' \cup Z} \in L_m^{S' \cup Z}) \mathbb{1}_{L_m^{S' \cup Z}}(\mathbf{x}_{S' \cup Z}) \right] \right]. \end{aligned}$$

If $Z \subseteq \bar{S}_m$ and $S \subseteq S_m$, we have

$$\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_{Z \cup S} \in L_m^{Z \cup S}) = \mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_S \in L_m^S). \quad (2.14)$$

Therefore, ϕ_{x_i} can be rewrite as follows:

$$\begin{aligned} & \phi_{x_i}(f^{(Leaf)}) \\ &= \frac{1}{p} \sum_{m \in C(\mathbf{x})} \sum_{S' \subseteq S_m \setminus \{i\}} \left[\binom{p-1}{|S'|}^{-1} + \sum_{Z \neq \emptyset, Z \subseteq \bar{S}_m \cup i} \binom{p-1}{|Z| + |S'|}^{-1} \right] \\ & \quad \times f_m \left[\mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_{S' \cup i} \in L_m^{S' \cup i}) \mathbb{1}_{L_m^{S' \cup i}}(\mathbf{x}_{S' \cup i}) - \mathbb{P}_{P_{\mathbf{X}}}(L_m | \mathbf{X}_{S'} \in L_m^{S'}) \mathbb{1}_{L_m^{S'}}(\mathbf{x}_{S'}) \right] \\ &= \sum_{m \in C(\mathbf{x})} \phi_{x_i}^m(f^{(Leaf)}) \end{aligned}$$

□

4 Comparison of the estimators

To compare the different estimators, we need a model where conditional expectations can be calculated exactly. If $X \sim \mathcal{N}(\mu, \Sigma)$ then $X_{\bar{S}}|X_S$ is also a multivariate Gaussian with explicit mean vector and covariance matrix. We do not include any comparisons with KernelSHAP as our main goal is to improve upon TreeSHAP which is the state-of-the-art for tree-based models. In addition, most implementations of KernelSHAP are based on the marginal distribution, as its aim is to be model-agnostic.

Experiment 1. In the first experiment, we consider a dataset $\mathcal{D}_n = \{\mathbf{X}_i, Y_i\}_{i=1}^n$ with $n = 10^4$ generated by a linear regression model with $\mathbf{X} \in \mathbb{R}^p$ following a multivariate Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix $\Sigma = \rho J_p + (\rho - 1)I_p$, where $p = 5$, $\rho = 0.7$, J_p is the all-ones matrix and I_p is the identity matrix. The response variable is $Y = B^t \mathbf{X}$, where $B \in \mathbb{R}^p$. We trained a Random Forest f on \mathcal{D}_n and obtained an MSE of 4.28. The detailed parameters can be found in the Supplementary Material (6.2). Since the law of \mathbf{X} is known, we can compute the exact Shapley Values (SV) of f using a Monte Carlo estimator.

We aim to compare the true Shapley Value $\phi_{x_i}(f)$ with the Shapley Value estimated by the different methods $\phi_{x_i}(\hat{f}^{(method)})$, where the *method* can be SHAP, Leaf, or D. To quantify the differences between these estimators, we consider two evaluation metrics. First, we compute the Relative Absolute Error (R-AE) as defined in Equation (2.7). Second, we measure the True Positive Rate (TPR) to assess whether the ranking of the top $k = 3$ highest and lowest Shapley Values is preserved across different estimators.

In Figure 2.3a, we compute the SV $\phi_{x_i}(\hat{f}^{SHAP})$, $\phi_{x_i}(\hat{f}^{Leaf})$ on a new dataset of size 1000 generated by the synthetic model. We observe that the estimator \hat{f}^{Leaf} is more accurate than TreeSHAP \hat{f}^{SHAP} by a large margin. TreeSHAP has an average R-AE= 3.31 and TPR= 86% ($\pm 17\%$) while Leaf estimator gets R-AE= 0.90 and TPR= 94% ($\pm 12\%$).

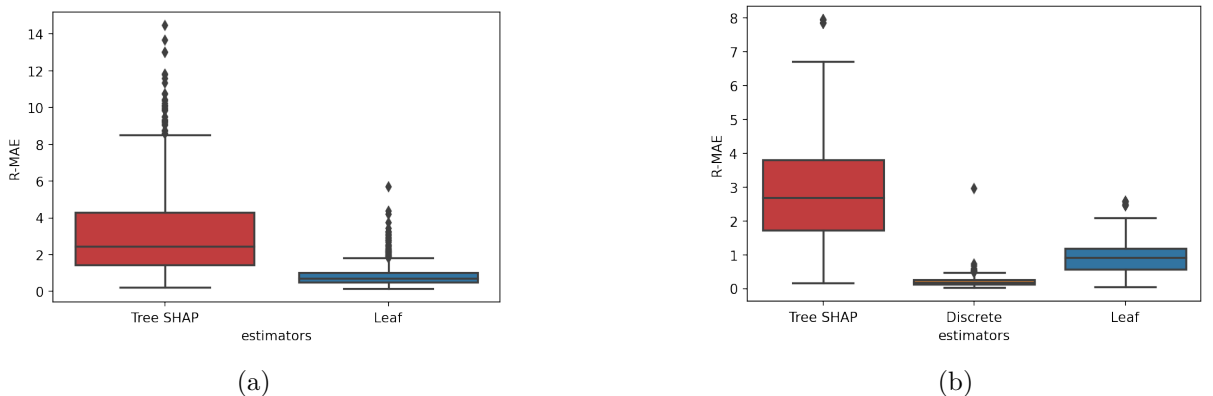


Figure 2.3: R-AE on 1000 new observations sampled from the synthetic model, $p=5$ using continuous variables (a) and discretized variables (b).

In Figure 2.3b, we compare the SV of the Discrete unbiased estimator $\phi_{x_i}(\hat{f}^{(D)})$, TreeSHAP $\phi_{x_i}(\hat{f}^{SHAP})$ and Leaf estimator $\phi_{x_i}(\hat{f}^{Leaf})$ with the True $\phi_{x_i}(f)$, where the model f was trained

on the discretized version of \mathcal{D}_n . As demonstrated in Figure 2.3b, the Discrete estimator also outperforms TreeSHAP by a significant margin.

Experiment 2. Here, we investigate the impact of feature dependence on the performance of the different estimators. We use the model of Experiment 1, but we vary the correlation coefficient ρ from 0 to 0.99, representing increasing positive correlations among the features. As shown in Figure 2.4, TreeSHAP performs well when the features are independent ($\rho = 0$), but it is outperformed by Leaf estimator as the dependence between the features increases.

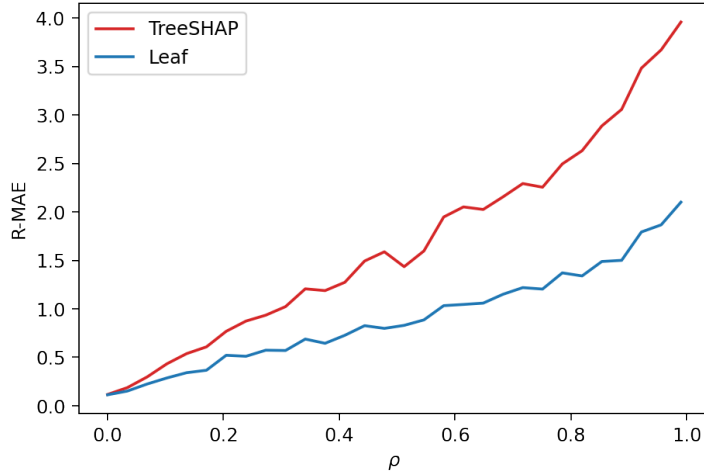


Figure 2.4: R-AE of the different estimators given the correlation coefficient $\rho \in [0, 0.99]$

Furthermore, we conduct a runtime comparison of computing SV with Leaf and TreeSHAP on three datasets with different shapes: Boston ($n = 506, p = 13$), Adults ($n = 32561, p = 12$) [Dua, 2017a], and a linear model ($n = 10000, p = 500$), where n is the number of observations and p is the number of variables. We train XGBoost with default parameters on these datasets and compute the SV of 1000 observations for Adults, the linear model, and 506 observations for Boston. As expected, Table 2.1 shows that TreeSHAP is much faster than the Leaf estimator. This difference in runtime can be partly explained by the Leaf estimator having to go through all the data for each leaf, whereas TreeSHAP uses the information stored in the trees. However, the Leaf estimator is not very affected by the dimension of the variables, as it succeeds in computing the SV when $p = 500$ in a reasonable time.

Table 2.1: Run-time of TreeSHAP and Leaf estimator on Adults (A), Boston (B) and the toy (T) datasets.

DATASETS	LEAF	TREE SHAP
A (P=12)	1 MIN 4 s \pm 1.73 s	3.33 s \pm 39.9 MS
B (P=13)	8.82 s \pm 204 MS	129 MS \pm 6.91 MS
T (P=500)	1MIN 5s \pm 1.73 s	101 MS \pm 4.54 MS

5 Discussion and Future works

We have demonstrated that the current implementation of Shapley Values can lead to unreliable explanations due to biased estimators or inappropriate handling of categorical variables. To address these issues, we have proposed new estimators and provided a correct method for handling categorical variables. Our results show that even in simple models, the difference between the state-of-the-art (TreeSHAP) and proposed methods can be significant. Despite growing interest in trustworthy AI, the impact of these inaccuracies in explanations is not well understood. One reason for this may be the difficulty in systematically and quantitatively evaluating the quality of an explanation, as it depends on the law of the data, which can be difficult to approximate. Furthermore, such analyses may be influenced by confirmation bias.

We also believe that the quality of the estimates is not the only drawback of SV. In fact, we demonstrate in Proposition 5.1 that SV explanations are not local explanations, but remain global, even in simple piecewise linear models.

Proposition 5.1. *Let us assume that we have $X \in \mathbb{R}^p$, $\mathbf{X} \in \mathcal{N}(0, I_p)$ independent Gaussian features, and a linear predictor f defined as:*

$$f(X) = (a_1X_1 + a_2X_2)\mathbb{1}_{X_5 \leq 0} + (a_3X_3 + a_4X_4)\mathbb{1}_{X_5 > 0}. \quad (2.15)$$

Even if we choose an observation \mathbf{x} such that $x_5 \leq 0$ and the predictor only uses x_1, x_2 , the SV of ϕ_{x_3}, ϕ_{x_4} is not necessarily zero. Indeed, for all $i \in \{3, 4\}$

$$\begin{aligned} \phi_{x_i} &= \frac{1}{p} \mathbb{P}(\mathbf{X}_5 > 0) \sum_{S \subseteq [p] \setminus \{i, 5\}} \binom{p-1}{|S|}^{-1} (a_i(\mathbf{x}_i - \mathbb{E}[\mathbf{X}_i])) \\ &= K (a_i(\mathbf{x}_i - \mathbb{E}[\mathbf{X}_i])) \quad , \end{aligned}$$

where K is a constant. The proof is in the Supplementary Material. Proposition 5.1 highlights that the SVs are not truly local measures, but rather have a global effect. This occurs because when calculating the SV for $X_3 = x_3$ or $X_4 = x_4$, we also consider subsets S that do not contain X_5 . By marginalizing and changing the sign of X_5 , we use the other linear model not used for the given observation. Such findings pose significant challenges in the interpretation of SV, and we believe that they are often overlooked due to the lack of precision and understanding of Shapley Values in practice.

Chapter 3

Please stop using SHAP and LIME and use Regional Explanations instead

Abstract

In this chapter, we criticize the two most popular local attribution methods, namely SHAP and LIME, which aim to quantify the contribution of a feature x_i to a specific prediction $f(x_1, \dots, x_p)$. We present arguments demonstrating the inability of these methods to detect the local important variables of a given prediction. Even in an ideal scenario, where there are no dependencies between the variables and the methods are computed exactly. Therefore, we propose an alternative approach called Regional Explanations that is between local and global explanations. Our method involves partitioning the input space into regions, wherein observations within the same region exhibit similar local importance measures. Subsequently, we employ global attribution methods within each region to determine the importance of each feature, thus establishing the contributions of the features of an observation based on its assigned group. The primary advantage of this approach is that it leverages the well-established understanding of global attribution measures to define local/regional attributions that have sound statistical properties.

Contents

1	Introduction	52
2	Stop using Local Shapley Values	53
3	Stop using LIME	56
4	From Global Explanations to Regional Explanations	58
5	Experiments	61
6	Discussion	65

1 Introduction

Machine learning models are widely recognized for their predictive power, but often lack transparency, making it difficult to understand the rationale behind their predictions. As these models are increasingly used in sensitive areas such as medicine and justice, there is a growing demand for tools that can elucidate the "why" behind their predictions. In response, the eXplainable AI (XAI) community has emerged to develop tools that help explain machine learning models.

These tools can be categorized into local and global methods. Local explanations aim to provide insights into individual predictions, while global explanations focus on understanding the overall behavior of a model across the entire input space. Popular local methods, such as SHAP [Lundberg, 2017b] and LIME [Ribeiro, 2016a], seek to create a local linear approximation of the model within the vicinity of a given instance. In contrast, global methods primarily consist of "leave-one-out" approaches [Lei, 2016; Rinaldo, 2019; Williamson, 2020; Covert, 2020c; Gan, 2022], which evaluate performance loss when a variable is removed.

This chapter focuses on local methods, particularly the widely-used SHAP and LIME, as they are widely used and have limited theoretical understanding. The ideas behind these methods are appealing, but the quantities they estimate are unclear. Apart from the linear or additive model [Bordt, 2023; Garreau, 2020], no work demonstrates what quantities these methods compute. The only theoretical study on LIME, by [Garreau, 2020], shows that in the case of a linear model, the LIME coefficients are proportional to the partial derivatives. However, it also reveals that the coefficient of important variables can vanish by simply changing a parameter of the method. In addition to their lack of theoretical understanding, we demonstrate in this work that these methods are ineffective in identifying important local variables, even in an ideal scenario. In our analysis, we consider the variables to be independent to eliminate any potential bias in detecting important variables when dependencies exist. We also assume no estimation bias, meaning that the limitations we raise persist even with perfect knowledge of the data distribution, the regression function and exact computation of the methods. Thus, the issues that we highlight stem from the inherent nature of these methods.

In contrast, most global methods are supported by the existing literature on feature importance [Breiman, 2001; Lei, 2016; Williamson, 2020] and global sensitivity analysis [Iooss, 2015], and are backed by strong consistency and inference results. Our ultimate goal is to leverage global methods to define local attributions for each individual while benefiting from the advantages of global methods. Essentially, we aim to find a partition of the input space where observations in each cell of the partition exhibit the same behavior concerning the local importance measure, and by associating each observation with the global importance measure conditional on the cell it belongs to, we can derive a local importance that possesses sound statistical properties.

Notations. Consider a dataset represented as $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, where $\mathbf{X}_i = (X_{i1}, \dots, X_{ip}) \in \mathcal{X} \subseteq \mathbb{R}^p$ denotes the input variables and $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ represents the output, and (\mathbf{X}_i, Y_i) are i.i.d. observations of $(\mathbf{X}, Y) \sim P = P_{\mathbf{X}}P_{Y|\mathbf{X}}$. We use $\mathbf{X}_S = (X_i)_{i \in S}$ to denote the subset of features, and $[p] = \{1, \dots, p\}$, and $\mathcal{P}(D)$ represents the power set of a set D .

2 Stop using Local Shapley Values

A cooperative game is a pair (D, v) , where $D = \{X_1, \dots, X_p\}$ represents a set of p players, and $v : \mathcal{P}(\{1, \dots, p\}) \rightarrow \mathbb{R}$ denotes a value function that assigns a value to every possible coalition of players, reflecting the worth of each group. Typically, the value function v is assumed to be positive and increasing monotonically, which means that if $A \subseteq B$, then $v(A) \leq v(B)$. Here, $v(A)$ represents the value of \mathbf{X}_A . A key concept in the definition of Shapley Values is the marginal contribution, denoted as $\Delta_i(S) = v(S \cup i) - v(S)$. The marginal contribution is the improvement of the value of a coalition S when a given player i is added to the coalition. The Shapley Value of X_i is the weighted average of the marginal contributions of X_i across all subsets, expressed as:

$$\phi_{X_i} = \sum_{S \subseteq D \setminus \{i\}} w(S) \Delta_i(S) = \frac{1}{p} \sum_{S \subseteq D \setminus \{i\}} \binom{|D| - 1}{|S|}^{-1} [v(S \cup i) - v(S)]. \quad (3.1)$$

We can establish feature importance, by defining the value function. In the global sensitivity literature, a frequently used value function is $v(S) = \mathbb{V}(\mathbb{E}[Y | \mathbf{X}_S]) / \mathbb{V}(Y)$, which represents the explained variance or the variance of the best approximation of Y given \mathbf{X}_S . This value function is nonnegative and monotonically increasing, resulting in a positive global importance measure. When features are independent, this Shapley Value is closely related to the functional ANOVA decomposition [Efron, 1981; Hoeffding, 1948] and Sobol indices [Sobol, 1990; Chastaing, 2012; Hooker, 2007]. The resulting Shapley Values are commonly referred to as Shapley Effects [Owen, 2014; Owen, 2017].

In contrast to the global Shapley Values (SV) approach, known as Shapley Effects, [Lundberg, 2017a; Lundberg, 2020a] adopts the game theory paradigm to explain a specific prediction $f(x_1, \dots, x_p)$ with players $D = \{X_1 = x_1, \dots, X_p = x_p\}$ using the value function $v(S) = \mathbb{E}[f(\mathbf{x}) | \mathbf{X}_S = \mathbf{x}_S]$ or $v(S) = \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}})]$. Although debates continue over the choice between these two value functions [Heskes, 2020; Janzing, 2020; Chen, 2020], we assume in this work that the variables are independent, making these value functions equivalent. We refer to the resulting Shapley Values in this context as Local Shapley Values (L-SV).

A key distinction between the global SV approach (Shapley Effects) and the local approach (L-SV) hinges on the definitions of their respective value functions. While the global value function $v(S) = \mathbb{V}(\mathbb{E}[Y | \mathbf{X}_S]) / \mathbb{V}(Y)$ serves as an effective measure of the predictive power of variables \mathbf{X}_S for the overall model, it is unclear whether the local value function $v(S) = \mathbb{E}[f(\mathbf{x}) | \mathbf{X}_S = \mathbf{x}_S]$ genuinely represents the predictive power of $\mathbf{X}_S = \mathbf{x}_S$ for the specific prediction $f(\mathbf{x})$. In the global case, a high value of $v(S) = \mathbb{V}(\mathbb{E}[Y | \mathbf{X}_S]) / \mathbb{V}(Y)$ indicates a strong predictive power of \mathbf{X}_S , while the values taken by $v(S) = \mathbb{E}[f(\mathbf{x}) | \mathbf{X}_S = \mathbf{x}_S]$ in the local case do not have any intuitive order, i.e., a high or low value of $v(S)$ does not necessarily mean that $\mathbf{X}_S = \mathbf{x}_S$ is important or not for the specific prediction $f(\mathbf{x})$.

Moreover, the value function of L-SV, $v(S) = \mathbb{E}[f(\mathbf{X}) | \mathbf{X}_S = \mathbf{x}_S]$, can be negative and does not satisfy the monotonic property, which may result in negative L-SV. The interpretation of a

negative Shapley Value is unclear. One might assume that a negative Shapley Value indicates that, on average, including this variable in a subset tends to lower the predictions. However, in a regression problem, the model may take negative values, and the negative L-SV could result from the model tending to produce negative values over numerous subsets. Additionally, a canceling effect can occur, where a variable that influences the decision ends up with a zero or low Shapley Value because the $\Delta_i(S)$ values across all subsets cancel each other out. It is important to note that Shapley Effects do not encounter the issues mentioned above, as they satisfy the non-negativity criterion suggested for feature importance [Johnson, 2004; Grömping, 2007; Feldman, 2005]. In fact, [Feldman, 2005] emphasized that an importance measure should be positive, as it evaluates the relative information a variable contributes to the model, and information is inherently non-negative.

Another limitation of Local Shapley Values (L-SV) is the ambiguity surrounding the quantities they estimate, which significantly impedes their interpretation. For instance, what does it mean when an L-SV equals 5? If the L-SV of $X_1 = x_1$ is twice the one of $X_2 = x_2$, can we conclude that $X_1 = x_1$ contributes twice as much as $X_2 = x_2$ to the prediction $f(x_1, \dots, x_p)$? [Verdinelli, 2023] make similar arguments against Shapley Effects.

Lastly, there is no strong justification for calculating contributions across all possible subsets, as some of these subsets might be poor predictors, thus introducing noise into the feature importance. The average performance of a feature across all submodels may not be indicative of the particular performance of that feature in the set of optimal submodels. Additionally, averaging over all subsets tends to reduce the local aspect of the contribution. To demonstrate this, let's assume we have $\mathbf{X} \in \mathbb{R}^p$, such that $\mathbf{X} \sim \mathcal{N}(0, I_p)$, and a piece-wise linear predictor f defined as:

$$f(X) = (a_1X_1 + a_2X_2)\mathbb{1}_{X_5 \leq 0} + (a_3X_3 + a_4X_4)\mathbb{1}_{X_5 > 0}. \quad (3.2)$$

Even if we choose an observation $\mathbf{x} = (x_1, \dots, x_p)$ such that $x_5 \leq 0$ and the predictor only uses x_1, x_2 , the L-SV of ϕ_{x_3}, ϕ_{x_4} is not necessarily zero. Proposition 5.1 of Chapter 2 shows that for all $i \in \{3, 4\}$

$$\phi_{x_i} = K \left(a_i(x_i - \mathbb{E}[\mathbf{X}_i]) \right),$$

where K is a constant. This highlights that the L-SV are not purely local measures but also exhibit global influences. This occurs because when calculating the L-SV of $X_3 = x_3$ or $X_4 = x_4$, we also consider subsets S that do not contain X_5 . By marginalizing and changing the sign of X_5 , we use the other linear model not used for this observation. We can extend the result above to show a similar issue with continuous piece-wise linear function.

Definition 2.1 ([Chua, 1988; Ovchinnikov, 2000; Chen, 2022]). A function $f : \mathcal{X} \rightarrow \mathcal{Y}$ is a m -Continuous Piece-Wise Linear Function (m -CPWL) if there exists K finite set of disjoint convex polytopes $\{A_k\}_{k=1}^m$ such that $\cup_{k=1}^m A_k = \mathcal{X}$ and f restricted to the domain A_k , denoted as $f|_{A_k} : A_k \ni \mathbf{x} \mapsto f(\mathbf{x})$ is affine for each $k \in \{1, \dots, m\}$.

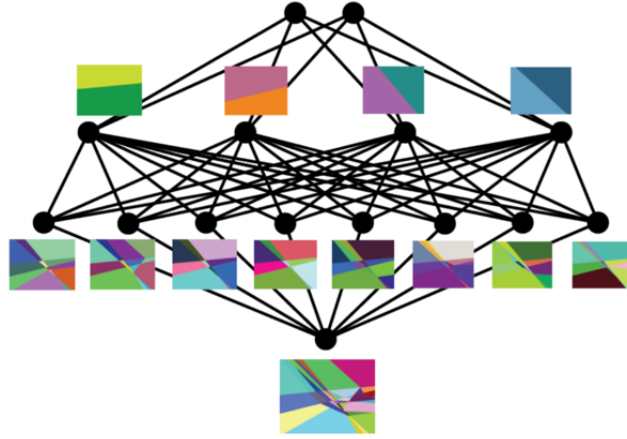


Figure 3.1: Evolution of linear regions within a ReLU network for 2-dimensional input. Each neuron in the first layer defines a linear boundary that partitions the input space into two regions. Neurons in the second layer combine and split these linear boundaries into higher level patterns of regions, and so on [Hanin, 2019a].

This class of functions is quite versatile, as it encompasses neural networks with piece-wise linear activations such as ReLU or hard tanh which correspond to $\max(0, x)$ and $\max(-1, \min(1, x))$ respectively. Indeed, we can view feedforward neural networks as piece-wise linear functions that divide the input space into multiple linear regions, where the network itself behaves as an affine function within each region [Pascanu, 2013; Hanin, 2019a; Hanin, 2019b; Chen, 2022]. Figure 3.1 shows an example of the evolution of linear regions within a feedforward neural networks with ReLU activation in 2 dimensions.

The important local variables of this model correspond to the coefficients of the linear model associated with the region A_k to which the observation belongs. However, in the following, we provide an impossibility result demonstrating that these explanations cannot be retrieved using Local Shapley Values.

Theorem 2.2. *Let f be a piecewise linear function with m components defined by the collection $\{f_{|A_1}, \dots, f_{|A_m}\}$, where $\cup_{k=1}^m A_k = \mathcal{X}$. The regions A_k are disjoint hyperrectangles, specifically $A_k = \otimes_{i=1}^p A_{i,k}$, where $A_{i,k} = [l_{i,k}, r_{i,k}]$ with $l_{i,k}, r_{i,k} \in \overline{\mathbb{R}}$. Each component $f_{|A_k}$ is represented as $f_k(\mathbf{X}) = \sum_{i=1}^p a_{i,k} X_i + b_k$, where the coefficients $a_{i,k}$ and b_k are real numbers. Consequently, f is defined as:*

$$f(\mathbf{X}) = \sum_{k=1}^m \left(\sum_{i=1}^p a_{i,k} X_i + b_k \right) \mathbf{1}_{A_k}(\mathbf{X}).$$

Consider an observation $\mathbf{x} = (x_1, \dots, x_p) \in A_{k^*}$, where $k^* \in \{1, \dots, m\}$, sampled from a distribution $P_{\mathbf{X}}$ with independent covariates such that the model only used $f_{k^*}(\mathbf{x})$ as $f(\mathbf{x}) = f_{k^*}(\mathbf{x})$ on A_{k^*} . The Local SV of a given feature-value $X_l = x_l$ is equal to

$$\phi_{x_l} = \sum_{k=1}^m \phi_{x_l}^k,$$

and $\phi_{x_l}^k$ is defined as

$$\begin{aligned} \phi_{x_l}^k = & \left(\frac{\mathbb{1}_{A_{l,k}}(x_l)}{\mathbb{P}(X_l \in A_{l,k})} - 1 \right) \sum_{S \subseteq D \setminus \{l\}} w(S) v_k(S) \\ & + a_{l,k} \left(x_l - \frac{\mathbb{E}[X_l \mathbb{1}_{A_{l,k}}(X_l)]}{\mathbb{P}(X_l \in A_{l,k})} \right) \sum_{S \subseteq D \setminus \{l\}} w(S) \times \prod_{i \in S \cup l} \mathbb{1}_{A_{i,k}}(x_i) \prod_{j \in \bar{S}} \mathbb{P}(X_j \in A_{j,k}), \end{aligned} \quad (3.3)$$

where $w(S) = \frac{1}{p} \binom{|D|-1}{|S|}^{-1}$ and $v_k(S) = \mathbb{E}[f_k(\mathbf{X}) \mathbb{1}_{A_k}(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$. Equation (3.3) demonstrates that even if the model only uses $f_{k^*}(\mathbf{x})$ for a given observation \mathbf{x} , the Local SV \mathbf{x} may depend on the coefficients of the unused linear models f_k for $k \in \{1, \dots, m\} \setminus \{k^*\}$.

See the proof in the Appendix (7). Theorem 2.2 demonstrated that SV also present difficulties in expressing local importance measures for neural networks with piece-wise linear activation layers and, more generally, for continuous piece-wise linear functions.

3 Stop using LIME

The main idea behind LIME [Ribeiro, 2016a] is to approximate a complex model using a simpler, more interpretable model, such as a linear model, in the vicinity of a given input instance. Given a model f , the local explanation $\xi(\mathbf{x}^*)$ of an instance \mathbf{x}^* is an interpretable model $g \in G$, where G is the set of linear models, such that

$$\xi(\mathbf{x}^*) = \arg \min_{g \in G} \mathcal{L}(f, g, \pi_{\mathbf{x}^*}^h) + \Omega(g), \quad (3.4)$$

where $\mathcal{L}(f, g, \pi_{\mathbf{x}^*}^h)$ measure of how unfaithful g is in approximating f over $\pi_{\mathbf{x}^*}^h$, a measure of locality around x^* with width h , and $\Omega(g)$ is a measure of complexity of the local model g . The loss \mathcal{L} is defined as

$$\mathcal{L}(f, g, \pi_{\mathbf{x}^*}^h) = \sum_{\mathbf{x}' \sim P'} [f(\mathbf{x}') - g(\mathbf{x}')]^2 \pi_{\mathbf{x}^*}^h(\mathbf{x}').$$

In the original implementation [Ribeiro, 2016a], $\pi_{\mathbf{x}^*}^h$ is a Gaussian kernel, and the sum is taken over samples $\mathbf{x}' \sim P'$ where $P' = \prod_i P_{X_i}$ is the marginal law of the features.

A primary concern with LIME stems from its reliance on arbitrary heuristics in its definition. Specifically, choosing the sampling distribution P' poses challenges, as the commonly used distribution disregards feature dependencies, and there is no guarantee that the model's local behavior on P' will be consistent with that on the observed data. Another significant issue lies in defining the neighborhood $\pi_{\mathbf{x}^*}^h$ and tuning the kernel width h , especially in high-dimension. The stability and sensitivity of LIME are heavily influenced by the selection of the perturbation sampling distribution, the definition of proximity $\pi_{\mathbf{x}^*}^h$ and the bandwidth h , which may result in varying explanations for the same instance under slightly varying settings. To illustrate this issue, we apply LIME on the piece-wise linear model defined in (3.2), with $a_1 = 0, a_2 = 2, a_3 = 0, a_4 = 5$.

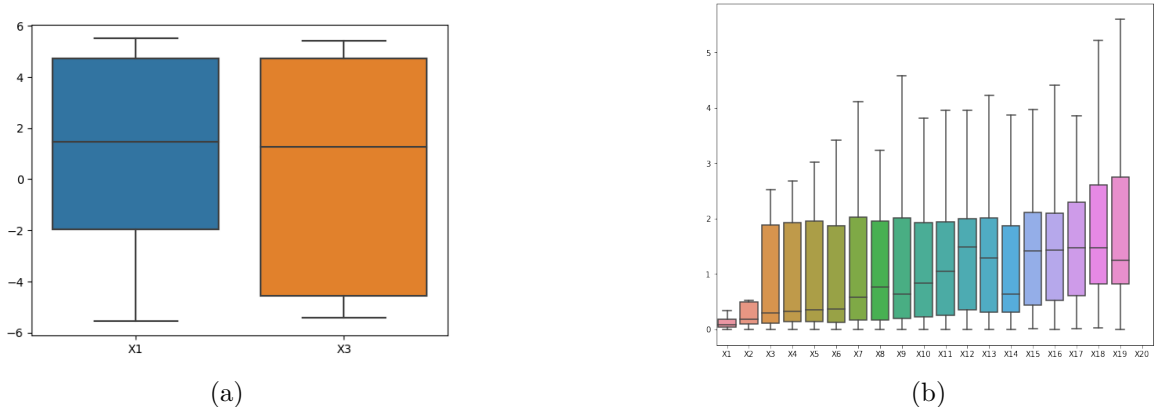


Figure 3.2: (a) LIME coefficients for X_1 and X_3 of the piece-wise linear model on observations that have $X_5 \leq 0$, (b) Relative Absolute Error of LIME coefficients on German credit dataset.

In Figure 3.2a, the distribution of LIME coefficients for X_1 and X_3 among observations with negative x_5 values shows that LIME assigns a non-null score to both X_1 and X_3 , although the latter is not used locally for the displayed observations. Thus, LIME shares the same problem as L-SV in the piece-wise linear model with independent variables. It is also important to note that such discontinuities are not uncommon, as tabular data often contain discontinuities with categorical variables, and the meaning of constructing a linear approximation in such cases remains unclear. Besides this empirical evidence, we are currently working to compute the theoretical quantity of LIME coefficient for continuous piece-wise-linear functions, yielding impossible results similar to Theorem 2.2 for Local Shapley Values.

To demonstrate the sensitivity of LIME to bandwidth selection, we applied it to the German credit dataset ($n = 1000, p = 20$) from UCI [Dua, 2017a]. We trained a RF on 80% of the dataset and computed LIME coefficients on the remaining data using two bandwidths, h and h' , of the proximity kernel $\pi_{\mathbf{x}^*}^h$. We set h using the median heuristic [Fukumizu, 2009; Flaxman, 2016; Garreau, 2017], where h is the median of the pairwise distance of $\|\mathbf{X}_i - \mathbf{X}_j\|_2$ and assigned $h' = \frac{1}{2}h$. In Figure 3.2b, we compare the relative absolute error of the LIME coefficients for all variables using the two bandwidths, i.e., $(L_{x_i}^h - L_{x_i}^{h'})/L_{x_i}^h$, where $L_{x_i}^h, L_{x_i}^{h'}$ represent the LIME coefficient of the variable X_i using bandwidth h, h' respectively. It reveals that the difference between the LIME coefficients could be much different after slightly modifying the bandwidth. Furthermore, we observed that 20% of the coefficients also changed signs. In real-world scenarios, we lack information about the true local importance, making the bandwidth selection process indefinite. Moreover, LIME exhibits issues related to instability or irreproducibility. For example, [Zhang, 2019; Zafar, 2019; Visani, 2020] have shown that repeated runs of the same setting of the algorithm on the same model and data point can yield different results. This inconsistency stems from the randomness introduced during the generation of the synthetic sample around the input. Several works [Zafar, 2019; Zhou, 2021] attempt to address this issue, using asymptotic analysis of the method to identify the minimum number of the sampled observations required for stability. Another aspect of stability is related to input perturbations. [Alvarez-Melis, 2018] show that nearby observations may have completely different LIME coefficients.

4 From Global Explanations to Regional Explanations

In this section, we follow the presentation of [Williamson, 2021], which introduces a general framework for global variable importance. We consider a comprehensive class \mathcal{F} of functions mapping from \mathcal{X} to \mathcal{Y} , and \mathcal{F}_{-j} be the subset of \mathcal{F} containing all functions that disregard the variable X_j . The conformity score $V(f(\mathbf{X}), Y)$ assesses the predictiveness of a prediction function $f \in \mathcal{F}$ on the observation (\mathbf{X}, Y) , where a high value implies high predictiveness. We define the oracle predictor with respect to the conformity score V and distribution $P = P_{\mathbf{X}}P_{Y|\mathbf{X}}$ as follows:

$$f_0 = \arg \max_{f \in \mathcal{F}} \mathbb{E}_P[V(f(\mathbf{X}), Y)]$$

In a similar manner, we define f_{-j} as the function maximizing $\mathbb{E}_P[V(f(\mathbf{X}), Y)]$ over all $f \in \mathcal{F}_{-j}$. Subsequently, we define the *population-level importance* for the variable X_j as the decrease in predictiveness when excluding X_j from $\mathbf{X} = (X_1, \dots, X_p)$. This is commonly referred to as the Leave Out COvariates (LOCO) importance in existing literature [Lei, 2016]. The LOCO importance for X_j is defined as:

$$\Psi_j(P) = \mathbb{E}_P[\Delta_j(\mathbf{X}, Y)] \quad \text{where} \quad \Delta_j(\mathbf{X}, Y) = V(f_0(\mathbf{X}), Y) - V(f_{-j}(\mathbf{X}), Y).$$

We can use any conformity score to measure variable importance, depending on the problem. For regression tasks, we can use $V(f(\mathbf{X}), Y) = 1 - [Y - f(\mathbf{X})]^2/\sigma^2$, where $\sigma^2 = \mathbb{E}_P[Y - \mathbb{E}_P[Y]]^2$ represents the variance of the target variable Y . This conformity score corresponds to the traditional R^2 score at the population level, i.e., $R^2 = \mathbb{E}_P[V(f(\mathbf{X}), Y)]$. Alternatively, for binary classification problems, we can use $V(f(\mathbf{X}), Y) = \mathbb{1}_{Y=f(\mathbf{X})}$, which corresponds to the accuracy score at the population-level. Regarding the choice of the conformity score, there is no ground truth for variable importance as there are multiple definitions of what makes a variable important [Hooker, 2019; Hama, 2022; Verdinelli, 2023]. Consequently, the choice of a conformity score should be contingent upon the specific context and goals of the analysis.

Our approach aims to derive local explanations for a specific observation (\mathbf{X}, Y) from the population-level importance $\Psi_j(P)$. This involves identifying a partition $\cup_i A_i = \mathcal{X}$ containing observations with similar explanations, which means $\mathbb{V}_P(\Delta_j(\mathbf{X}, Y) \mid \mathbf{X} \in A_i) \approx 0$ for all $j \in \{1, \dots, p\}$ simultaneously. In other words, the observations within each partition have low variance in their feature importance. As a result, we can use $\Psi_j(P_{A_i}) = \mathbb{E}_{P_{A_i}}[\Delta_j(\mathbf{X}, Y)] = \mathbb{E}_P[\Delta_j(\mathbf{X}, Y) \mid \mathbf{X} \in A_i]$ as local explanations for all observations in A_i , since the random variable $\Delta_j(\mathbf{X}, Y)$ exhibits low variance within A_i . Essentially, our approach involves identifying a homogeneous group with respect to importance measure $\Delta_j(\mathbf{X}, Y)$ and attributing the global importance of this group as the local explanations of its members. Hence, it permits us to have local explanations while benefiting from all the inference results available for global explanations.

4.1 Estimation and inference

To compute our local explanations, we need to compute two quantities: the LOCO importance $\Psi_j(P) = \mathbb{E}_P[\Delta_j(\mathbf{X}, Y)]$ for any distribution P and the partition $\cup_i A_i = \mathcal{X}$ that group observations by their feature importance similarity. The former has been extensively studied in [Williamson, 2021], where the authors proposed a nonparametric efficient estimation procedure using the following plug-in estimator:

$$\widehat{\Psi}_j(\widehat{P}) = \mathbb{E}_{\widehat{P}}[\widehat{\Delta}_j(\mathbf{X}, Y)] \quad \text{where} \quad \widehat{\Delta}_j(\mathbf{X}, Y) = V(\widehat{f}_0(\mathbf{X}), Y) - V(\widehat{f}_{-j}(\mathbf{X}), Y), \quad (3.5)$$

$\widehat{P} = 1/n \sum_{i=1}^n \delta_{(\mathbf{X}_i, Y_i)}$ is the empirical distribution of P based on $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ and $\widehat{f}_0, \widehat{f}_{-j}$ are estimators of the population minimizers f_0 and f_{-j} respectively. $\widehat{f}_0, \widehat{f}_{-j}$ are obtained by building a predictive model for Y using all features in \mathbf{X} and after removing the variable X_j respectively. This can be achieved using any machine learning algorithm. [Williamson, 2021] demonstrated that the estimator defined in Equation (3.5), is asymptotically efficient and enables valid statistical inference under regularity conditions.

Having obtained a consistent estimator for $\Psi_j(P)$, we now propose a method to derive the partition $\cup_i A_i = \mathcal{X}$. The initial step involves creating a new representation for each observation $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$ in \mathcal{D}_n using the conformity score V . This is expressed as $\tilde{\mathbf{X}}_i = (\widehat{\Delta}_1(\mathbf{X}_i, Y_i), \dots, \widehat{\Delta}_p(\mathbf{X}_i, Y_i))$. Representing the data in the space of feature importance, rather than the original covariate space, allows for the grouping of observations that exhibit consistent behavior with respect to the importance measure $\widehat{\Delta}_j(\mathbf{X}, Y) = V(\widehat{f}_0(\mathbf{X}), Y) - V(\widehat{f}_{-j}(\mathbf{X}), Y)$ across all variables $j \in \{1, \dots, p\}$ simultaneously.

The subsequent step involves clustering similar observations based on their new representations $\tilde{\mathbf{X}}$. This can be achieved using any clustering algorithm, such as K-means, DBSCAN, or Affinity Propagation, ultimately resulting in a partition of \mathcal{D}_n into K sets $\mathbf{C} = \{C_1, \dots, C_K\}$. A comprehensive overview of these methods can be found in [Schaeffer, 2007].

The final step entails using the identified clusters \mathbf{C} to define a partition of \mathcal{X} . This is accomplished by assigning any new points $\mathbf{x} \in \mathcal{X}$ to their nearest partition C_k with respect to a similarity function $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$. We can use any similarity function such as Euclidean or Manhattan distance. More formally, we define the corresponding region \widehat{A}_k of cluster C_k as

$$\widehat{A}_k = \left\{ \mathbf{x} \in \mathcal{X} : \sum_{(\mathbf{X}_i, Y_i) \in C_k} d(\mathbf{x}, \mathbf{X}_i) < \sum_{(\mathbf{X}_i, Y_i) \in C_l} d(\mathbf{x}, \mathbf{X}_i) \quad \text{for all } k \neq l \right\}.$$

In order to establish a partition of \mathcal{X} , we must also account for the set of "undecidable" observations, which are those that simultaneously belong to multiple groups. We define this set as follows:

$$\widehat{A}_{K+1} = \left\{ \mathbf{x} \in \mathcal{X} : \exists k, l \in \{1, \dots, K\}, \sum_{(\mathbf{X}_i, Y_i) \in C_k} d(\mathbf{x}, \mathbf{X}_i) = \sum_{(\mathbf{X}_i, Y_i) \in C_l} d(\mathbf{x}, \mathbf{X}_i) \right\}.$$

Hence, the local explanations of a given observation \mathbf{X}_i that belongs to the region \hat{A}_k can be represented by the vector $(\hat{\Psi}_1(\hat{P}_{\hat{A}_k}), \dots, \hat{\Psi}_p(\hat{P}_{\hat{A}_k}))$, where the contribution of the feature X_j , $\hat{\Psi}_j(\hat{P}_{\hat{A}_k})$, is defined as follows:

$$\hat{\Psi}_j(\hat{P}_{\hat{A}_k}) = \mathbb{E}_{\hat{P}_{\hat{A}_k}} [\hat{\Delta}_j(\mathbf{X}, Y)] \quad (3.6)$$

$$= \mathbb{E}_{\hat{P}} [\hat{\Delta}_j(\mathbf{X}, Y) \mid \mathbf{X} \in \hat{A}_k]. \quad (3.7)$$

$$= \sum_{(\mathbf{X}_i, Y_i) \in C_k} \frac{\hat{\Delta}_j(\mathbf{X}_i, Y_i)}{|C_k|}$$

We called this approach Regional LOCO (R-LOCO) and its outline is the following:

1. Apply the transformation $\hat{\Delta}_j(\mathbf{X}, Y)$ to the features

$$\begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,p} \\ X_{2,1} & X_{2,2} & \dots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \dots & X_{n,p} \end{bmatrix} \xrightarrow{\text{Apply } \hat{\Delta}_j(\mathbf{X}, Y)} \begin{bmatrix} \hat{\Delta}_1(\mathbf{X}_1, Y_1) & \hat{\Delta}_2(\mathbf{X}_1, Y_1) & \dots & \hat{\Delta}_p(\mathbf{X}_1, Y_1) \\ \hat{\Delta}_1(\mathbf{X}_2, Y_2) & \hat{\Delta}_2(\mathbf{X}_2, Y_2) & \dots & \hat{\Delta}_p(\mathbf{X}_2, Y_2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Delta}_1(\mathbf{X}_n, Y_n) & \hat{\Delta}_2(\mathbf{X}_n, Y_n) & \dots & \hat{\Delta}_p(\mathbf{X}_n, Y_n) \end{bmatrix}$$

Instead of computing the LOCO importance by averaging the transformed features column-wise as follows using $\hat{\Psi}_j(\hat{P}) = \mathbb{E}_{\hat{P}} [\hat{\Delta}_j(\mathbf{X}, Y)] = \sum_{i=1}^n \frac{\hat{\Delta}_j(\mathbf{X}_i, Y_i)}{n}$ for all $j \in \{1, \dots, p\}$:

$$\begin{bmatrix} \hat{\Delta}_1(\mathbf{X}_1, Y_1) & \hat{\Delta}_2(\mathbf{X}_1, Y_1) & \dots & \hat{\Delta}_p(\mathbf{X}_1, Y_1) \\ \hat{\Delta}_1(\mathbf{X}_2, Y_2) & \hat{\Delta}_2(\mathbf{X}_2, Y_2) & \dots & \hat{\Delta}_p(\mathbf{X}_2, Y_2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Delta}_1(\mathbf{X}_n, Y_n) & \hat{\Delta}_2(\mathbf{X}_n, Y_n) & \dots & \hat{\Delta}_p(\mathbf{X}_n, Y_n) \end{bmatrix} \xrightarrow[\text{by column}]{\text{Average}} \begin{array}{cccc} \hat{\Delta}_1(\mathbf{X}_1, Y_1) & \hat{\Delta}_2(\mathbf{X}_1, Y_1) & \dots & \hat{\Delta}_p(\mathbf{X}_1, Y_1) \\ \hat{\Delta}_1(\mathbf{X}_2, Y_2) & \hat{\Delta}_2(\mathbf{X}_2, Y_2) & \dots & \hat{\Delta}_p(\mathbf{X}_2, Y_2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Delta}_1(\mathbf{X}_n, Y_n) & \hat{\Delta}_2(\mathbf{X}_n, Y_n) & \dots & \hat{\Delta}_p(\mathbf{X}_n, Y_n) \\ \hline \hat{\Psi}_1(\hat{P}) & \hat{\Psi}_2(\hat{P}) & \dots & \hat{\Psi}_p(\hat{P}) \end{array}$$

2. We compute the average of the transformed features column-wise using the clusters $\cup_{k=1}^{K+1} A_k = \mathcal{X}$, $\hat{\Psi}_j(\hat{P}_{\hat{A}_k}) = \sum_{(\mathbf{X}_i, Y_i) \in C_k} \frac{\hat{\Delta}_j(\mathbf{X}_i, Y_i)}{|C_k|}$, $j \in \{1, \dots, p\}$, $k \in \{1, \dots, K\}$ to have feature attributions for each observation,

$$\begin{array}{l} \hat{A}_k \\ \hat{A}_k \\ \vdots \\ \hat{A}_3 \end{array} \left| \begin{array}{cccc} \hat{\Delta}_1(\mathbf{X}_1, Y_1) & \hat{\Delta}_2(\mathbf{X}_1, Y_1) & \dots & \hat{\Delta}_p(\mathbf{X}_1, Y_1) \\ \hat{\Delta}_1(\mathbf{X}_2, Y_2) & \hat{\Delta}_2(\mathbf{X}_2, Y_2) & \dots & \hat{\Delta}_p(\mathbf{X}_2, Y_2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Delta}_1(\mathbf{X}_n, Y_n) & \hat{\Delta}_2(\mathbf{X}_n, Y_n) & \dots & \hat{\Delta}_p(\mathbf{X}_n, Y_n) \end{array} \xrightarrow[\text{LOCO computed by its group } \hat{A}_k]{\text{Assigned each observation}} \begin{array}{l} \hat{\Psi}_1(\hat{P}_{\hat{A}_k}), \dots, \hat{\Psi}_p(\hat{P}_{\hat{A}_k}) \\ \hat{\Psi}_1(\hat{P}_{\hat{A}_k}), \dots, \hat{\Psi}_p(\hat{P}_{\hat{A}_k}) \\ \vdots \\ \hat{\Psi}_1(\hat{P}_{\hat{A}_3}), \dots, \hat{\Psi}_p(\hat{P}_{\hat{A}_3}) \end{array}$$

5 Experiments

In this section, we compare the regional feature attributions of our approach, Regional LOCO (R-LOCO), with SHAP and LIME on models f for which we have knowledge of the local important variables. To evaluate the effectiveness in detecting local important variables from feature attributions, it is essential to establish a significance threshold for considering a variable as important. The Shapley and LIME values lack a clear interpretation, making it challenging to choose an appropriate threshold value. In contrast, R-LOCO provides a clear interpretation: the drop in predictive performance when a variable is removed. This allows users to set a threshold value based on an acceptable performance drop for the specific problem at hand. However, to ensure a more objective analysis, we compare the methods by varying the thresholds for each method. We begin by projecting the feature attributions of each observation onto the simplex, by transforming the attribution of each method, $m \in \{\text{SHAP}, \text{LIME}, \text{R-LOCO}\}$, $\phi_{x_i}^m$ as $\tilde{\phi}_{x_i}^m = |\phi_{x_i}^m| / \sum_{j=1}^p |\phi_{x_j}^m|$. Subsequently, we use quantiles $q \in \{0.2, 0.3, 0.4, 0.5\}$ of the feature attributions of each observation as threshold values for each method m . We consider a variable-value $X_i = x_i$ to be significant for the prediction $f(x_1, \dots, x_p)$ by method m if $\tilde{\phi}_{x_i}^m \geq \mathcal{Q}(q, \{\tilde{\phi}_{x_i}^m\}_{j=1}^p)$, where $\mathcal{Q}(q, \{\tilde{\phi}_{x_i}^m\}_{j=1}^p)$ is the q -quantile of the values $\{\tilde{\phi}_{x_i}^m\}_{j=1}^p$. In all our experiments, we have at least 50% of the variables that are considered important for each prediction. As a result, we set $q = 0.5$ as the maximum quantile value to ensure that each method can choose at least 50% of the variables. Our evaluation metrics are the True Discovery Rate (TDR) and the False Discovery Rate (FDR), which assess the ability of each method to identify the model’s local important variables. The TDR (higher is better) reflects the proportion of truly important variables to which each method assigns a score higher than the threshold chosen using the quantile $q \in \{0.2, 0.3, 0.4, 0.5\}$. Conversely, the FDR (lower is better) represents the fraction of non-important local variables that incorrectly receive a score higher than the threshold by each method.

The comparative analysis encompasses three distinct experiments in which the variables are independent to eliminate any potential bias in detecting important variables when dependencies exist. We generate a set of n samples, denoted as $\{\mathbf{X}_i\}_{i=1}^n$, drawn from a distribution $\mathbf{X} \sim P_{\mathbf{X}} = \prod_{i=1}^p P_{X_i}$. By applying a function $Y = f(\mathbf{X})$, we construct a dataset $\mathcal{D}_n = \{(\mathbf{X}_i, f(\mathbf{X}_i))\}_{i=1}^n$. Subsequently, SHAP and LIME received the model f to be explained and the data \mathcal{D}_n as input. On the other hand, R-LOCO exclusively receives the observed data \mathcal{D}_n . For the R-LOCO method, we used Affinity Propagation [Frey, 2007] implemented via scikit-learn [Pedregosa, 2011] to identify clusters $\cup_{k=1}^{K+1} \hat{A}_k$ that were subsequently used to calculate LOCO by region as defined in Equation (3.6). All the regression functions used in our analysis are defined piece-wise, with each piece involving different variables. Thus, the true clusters correspond to the regions where each piece of the function is defined. It’s important to note that approximating these regions using clustering methods presents challenges. Consequently, we computed R-LOCO using the ground-truth clusters as well. This alternate approach, referred to as Regional LOCO - Truth Cluster (R-LOCO - TC), serves as a benchmark to compare against R-LOCO with approximated clusters.

Experiment 1: Piece-wise Linear Model. We propose a simple piece-wise linear model where the local important variables can be read out as in a classical linear model. Let $\mathbf{X} = (X_1, \dots, X_{10})$ with each components being independent, and $X_i \sim \mathcal{U}[-1, 1]$ for all $i \in \{1, \dots, 10\}$. The predictor function is defined as

$$f(\mathbf{X}) = \begin{cases} X_1 + X_2, & \text{if } X_{10} \leq 0 \\ X_5 + X_6 & \text{otherwise} \end{cases}$$

In Figure 3.3, the TDR and FDR scores remain constant across the observations for this example, and all four methods successfully identify the local important variables for all thresholds $q \in \{0.2, 0.3, 0.4, 0.5\}$. However, SHAP and LIME exhibited a high False Discovery Rate (FDR) of 50 and 60 percent, respectively. This indicates that for each observation, among the variables assigned score above the threshold, 50 to 60 percent are actually not important. Contrarily, Regional-LOCO and Regional-LOCO Truth Cluster, detected all the significant variables with an FDR of 0, demonstrating superior precision compared to SHAP and LIME.

Remark: The classic global importance measure, LOCO, can be calculated from R-LOCO as follows $\hat{\Psi}_j(\hat{P}) = \sum_{k=1}^K \frac{|C_k|}{n} \hat{\Psi}_j(\hat{P}_{\hat{A}_k})$ where $|C_k|$ represent the number of observations of \mathcal{D}_n in \hat{A}_k . By calculating it, for this example, we observe unsurprisingly that the model considers all the variables X_1, X_2, X_5, X_6 as important, showing the benefit of R-LOCO over LOCO that enables to identify the two regimes of this model.

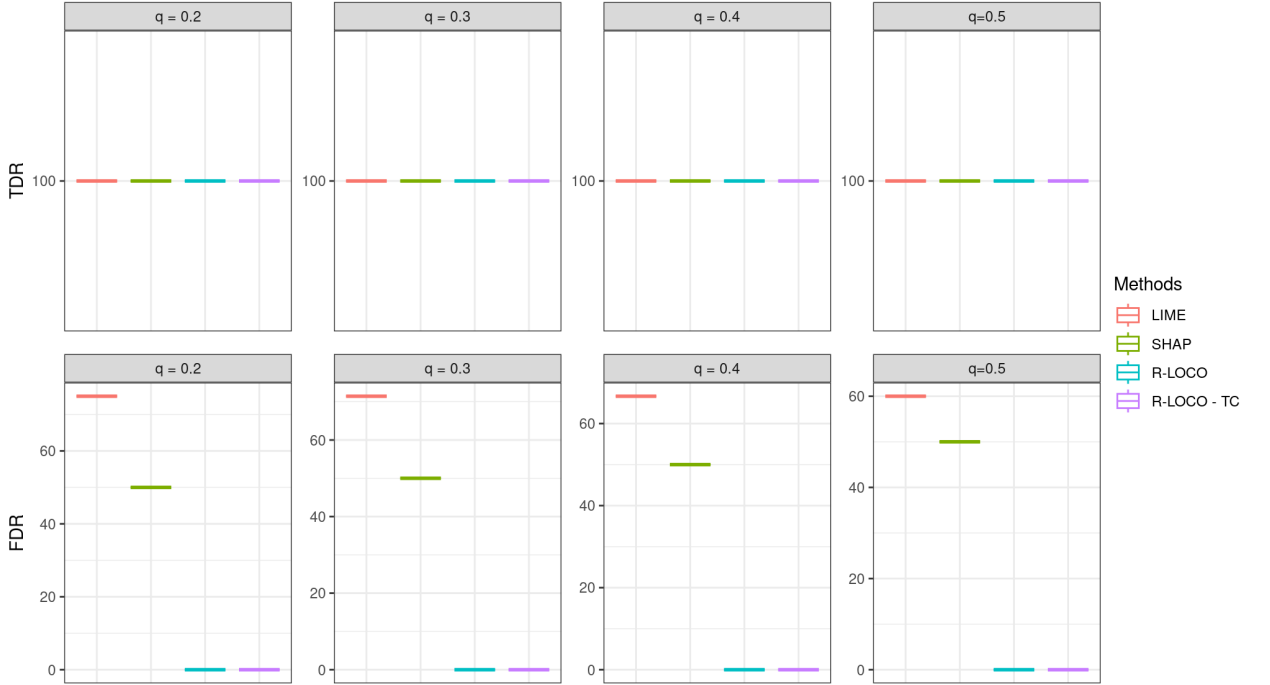


Figure 3.3: TDR and FDR of each method on the piece-wise linear model across 10000 observations of test set.

Experiment 2: High-order Interactions. Here, we use the independent variables of the previous experiment on a more complex function having interactions, used in [Bénard, 2021b], defined as follows:

$$f(\mathbf{X}) = 3\sqrt{3} \times X_1 X_2 \mathbb{1}_{X_3 > 0} + \sqrt{3} \times X_4 X_5 \mathbb{1}_{X_3 \leq 0} + 3 \times X_6 X_7 \mathbb{1}_{X_8 > 0} + X_9 X_{10} \mathbb{1}_{X_8 \leq 0}.$$

Within this model, we discern four distinct regimes. The potentially locally important variables are $\{X_1, X_2\}$, $\{X_4, X_5\}$, $\{X_6, X_7\}$, or $\{X_9, X_{10}\}$, subject to the sign of $\{X_3, X_8\}$. The ground-truth clusters are defined using the sign of $\{X_3, X_8\}$. As Figure 3.4 demonstrates, the distribution of the TDR and FDR exhibits more variance between observations in this model. For small values of the thresholds, $q = 0.2$ and 0.3 , the Shapley Values successfully identify the important variables. However, for $q = 0.4$ and 0.5 , the Shapley Values struggle, as indicated by the wider TDR interquartile varying between 50 – 75%. On the other hand, LIME fails to detect all the important variables regardless of the threshold value, with the TDR varying between 50 – 100% across the different thresholds. Conversely, the Regional LOCO method successfully detects all important variables, with TDR= 100% for all q . The interquartile range of the False Discovery Rate for R-LOCO, SHAP, and LIME falls within 40 – 60%, 20 – 40%, and 0 – 20%, respectively. In this example, we see a difference in performance between the R-LOCO and the R-LOCO - TC. Unlike R-LOCO, the R-LOCO - TC cluster has an FDR of 0, thus showing the importance of identifying the right clusters. Overall, our approach is much better than SHAP and LIME for detecting local important variables.

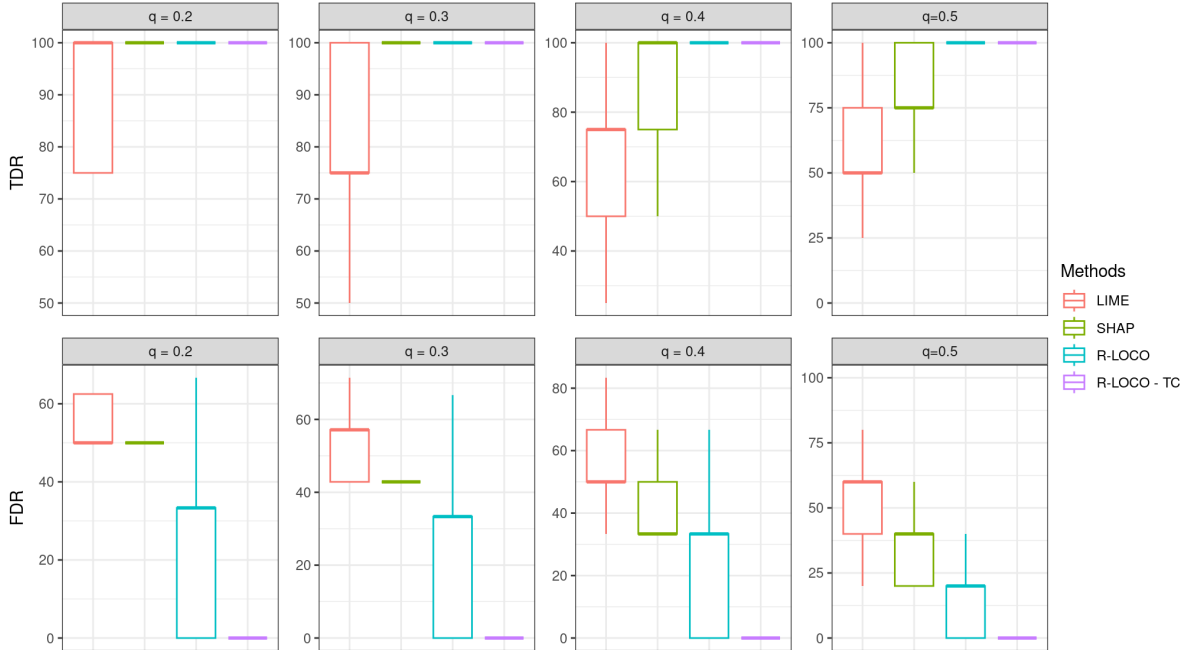


Figure 3.4: TDR and FDR of each method on the model with interactions across 10000 observations of test set.

Experiment 3: Linear Tree. In our final experiment, we utilize a linear tree - a variant of a decision tree where predictions are derived from a linear model in each leaf, in contrast to averaging the outputs of the observations that belong to the leaf. This model has been featured in several studies as an instrument to construct interpretable models [Künzel, 2022] or enhance random forest or boosting tree [Friedberg, 2020; Athey, 2019]. We use the package `linear-tree` and train a linear tree f on the California House Price dataset [Kelley Pace, 1997] ($n = 20640, p = 8$). The learned tree is displayed in Figure 3.5.

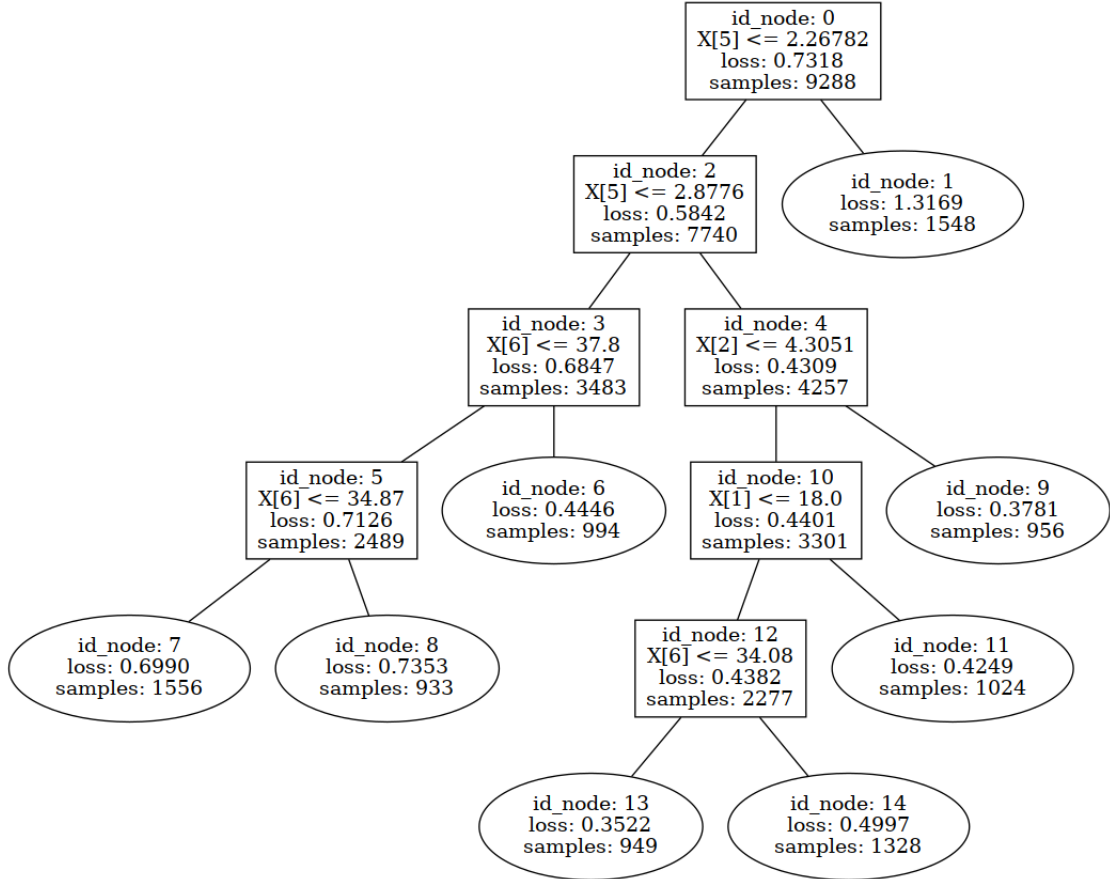


Figure 3.5: Learned linear tree on California house price dataset. Each node of the tree displays the loss, split variable, and the number of samples.

To mitigate any potential bias in detecting important variables when dependencies exist, we shuffle each column of the dataset, generating new inputs with independent covariates. More precisely, assuming $P_{\mathbf{X}}$ represents the distribution of the features, we generate new inputs $\{\mathbf{X}_i^\circ\}_{i=1}^n$ using the marginal distribution of the features $\mathbf{X}_i^\circ \sim \prod_i^p P_{X_i}$, and then aim to explain the predictions $\{f(X_i^\circ)\}_{i=1}^n$. In this model, the local important variables are the non-null coefficients of the linear model at the node where the observation is located. We evaluate the model's ability to discern the non-zero coefficients of the linear model of the leaf where the observation belongs. Similar to the previous experiments, Figure 3.6 suggests that R-LOCO outperforms SHAP, which in turn outshines LIME. However, R-LOCO with the truth cluster shows a slightly superior performance to R-LOCO, underscoring the importance of finding the correct cluster.

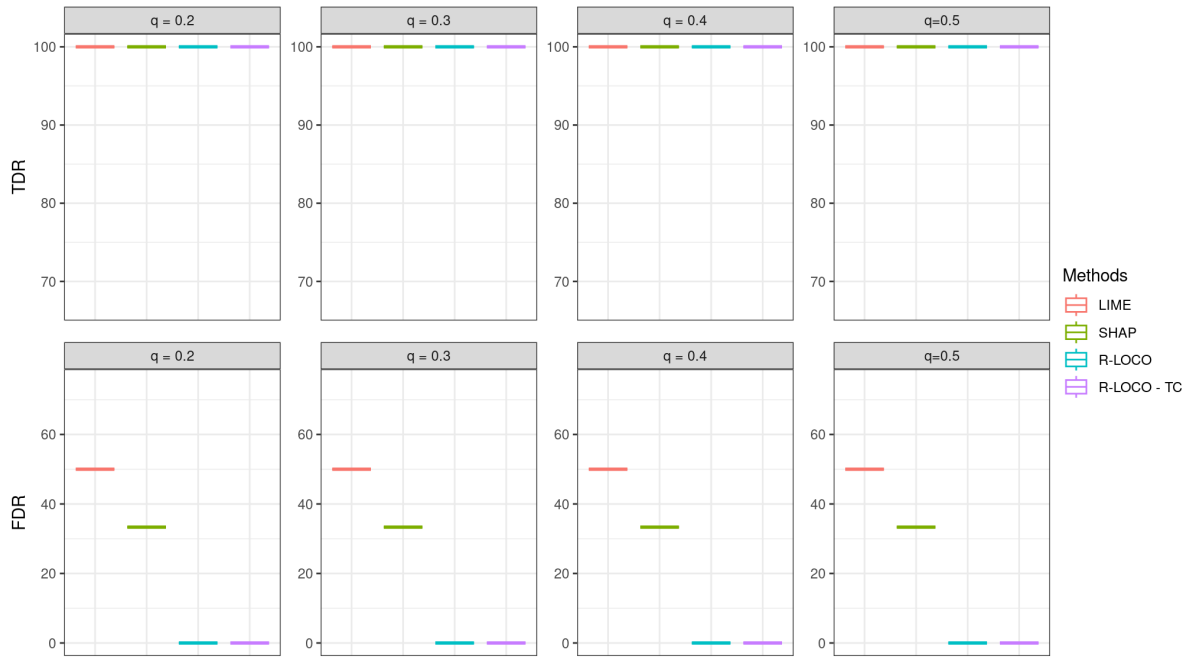


Figure 3.6: TDR and FDR of each method on the Linear Tree

6 Discussion

In this chapter, we have introduced a technique for identifying distinct influence zones of the model under examination, which are employed to calculate overall importance on a regional basis. These regions provide us with a more detailed understanding compared to global importance measures, although they are not specific to each observation. This regional importance approach strikes a balance between the local and global approaches. The key challenge in this approach lies in accurately identifying the diverse zones of influence within the model, which is no easy task. One potential improvement to enhance cluster identification is by enriching our representation, denoted as $\tilde{\mathbf{X}}_i = (\hat{\Delta}_1(\mathbf{X}_i, Y_i), \dots, \hat{\Delta}_p(\mathbf{X}_i, Y_i))$, through the addition of second-order interactions. This enriched representation would take the form of $\tilde{\mathbf{X}}_i = (\hat{\Delta}_1(\mathbf{X}_i, Y_i), \dots, \hat{\Delta}_p(\mathbf{X}_i, Y_i), \hat{\Delta}_{1,1}(\mathbf{X}_i), \hat{\Delta}_{1,2}(\mathbf{X}_i), \dots, \hat{\Delta}_{p,p}(\mathbf{X}_i))$, and may even incorporate higher-order interactions. Another significant challenge is ensuring that each region contains a minimum number of observations to guarantee stability. Overall, we propose this approach as an alternative to other existing local explanation methods, such as SHAP and LIME, which are unreliable for the detection of local important variables. We have demonstrated through several examples with independent variables that R-LOCO outperforms these methods. Moreover, this approach avoids extrapolation with out-of-distribution observations as it uses only observed observations. Our objective was to show that SHAP and LIME do not work even when the variables are independent and to propose an alternative. However, it would also be interesting to see our method when the variables are dependent. We believe that the problems that arise when the variables are dependent are independent of the choice of methodology. Independent strategies should be used to correct them, such as grouping correlated variables, as explored by

[Verdinelli, 2023] in their analysis of LOCO. While the definition of a local contribution of a variable $X_i = x_i$ to a prediction $f(x_1, \dots, x_p)$ remains unclear, we suggest using this intermediate approach between the local and global approaches, aiming to identify different model regimes. However, from a more theoretical perspective, except for piece-wise functions, it is extremely challenging to define these zones of influence, particularly for general continuous functions. The existence and uniqueness of such partitions pose theoretical questions that we leave for future studies. In the next chapter, we also introduce an approach called Sufficient Rules, which permits finding partitions of observations based on their predictions. This method aims to identify groups of observations with similar predictions, further expanding our understanding of the model's behavior.

Chapter 4

Beyond Features attributions: Sufficient Explanations and Rules

Abstract

To explain the decision of any regression and classification model, we extend the concept of probabilistic sufficient explanations (P-SE). For each instance, this approach selects the minimal subset of features that is sufficient to yield the same prediction with high probability while removing other features. The crux of P-SE is to compute the conditional probability of maintaining the same prediction. Therefore, we introduce an accurate and fast estimator of this probability via Random Forest for any data and show its efficiency through a theoretical analysis of its consistency. Consequently, we extend the P-SE to address regression problems and deal with non-discrete features, without learning the distribution of input nor having the model for making predictions. Finally, we introduce local rule-based explanations for regression/classification based on the P-SE and compare our approaches with other explainable AI methods. These methods are available as a [Python package](#).

Contents

1	Introduction	68
2	Motivations and Related works	69
3	Probabilistic Sufficient Explanations for Regression	70
4	SDP, Sufficient Explanations and Sufficient Rules via Random Forest	72
5	Experiments	79
6	Conclusion	83

1 Introduction

Many methods have been proposed to explain specific predictions of machine learning models from different perspectives, such as feature attributions approaches [Lundberg, 2017b; Ribeiro, 2016a], decision rules [Ribeiro, 2018], counterfactual examples [Wachter, 2017] and logic-based [Shih, 2018; Darwiche, 2020].

Among these categories, the most popular are feature attributions approaches, in particular SHAP [Lundberg, 2017a], which is based on Shapley Values (SV) and aims at indicating the importance of each feature in the decision. One of the main reasons for SHAP’s success is its scalability, nice representations of the explanations, and game-theoretic foundations. However, SV used in SHAP does not guarantee the truthfulness of the important variables involved in a given decision. Indeed, it is possible to construct simple theoretical models with discontinuity for which SV cannot distinguish between local important and nonimportant variables; see Chapter 3. Similar difficulties have also been highlighted by [Ghalebikesabi, 2021] for SHAP and LIME [Ribeiro, 2016a]. This lack of guarantees is a major issue since the explanations may be used for high-stakes decisions. Moreover, additive explanations are not suitable when interactions occur in the model [Gosiewska, 2019].

An appealing solution to the problem above is to use decision rules [Ribeiro, 2018] or logic-based explanations [Darwiche, 2020; Shih, 2018], which gives local explanations that take into account interactions while ensuring minimality and guarantee on the outcome. However, these methods are not currently available in the general case (e.g., regression model, continuous features). Our objective is to extend these methods to more realistic cases by developing new consistent algorithms.

In this paper, we generalize the concept of Probabilistic Sufficient Explanations (P-SE) introduced by [Wang, 2020]. P-SE is a relaxation of logic-based explanation: it explains the classification of an example by choosing a minimal subset of features guaranteeing that, the model makes the same prediction with high probability, whatever the values of the remaining features under the data distribution. Such a subset is called a Sufficient Explanation (also known as sufficient reason or prime implicant [Shih, 2018; Darwiche, 2020]).

We make several contributions. We extend the concept of Same Decision Probability (SDP) to the regression setting so that we can extend Sufficient Explanations from classification to regression. We introduce a fast and efficient estimator of the SDP based on Random Forests and prove its uniform almost sure convergence. In contrast to [Wang, 2020], our approach can deal with non-discrete features and does not need the estimation of the distribution of \mathbf{X} . Our method can explain the data generating process (\mathbf{X}, Y) directly or any learnt model $(\mathbf{X}, f(\mathbf{X}))$. We introduce the probabilistic local explanatory importance which is the frequency of each feature to be in the set of all Sufficient Explanations. In particular, this summarizes the diversity of the Sufficient Explanations. We introduce local rule-based explanations for classification or regression which are simultaneously minimal and sufficient. We compare our approaches with other explainable AI methods and provide a [Python package](#) that computes all our methods.

2 Motivations and Related works

The methods used to explain the local behavior of machine learning models can be organized into 5 groups: features attributions, decision rules, instance-wise feature selection, logical reasoning approaches, data generation based or counterfactual examples. The benefits of feature attribution-based explanations, e.g., SHAP [Lundberg, 2020a] or LIME [Ribeiro, 2016a] is that they are easy to read, they can be applied to any model and are generally more scalable than their alternatives. On the other hand, they are sensitive to perturbations [Ignatiev, 2019], or can be fooled by adversarial attacks [Slack, 2020]. These downsides can be caused by the local perturbations used, which make them inconsistent with the data distribution.

Quite differently, instance-wise feature selection such as L2X [Chen, 2018] or INVASE [Yoon, 2018] aims at finding the minimal subset of variables that are relevant for a given instance \mathbf{x} and its label y . Interactions can be captured in that way. In addition, the identification of a minimal subset $S(\mathbf{x}) = S$ is well formalized and the objective is to find S such that $F_{Y|\mathbf{X}=\mathbf{x}} \approx F_{Y|\mathbf{X}_S=\mathbf{x}_S}$. However, these methods are not reliable because they are prone to approximation errors due to the training of several Neural Networks, and they do not provide any guarantees regarding the fidelity of the explanations [Jethani, 2021]. A similar approach is also developed in [Dhurandhar, 2018] called Pertinent Positive.

Anchors [Ribeiro, 2018] are local rule-based explanations that propose a solution to the reliability issue by providing an explanation with guarantees. It explains individual predictions of any classification model by finding a decision rule that reaches a given accuracy for a high percentage of the neighborhood of the instance. However, the method is only available for classification, requires discretizing the variables, is unstable, and tends to use more variables than needed.

Logical Reasoning Approaches such as Sufficient Reasons [Shih, 2018; Darwiche, 2020] select a minimal subset of features guaranteeing that, no matter what is observed for the remaining features, the decision will stay the same. It can be seen as an instance-wise feature selection but with guarantees of sufficiency and minimality (i.e., no subset of the set satisfies the sufficiency condition). However, since the guarantees are deterministic, it is often necessary to include many features into the explanation, making the explanation more complex, and thus less intelligible. A relaxation of this method is Sufficient Explanations [Wang, 2020] that gives probabilistic guarantees instead of deterministic guarantees, i.e., it requires that the prediction remain the same with high probability. It gives a simple local explanation with guarantees while considering feature interactions and the data distribution. However, it is limited to classification with binary features and requires learning the distribution of the features. Moreover, the Sufficient Explanations are not unique, which causes a selection problem as the whole set of explanations is not interpretable.

In this work, we propose a consistent method that efficiently finds the Sufficient Explanations of any data generating process (\mathbf{X}, Y) or any model $(\mathbf{X}, f(\mathbf{X}))$, without learning the distribution of \mathbf{X} . In particular, we don't need to have access to the model f , we need only the predictions or outputs, contrary to [Wang, 2020]. We propose local attributions that summarize the diversity of

the Sufficient Explanations. In addition, we propose local rule-based explanations for regression and classification models based on Sufficient Explanations. To the best of our knowledge, it is the first local rule-based explanations for regression tasks.

3 Probabilistic Sufficient Explanations for Regression

Let assume we have an i.i.d. samples $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ such that $(\mathbf{X}, Y) \sim P = P_{\mathbf{X}}P_{Y|\mathbf{X}}$ where $\mathbf{X} \in \mathcal{X}$ and $Y \in \mathbb{R}$. We use $[p]$ to represent the indices of the features, and for a given subset $S \subseteq [p]$, $\mathbf{X}_S = (X_i)_{i \in S}$ represents a subgroup of features, and we write $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_{\bar{S}})$.

We define as the explanations of an instance (\mathbf{x}, y) the minimal subsets $\mathbf{x}_S, S \subseteq [p]$ such that given only those features, the model yields "almost" the same prediction y as on the complete example with high probability, under the data distribution. The main probabilistic reasoning tool that we use for our explanations is the Same Decision Probability (SDP) [Chen, 2012]. For classification, it is defined as the probability that the classifier has the same output by ignoring some variables. To also explain regression models, we propose the following definition of the SDP.

Definition 3.1. (Same Decision Probability of a regressor). Given an instance (\mathbf{x}, y) , the Same Decision Probability at level t of the subset $S \subseteq [p]$, w.r.t $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_{\bar{S}})$ is

$$SDP_S(y; \mathbf{x}, t) = \mathbb{P}\left((Y - y)^2 \leq t \mid \mathbf{X}_S = \mathbf{x}_S\right).$$

In a regression setting, the SDP gives the probability to stay close to the same prediction y at level t , when we fix $\mathbf{X}_S = \mathbf{x}_S$ or when $\mathbf{X}_{\bar{S}}$ are missing. The higher the probability, the better the explanation powered by S . Note that for classification, the SDP is defined as $SDP_S(y; \mathbf{x}) = \mathbb{P}(Y = y \mid \mathbf{X}_S = \mathbf{x}_S)$. Although we present all the methods with the SDP for regression, they remain the same for classification, we only need to replace $SDP_S(y; \mathbf{x}, t)$ by $SDP_S(y; \mathbf{x})$. Now, we focus on the minimal subset of features such that the model makes the same or almost the same decision with a given (high) probability π .

Definition 3.2. (Minimal Sufficient Explanations). Given an instance (\mathbf{x}, y) , $S_\pi(\mathbf{x})$ is a Sufficient Explanation for probability π , if $SDP_{S_\pi(\mathbf{x})}(y; \mathbf{x}, t) \geq \pi$, and no subset Z of $S_\pi(\mathbf{x})$ satisfies $SDP_Z(y; \mathbf{x}, t) \geq \pi$. Hence, a Minimal Sufficient Explanation is a Sufficient Explanation with minimal size.

For a given instance, the Sufficient Explanation or Minimal Sufficient Explanation may not be unique [Darwiche, 2020]. Furthermore, there may be significant differences among the Sufficient Explanations or Minimal Sufficient Explanations. We denote SE as the set of all Sufficient Explanations and M-SE as the set of Minimal Sufficient Explanations. Thus, the number and the diversity of the explanations make the method less intelligible, as deriving one of them is not informative enough, and all of them are too complex to interpret. Therefore, we propose to compute the following local attributions that summarize the importance of each variable in SE/M-SE:

Definition 3.3. (Local eXplanatory Importance - LXI). Given an instance (\mathbf{x}, y) and its SE or M-SE. The local explanatory importance of \mathbf{x}_i is how frequent \mathbf{x}_i is chosen in the SE or M-SE.

Contrary to classical local feature attributions such as SHAP or LIME, the values of Local eXplanatory Importance does not depend on the range of values of the predictions and are interpretable by design. It corresponds to the frequency of apparition in the SE or M-SE, which allows to reason about the relative difference between the attribution of each feature. Indeed, we can easily discriminate between the importance of variables in terms of probabilities compared to arbitrary values of SHAP or LIME that depend on the model and its predictions. In our framework, a value equal to 1 means that this feature is present in all the SE/M-SE. Hence this feature is necessary to maintain the prediction. Moreover, the attributions of the features are sparse since they are based on the SE/M-SE.

Although Sufficient Explanations allow finding local relevant variables, we may want to know the logical reasons relating input and output. In essence, explaining a decision means giving the reasons that highlight why the decision has been made. Therefore, we propose to extend the Sufficient Explanations into local rules. A rule is a simple IF-THEN statement, e.g., IF the conditions on the features are met, THEN make a specific prediction. Recall that given an instance \mathbf{x} , a Sufficient Explanation is the minimal subset $S \subseteq [p]$, such that fixing the values $\mathbf{X}_S = \mathbf{x}_S$ permits to maintain the prediction with high probability. The idea is to find the largest rectangle $L_S(\mathbf{x}) = \prod_{i=1}^{|S|} [a_i, b_i]$, $a_i, b_i \in \mathbb{R}$ given the indexes of the Sufficient Explanation S such that $\mathbf{x}_S \in L_S(\mathbf{x})$ and for all $\mathbf{z} \sim P_{\mathbf{X}}$ with $\mathbf{z}_S \in L_S(\mathbf{x})$, $SDP_S(y; \mathbf{z}, t) \geq \pi$.

Definition 3.4. (Minimal Sufficient Rule). Given an instance (\mathbf{x}, y) , S a Minimal Sufficient Explanation, the rectangle $L_S(\mathbf{x}) = \prod_{i=1}^{|S|} [a_i, b_i]$, $a_i, b_i \in \mathbb{R}$ is a Minimal Sufficient Rule if $L_S(\mathbf{x}) = \arg \max_L Vol(L)$, $\mathbf{x}_S \in L_S(\mathbf{x})$ and for all $\mathbf{z} \sim P_{\mathbf{X}}$ with $\mathbf{z}_S \in L_S(\mathbf{x})$, $SDP_S(y; \mathbf{z}, t) \geq \pi$.

Intuitively, the Sufficient Rule is a generalization of the Sufficient Explanation, i.e., instead of satisfying the minimality/sufficiency conditions of Definition 3.2 if we fixed the values $\mathbf{X}_S = \mathbf{x}_S$, we want to satisfy these conditions on all the elements of a rectangle $L_S(\mathbf{x})$ that contains \mathbf{x}_S . We also want this rectangle to be of maximal volume such that it covers a large part of the input space. Thus, the Sufficient Rule captures the local behavior of the model around \mathbf{x} while ensuring the minimality of the rule and guarantees on the outcome. Note that the volume of the rectangle L can be defined as $Vol(L) = \mathbb{P}(\mathbf{X}_S \in L)$ or $\lambda(L)$, with λ the Lebesgue measure.

While Sufficient Rules are similar to Anchors introduced by [Ribeiro, 2018], we emphasize two major types of differences. The first is that our framework for constructing rules can address regression problems, deal with continuous features without discretization, and do not need access to the model f . Moreover, if we have a model f and an instance \mathbf{x} , Anchors search the largest rule (or rectangle) $L_S(\mathbf{x})$ such that $\mathbb{P}_Q(f(\mathbf{x}) = Y | \mathbf{X}_S \in L_S(\mathbf{x})) \geq \pi$ under an instrumental distribution Q , typically the marginal law $Q = \prod_i P_{X_i}$. This is different from the Sufficient Rule that requires the stability of the prediction for all the observations in the rectangle, i.e, for all $\mathbf{x}_S \in L_S(\mathbf{x})$, $\mathbb{P}(f(\mathbf{x}) = Y | \mathbf{X}_S = \mathbf{x}_S) \geq \pi$. The second major difference is that the Sufficient

Rule is based on the original distribution P as we use the conditional distribution $Y|\mathbf{X}_S$. In contrast, anchors use local sampling perturbations (introducing another distribution Q). As we discuss in the next section, the effective computation of these rules is very different. Anchors use a heuristic approach to find the minimal rule, which might produce instable and suboptimal minimal rules. The Sufficient Rules satisfy a minimality principle by definition, as they are based on Sufficient Explanations.

4 SDP, Sufficient Explanations and Sufficient Rules via Random Forest

In order to find the Sufficient Explanations $S_\pi(\mathbf{x})$ and the corresponding Sufficient Rules $L_{S_\pi}(\mathbf{x})$, we need to compute the SDP for any subset S . However, the computation of the SDP is known to be computationally hard; even for simple Naive Bayes model, the computation of SDP is NP-hard [Chen, 2013]. Consequently, approximate criteria based on expectations instead of probabilities have been introduced by [Wang, 2020]. They proposed to use a Probabilistic Circuit [Choi, 2020] to model the distribution of the features \mathbf{X} and compute a lower bound of the SDP.

In this section, we propose a consistent estimator of the SDP for any distribution (\mathbf{X}, Y) . It is based on two ideas: Projected Forest [Bénard, 2021b; Bénard, 2021e] and Quantile Regression Forest [Meinshausen, 2006]. The Projected Forest is an adaptation of the Random Forest algorithm that estimates $\mathbb{E}[Y|\mathbf{X}_S = \mathbf{x}_S]$ instead of $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$, and the Quantile Regression Forest uses the Random Forest algorithm to estimate the Conditional Distribution Function (CDF) $\mathbb{P}(Y \leq y|\mathbf{X} = \mathbf{x})$. The first step is to write the SDP as

$$SDP_S(y; \mathbf{x}, t) = \mathbb{P}((Y - y)^2 \leq t | \mathbf{X}_S = \mathbf{x}_S) = F_S(y + \sqrt{t} | \mathbf{X}_S = \mathbf{x}_S) - F_S(y - \sqrt{t} | \mathbf{X}_S = \mathbf{x}_S). \quad (4.1)$$

Equation (4.1) demonstrates that the primary challenge lies in estimating the Projected (or Conditional) Cumulative Distribution Function (CDF) $F_S(y|\mathbf{X}_S = \mathbf{x}_S) = \mathbb{P}(Y \leq y|\mathbf{X}_S = \mathbf{x}_S)$. The variation of the original Random Forest suggested by [Meinshausen, 2006], which estimates the CDF $F(y|\mathbf{X} = \mathbf{x}) = \mathbb{P}(Y \leq y|\mathbf{X} = \mathbf{x})$, is not directly applicable to our objective as we aim to estimate the Projected CDF $F_S(y|\mathbf{X}_S = \mathbf{x}_S)$ for any S . The recent works of [Bénard, 2021b; Bénard, 2021e] are more relevant as they permit the estimation of $\mathbb{E}[Y|\mathbf{X}_S = \mathbf{x}_S]$ from a classical Random Forest trained to predict $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. The idea is to extract a new Forest called Projected Forest from the original Forest, which is a projection of the original Forest along the S -direction.

We propose to combine the ideas of Quantile Regression Forest and Projected Forest to estimate the Projected CDF $F_S(y|\mathbf{X}_S = \mathbf{x}_S)$. In addition, we establish the consistency of this estimator.

4.1 Random Forest and Condition Distribution Function (CDF) Forest

A Random Forest (RF) is an ensemble of k randomized decision trees based on the CART procedure [Breiman, 1984]. The algorithm works as follows. For each tree, data points are drawn at random with replacement from the original data set of size n ; then, at each cell, a split variable is chosen by maximizing the CART criterion among a random subset of variables; finally, the construction of every tree is stopped when the number of observations in each leaf reaches a predefined value. For each new instance \mathbf{x} , the prediction of the l -th tree is:

$$m_l(\mathbf{x}; \Theta_l, \mathcal{D}_n) = \sum_{i=1}^n \frac{B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}; \Theta_l)} Y_i}{N_n(\mathbf{x}; \Theta_l)} \quad (4.2)$$

- $\Theta_l, l = 1, \dots, k$ are independent random vectors, distributed as a generic random vector $\Theta = (\Theta^1, \Theta^2)$ independent of \mathcal{D}_n . Θ^1 contains indices of the bootstrap samples used to build the tree, and Θ^2 contains the splitting candidate variables at each node.
- $A_n(\mathbf{x}; \Theta_l)$ is the leaf node containing \mathbf{x}
- $N_n(\mathbf{x}; \Theta_l)$ is the number of bootstrap elements that fall into $A_n(\mathbf{x}; \Theta_l)$
- $B_n(\mathbf{X}_i; \Theta_l)$ is the bootstrap component i.e., the number of times the observation \mathbf{X}_i has been chosen from the original data for the l -th tree.

The trees are then averaged to give the prediction of the forest as:

$$m(\mathbf{x}; \Theta_{1:k}, \mathcal{D}_n) = \frac{1}{k} \sum_{l=1}^k m_l(\mathbf{x}; \Theta_l, \mathcal{D}_n). \quad (4.3)$$

The Random Forest estimator (Eq. 4.3) can also be seen as an adaptive neighborhood procedure [Lin, 2006; Biau, 2010] or kernel methods [Breiman, 2000; Geurts, 2006; Scornet, 2016]. For every instance \mathbf{x} , the observations in \mathcal{D}_n are weighted by $w_{n,i}(\mathbf{x})$, $i = 1, \dots, n$. The prediction of Random Forests and the weights can be rewritten as

$$m(\mathbf{x}; \Theta_{1:k}, \mathcal{D}_n) = \sum_{i=1}^n w_{n,i}(\mathbf{x}) Y_i, \quad w_{n,i}(\mathbf{x}) = \frac{1}{k} \sum_{l=1}^k \frac{B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}; \Theta_l)}}{N_n(\mathbf{x}; \Theta_l)}.$$

Viewing a Random Forest as an adaptive nearest neighbor predictor offers natural estimates of more complex quantities such as cumulative hazard function [Ishwaran, 2008], treatment effect [Wager, 2017; Jocteur, 2023], and conditional density [Du, 2021]. Therefore, just as $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ is approximated by a weighted mean over the outputs Y_i , $\mathbb{E}[\mathbb{1}_{Y \leq y} | \mathbf{X} = \mathbf{x}]$ is approximated by the weighted mean over the $\mathbb{1}_{Y_i \leq y}$ using the same weights $w_{n,i}(\mathbf{x})$. The approximation is

$$\widehat{F}(y|\mathbf{X} = \mathbf{x}; \Theta_{1:k}, \mathcal{D}_n) = \sum_{i=1}^n w_{n,i}(\mathbf{x}) \mathbb{1}_{Y_i \leq y}. \quad (4.4)$$

To simplify notations, we omit $\Theta_1, \dots, \Theta_k, \mathcal{D}_n$ and we write $\widehat{F}(y|\mathbf{X} = \mathbf{x})$.

4.2 Projected Forest and Projected CDF Forest

We describe the Projected Forest (PRF) and show how we combined it with the Quantile Regression Forest to build the estimator of the Projected CDF. The PRF algorithm has been introduced in [Bénard, 2021e; Bénard, 2021b]. The idea is to project the partition of each tree of the forest on the subspace spanned by the variables in S , thus we can estimate $\mathbb{E}[Y|\mathbf{X}_S]$ rather than $\mathbb{E}[Y|\mathbf{X}]$. The computation of these partitions for each S can be computationally expensive in high dimension. However, [Bénard, 2021b] proposes a simple algorithm to efficiently derive the output of the Projected Forest without explicitly computing its partitions. To compute the prediction of a tree projected along the S direction, the algorithm ignores splits that use variables that are not contained in S . It works as follows: when an observation is dropped down in the tree, and it encounters a split involving a variable $i \notin S$, the observation is sent both to the left and right children nodes. As a result, each observation falls in multiple terminal leaves of the tree. Thus, to compute the prediction of an instance \mathbf{x}_S , we collect the set of terminal leaves where it falls, and average the output Y_i of the training observations which belong to every terminal leaf of this collection. $\mathbb{E}[Y|\mathbf{X}_S = \mathbf{x}_S]$ is estimated as the average outputs of the training observations in the intersection of the leaves where \mathbf{x}_S falls.

The PRF algorithm is detailed in the Appendix (12) and the corresponding PRF is denoted $m^{(S)}(\mathbf{x}_S) = \sum_{i=1}^n w_{n,i}(\mathbf{x}_S)Y^i$ where the weights are defined by

$$w_{n,i}(\mathbf{x}_S) = \frac{1}{k} \sum_{l=1}^k \frac{B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{\mathbf{X}_i \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)}}{N_n^{(S)}(\mathbf{x}; \Theta_l)}, \quad (4.5)$$

where $A_n^{(S)}(\mathbf{x}_S; \Theta_l)$ is the leaf of the projected l -th tree given S where \mathbf{x}_S falls and $N_n^{(S)}(\mathbf{x}; \Theta_l)$ denoted the number of bootstrap observations that falls in $A_n^{(S)}(\mathbf{x}_S; \Theta_l)$. Consequently, we approximate the Projected CDF $F_S(y|\mathbf{X}_S = \mathbf{x}_S) = \mathbb{P}(Y \leq y|\mathbf{X}_S = \mathbf{x}_S)$ as in Eq. (4.4) by using the weights of the Projected Forest defined in Eq. (4.5). The estimator of the Projected CDF is defined as $\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S) = \sum_{i=1}^n w_{n,i}(\mathbf{x}_S)\mathbb{1}_{Y_i \leq y}$.

4.3 Consistency of the Projected CDF Forest

In this section, we state our main result, which is the uniform a.s. convergence of the estimator $\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S)$ to $F_S(y|\mathbf{X}_S = \mathbf{x}_S)$. [Meinshausen, 2006] showed the uniform convergence in probability of a simplified version of the estimator of the CDF defined in Eq. (4.4), where the weights $w_{n,i}(\mathbf{x}_S)$ are in fact considered to be non-random while they are indeed random variables depending on $(\Theta_l)_{l=1,\dots,k}, \mathcal{D}_n$. Moreover, the bootstrap step was replaced by subsampling without replacement as in most studies that analyze the asymptotic properties of Random Forests [Scornet, 2015; Wager, 2017; Goehry, 2020]. However, [Elie-Dit-Cosaque, 2022] showed the almost surely uniform convergence of both estimators (the simplified and the one defined in Eq. 4.4) under mild assumptions with all the randomness and bootstrap samples. We follow their works to prove the consistency of the PRF CDF $\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S)$ based on the following assumptions.

Assumption 4.1. For all $x \in \mathbb{R}^d$, the conditional cumulative distribution function $F(y|X = x)$ is continuous.

Assumption 4.1 is necessary to get uniform convergence of the estimator.

Assumption 4.2. For $l \in [k]$, we assume that the variation of the conditional cumulative distribution function within any cell goes to 0.

$$\forall x \in \mathbb{R}^d, \forall y \in \mathbb{R}, \sup_{z \in A_n(x; \Theta_l)} |F(y|z) - F(y|x)| \xrightarrow[n \rightarrow +\infty]{a.s.} 0$$

Assumption 4.2 allows to control the approximation error of the estimator. If for all y , $F(y|\cdot)$ is continuous, Assumption 4.2 is satisfied provided that the diameter of the cell goes to zero. Note that the vanishing of the diameter of the cell is a common condition used to prove the consistency of general partitioning estimator (see chapter 4 in [Györfi, 2002]). [Scornet, 2015] show that this is true when the data come from additive regression models [Stone, 1985b], and [Elie-Dit-Cosaque, 2022] show that it holds for a more general class, such as product functions or sums of product functions. The result is also valid for all regression functions, with a slightly modified version of RF, where each child node contains at least a small fraction of the observations in the parent node, and the probability that each variable $j = 1, \dots, p$ is chosen for the split is positive in each node. Under these small modifications, Lemma 2 from [Meinshausen, 2006] gives that the diameter of each leaf node vanishes.

Assumption 4.3. Let k the number of trees and $N_n(\mathbf{x}; \Theta_l)$ number of bootstrap observations in a leaf node, and assume that $k = \mathcal{O}(n^\alpha)$, with $\alpha > 0$, and for all $\mathbf{x} \in \mathbb{R}^d$, $N_n(\mathbf{x}; \Theta_l) = \Omega^1(\sqrt{n}(\ln(n))^\beta)$, with $\beta > 1$ a.s.

Assumption 4.3 allows us to control the estimation error and means that the cells should contain a sufficiently large number of points so that averaging among the observations is effective.

To prove the consistency of the PRF CDF $\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S)$, we only need to verify the assumptions 4.1, 4.2, 4.3 on the parameters of the Projected Forest and the Projected CDF $F_S(y|\mathbf{X}_S = \mathbf{x}_S) = \mathbb{P}(Y \leq y|\mathbf{X}_S = \mathbf{x}_S)$.

Assumptions 4.1 and 4.2 are satisfied for the Projected CDF and the PRF Forest's leaves. Since by definition $A_n^{(S)}(\mathbf{x}_S; \Theta_l)$ is included in $A_n(\mathbf{x}; \Theta_l)$, if the diameter goes to zero within the cells of the RF, it also vanishes in the Projected Forest. In addition, if the CDF $F(y|\mathbf{X} = \mathbf{x}) = F(y|\mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}})$ is continuous with respect to \mathbf{x} , an analysis of parameter-dependent integral shows that the Projected CDF $F_S(y|\mathbf{X}_S = \mathbf{x}_S) = \int F(y|\mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}})\mathbb{P}(\mathbf{x}_{\bar{S}}|\mathbf{x}_S)d\mathbf{x}_{\bar{S}}$ is also continuous. As we control the minimal number of observations in the leaf of the Projected Forest by construction, Assumption 4.3 is also verified. Then, the PRF CDF satisfies also Assumption 4.1-4.3 which ensures its consistency thanks to Theorem 4.4.

¹ $f(n) = \Omega(g(n)) \iff \exists c > 0, \exists n_0 > 0 \mid \forall n \geq n_0, |f(n)| \geq c|g(n)|$

Theorem 4.4. Consider a RF satisfying Assumptions 4.1 to 4.3. Then,

$$\forall \mathbf{x} \in \mathbb{R}^d, \sup_{y \in \mathbb{R}} |\widehat{F}_S(y | \mathbf{X}_S = \mathbf{x}_S) - F_S(y | \mathbf{X}_S = \mathbf{x}_S)| \xrightarrow[n \rightarrow +\infty]{a.s.} 0$$

The proof and convergence simulations are detailed in the Appendix (8).

4.4 Estimation of SDP, Sufficient Explanations and Sufficient Rules

In this section, we show how we compute the SDP, Sufficient Explanations, and Sufficient Rules using the PRF CDF estimator. We derive from the previous section the following consistent estimator of any $SDP_S(y; \mathbf{x}, t)$:

$$\widehat{SDP}_S = \widehat{F}_S(y + \sqrt{t} | \mathbf{X}_S = \mathbf{x}_S) - \widehat{F}_S(y - \sqrt{t} | \mathbf{X}_S = \mathbf{x}_S).$$

Sufficient Explanations. Finding the SE/M-SE using a greedy algorithm is computationally hard, since the number of subsets is exponential. Therefore, we propose to reduce the number of variables by focusing only on the most influential variables. We search the Sufficient Explanations in the subspace of the $s = 10$ variables frequently selected in the RF used to estimate the SDP, reducing the complexity from 2^p to 2^s . This preselection procedure is already used in [Strobl, 2007; B enard, 2021d; B enard, 2021b], and it is mainly based on Proposition 1 of [Scornet, 2015], which highlights the fact that RF naturally splits the most influential variables. However, any RF’s importance measure such as Sobol-MDA [B enard, 2021e] can be used as the RF algorithm is known to adapt to the intrinsic dimension [Scornet, 2015; Klusowski, 2020]. Therefore, the choice of s is directly driven by the computation power available to explore the subsets. In practice, we have always found Sufficient Explanations with a probability above $\pi = 0.9$ with $s = 10$ for many real-world datasets. Instead of exhaustively searching through all 2^s possible subsets, an alternative approach would involve sampling a sufficient number of subsets, usually a few thousand, which are present in the decision paths of the trees of the Projected Forest. Due to their inherent construction, these subsets are likely to include influential variables and interactions. This strategy was employed in [Basu, 2018; B enard, 2021b].

Sufficient Rules. We utilized the SDP’s estimator \widehat{SDP}_S to identify the Sufficient Rules. As the PRF CDF estimator is a tree-based model, \widehat{SDP}_S also partitions the input space in a manner similar to a tree or a Random Forest. This allows us to avoid discretizing the input space to find the rule satisfying Definition (3.4). Instead, we only need to locate the leaves that are compatible with the conditions of the Sufficient Rule defined in (3.4).

Given a Minimal Sufficient Explanation S of an instance (\mathbf{x}, y) at level π , we already have a rectangle $L_S(\mathbf{x})$ defined by the PRF CDF or \widehat{SDP}_S that is the largest rectangle such that $\mathbf{x}_S \in L_S(\mathbf{x})$ and for all \mathbf{z} with $\mathbf{z}_S \in L_S(\mathbf{x})$, $\widehat{SDP}_S(y; \mathbf{z}, t) = \widehat{SDP}_S(y; \mathbf{x}, t) \geq \pi$. By definition, it is the intersection of the leaves of the trees where \mathbf{x}_S falls, namely $\cap_{l=1}^k A_n^{(S)}(\mathbf{x}_S; \Theta_l)$. Thus, starting from $\cap_{l=1}^k A_n^{(S)}(\mathbf{x}_S; \Theta_l)$, which is also a leaf of the Projected Forest, we only need to find all the neighboring leaf that can be merged with it to obtain the rule satisfying (3.4).

An exhaustive search for compatible leaves is not feasible due to the exponential number of leaves that a forest can possess. The number of leaves of a forest with k trees is bound by $t_n^1 \times \dots \times t_n^k$, where $t_n^l, l \in [k]$ is the number of leaves of the l -th tree, as each input reaches exactly one leaf in each tree. However, our focus lies on nonempty cells - leaves that contain at least one training observation. Consequently, we can give each training sample of \mathcal{D}_n to the forest and determine the corresponding leaf it reaches in each tree. The intersection of these leaves corresponds to a leaf of the forest. As multiple observations can fall into the same leaf, we can have a maximum of n nonempty regions. We can further reduce the search space by only considering the leaves of observations whose outputs are close to our target y . Then, we merge the compatible leaves that satisfy Definition (3.4) with $\cap_{l=1}^k A_n^{(S)}(\mathbf{x}_S; \Theta_l)$ to find the Sufficient Rule. The utilization of the leaves of the Projected Forest to identify the Sufficient Rule is a significant advantage of our method. Discovering meaningful rules, especially in high-dimensional spaces, is a challenging task. However, by leveraging the partition already learned by the PRF CDF, we simplify the problem by directly focusing on the regions or leaves of PRF CDF where the observations are located.

Nevertheless, the merging step poses a significant challenge. It involves identifying the largest hyperrectangle in a collection of hyperrectangles, which may not necessarily be connected or convex. This task is closely related to a well-known computational geometric problem that consists of finding the minimal area axis-aligned rectangle that can encompass a given set of points [Kaplan, 2019; Chan, 2021; Lin, 2018; Aggarwal, 1991; Eppstein, 1994; Datta, 1995]. An approximate solution to the Sufficient Rule entails locating the smallest hyperrectangle that contains all the training points within the set of compatible leaves that meet the Definition (3.4).

A well-known variant of this problem is the largest empty hyperrectangle problem [Chan, 2023; Abo-Alsabeh, 2023; Naamad, 1984]. Given a hyperrectangle R containing a set of points, the objective is to find the largest hyperrectangle that can be embedded in R without containing any of the points. While this problem has been extensively explored in lower dimensions, it remains unsolvable in higher dimensions. Finding holes in one-dimensional datasets is an easy task, and faster algorithms has been proposed for 2, and 3 dimensions [Aggarwal, 1987; Datta, 2000]. However, the problem turns out to be NP-hard when the dimension is above 4 [Backer, 2009; Eckstein, 2002; Dumitrescu, 2013]. Consequently, we adopt an approximating procedure using a Monte-Carlo based algorithm, specifically the simulated annealing algorithm, to compute the Sufficient Rule. We leave the investigation of better algorithms for future work.

Let's define the set of compatible leaves as $C = \{(\mathbf{l}_1, \mathbf{u}_1), \dots, (\mathbf{l}_m, \mathbf{u}_m)\}$ where the tuple $(\mathbf{l}_m, \mathbf{u}_m) \in \mathbb{R}^p \times \mathbb{R}^p$ represent the lower and upper bound of the m -th compatible leaf. Hence, an observation \mathbf{x} falls into compatible leaf m if it satisfies $\mathbf{l}_m \leq \mathbf{x} \leq \mathbf{u}_m$ elementwise, meaning $l_{m,j} \leq x_j \leq u_{m,j}$ for all $j \in \{1, \dots, p\}$. Similarly, we represent $\cap_{l=1}^k A_n^{(S)}(\mathbf{x}_S; \Theta_l)$ as $(\mathbf{l}^{inters}, \mathbf{u}^{inters})$. We denote $vol(\mathbf{l}, \mathbf{u})$ as the measure of a rule, it is either the number of observations that falls into the rule or the Lebesgue measure. Algorithm 2 describes the merging step using simulated annealing to find the maximal volume hyperrectangle approximating the Sufficient Rule, and the proposal distribution of the simulated annealing is presented in Algorithm 3.

Algorithm 2: Merging step to generate Sufficient Rule

Input : $(\mathbf{l}^{inters}, \mathbf{u}^{inters}) = \cap_{l=1}^k A_n^{(S)}(\mathbf{x}_S; \Theta_l)$, $C = \{(\mathbf{l}_1, \mathbf{u}_1), \dots, (\mathbf{l}_m, \mathbf{u}_m)\}$ set of compatible leaves, \mathcal{D}_n training data set, \mathcal{D}^{in} observations of \mathcal{D}_n that fall into one of the leaves in C , $vol(\mathbf{l}, \mathbf{u})$ represents the measure of a rule (\mathbf{l}, \mathbf{u}) , $maxIter$ max number of iterations, $p \in (0, 1)$ probability of proposing a rule larger than $(\mathbf{l}^{current}, \mathbf{u}^{current})$ otherwise larger than $(\mathbf{l}^{inters}, \mathbf{u}^{inters})$, T initial temperature, r cooling rate.

Output: Approximation of the largest hyperrectangle $(\mathbf{l}^{best}, \mathbf{u}^{best})$ contained within the union of rectangles in C and that contains $(\mathbf{l}^{inters}, \mathbf{u}^{inters})$

```
1: Initialize  $(\mathbf{l}^{current}, \mathbf{u}^{current}) \leftarrow (\mathbf{l}^{inters}, \mathbf{u}^{inters})$ ,  $(\mathbf{l}^{best}, \mathbf{u}^{best}) \leftarrow (\mathbf{l}^{inters}, \mathbf{u}^{inters})$ 
2: for  $i$  from 1 to  $maxIter$  do
3:    $(\mathbf{l}^{gen}, \mathbf{u}^{gen}) = \text{GenerateRule}(\mathbf{l}^{current}, \mathbf{u}^{current}, \mathbf{l}^{inters}, \mathbf{u}^{inters}, C, \mathcal{D}_n, \mathcal{D}^{in}, maxIter, p)$ ;
   /* Generate a valid rule larger than  $(\mathbf{l}^{current}, \mathbf{u}^{current})$  with probability  $p$ .
   Otherwise, generate a valid rule larger than  $(\mathbf{l}^{inters}, \mathbf{u}^{inters})$ . */
4:   Compute the volume difference  $\Delta V = vol(\mathbf{l}^{current}, \mathbf{u}^{current}) - vol(\mathbf{l}^{gen}, \mathbf{u}^{gen})$ 
5:   if  $\Delta V < 0$  or  $exp(-\Delta V/T) > random(0, 1)$  then
6:     Set  $(\mathbf{l}^{current}, \mathbf{u}^{current}) \leftarrow (\mathbf{l}^{gen}, \mathbf{u}^{gen})$ ; /* Accept the new hyperrectangle */
7:     if  $vol(\mathbf{l}^{best}, \mathbf{u}^{best}) < vol(\mathbf{l}^{current}, \mathbf{u}^{current})$  then
8:       Set  $(\mathbf{l}^{best}, \mathbf{u}^{best}) \leftarrow (\mathbf{l}^{current}, \mathbf{u}^{current})$ ; /* Update the best hyperrectangle */
9:     Decrease  $T$  by  $T = T * r$ ; /* Cooling step */
10: return  $(\mathbf{l}^{best}, \mathbf{u}^{best})$ 
```

Algorithm 3: GenerateRule($\mathbf{l}^{current}, \mathbf{u}^{current}, C, \mathcal{D}_n, \mathcal{D}^{in}, maxIter, p$)

Input : $(\mathbf{l}^{current}, \mathbf{u}^{current})$, $C = \{(\mathbf{l}_1, \mathbf{u}_1), \dots, (\mathbf{l}_m, \mathbf{u}_m)\}$ set of compatible leaves, \mathcal{D}_n training data set, \mathcal{D}^{in} observations of \mathcal{D}_n that fall into one of the leaves in C , $maxIter$ max number of iterations, $p \in (0, 1)$ probability of proposing a rule larger than $(\mathbf{l}^{current}, \mathbf{u}^{current})$ otherwise larger than $(\mathbf{l}^{inters}, \mathbf{u}^{inters})$

Output: New valid hyperrectangle $(\mathbf{l}^{prop}, \mathbf{u}^{prop})$

```
1: Initialize  $(\mathbf{l}^{gen}, \mathbf{u}^{gen}) \leftarrow (\mathbf{l}^{current}, \mathbf{u}^{current})$ ,  $(\mathbf{l}^{prop}, \mathbf{u}^{prop}) \leftarrow (\mathbf{l}^{current}, \mathbf{u}^{current})$ ,
    $found = \text{False}$ ,  $it = 0$ 
2: while not  $found$  and  $it \leq maxIter$  do
3:   if  $random(0, 1) \leq p$  then
4:     for  $j$  from 1 to  $length(\mathbf{l}^{gen})$  do
5:        $l_j^{gen} \leftarrow \text{sample uniformly from the set } \{l_{i,j} : i = 1, \dots, m \text{ and } l_{i,j} \leq l_j^{current}\}$ 
6:        $u_j^{gen} \leftarrow \text{sample uniformly from the set } \{u_{i,j} : i = 1, \dots, m \text{ and } u_{i,j} \geq u_j^{current}\}$ 
7:     else
8:       for  $j$  from 1 to  $length(\mathbf{l}^{gen})$  do
9:          $l_j^{gen} \leftarrow \text{sample uniformly from the set } \{l_{i,j} : i = 1, \dots, m \text{ and } l_{i,j} \leq l_j^{inters}\}$ 
10:         $u_j^{gen} \leftarrow \text{sample uniformly from the set } \{u_{i,j} : i = 1, \dots, m \text{ and } u_{i,j} \geq u_j^{inters}\}$ 
11:       if no observation of  $\mathcal{D}_n \setminus \mathcal{D}^{in}$  falls in the rule  $(\mathbf{l}^{gen}, \mathbf{u}^{gen})$  then
12:          $found = \text{True}$ 
13:          $(\mathbf{l}^{prop}, \mathbf{u}^{prop}) = (\mathbf{l}^{gen}, \mathbf{u}^{gen})$ 
14:          $it = it + 1$ 
15: return  $(\mathbf{l}^{prop}, \mathbf{u}^{prop})$ 
```

How to choose the hyperparameters. The main hyperparameters are: π the minimal probability of changing the decision and t which corresponds to the radius of the region center at the prediction in the definition (3.1) of the SDP for regression problems.

We propose choosing $\pi = 0.9$ as it is an acceptable level of risk, but the user can increase or decrease this probability depending on the use case.

The hyperparameter most challenging to choose is t ; we recommend having an adaptive radius $t(\mathbf{x})$ using the quantile of the conditional distributions $Y|\mathbf{X} = \mathbf{x}$, which is a by-product of the Quantile Regression Forest used for computing the SDP. For each observation, we choose the region $t(\mathbf{x}) = [\hat{q}_{\alpha_1}(\mathbf{x}), \hat{q}_{1-\alpha_2}(\mathbf{x})]$ with $\alpha_1 + \alpha_2 = \alpha$ where $\hat{q}_{1-\alpha_2}$ is an estimator of the α -quantile of $Y|\mathbf{X} = \mathbf{x}$ using the Quantile Regression Forest. This allows us to construct a predictive interval with varying lengths but consistent confidence level. This positions our approach as a natural extension of the SDP in the classification case, focusing on the model’s uncertainty rather than predictions for the explanations. In that case, the Sufficient Explanation should be read: "if $\mathbf{X}_S = \mathbf{x}_S$ is fixed, then there is a probability at least π of having the same uncertainty as with all the features, or so that $Y \in [\hat{q}_{\alpha_1}(\mathbf{x}), \hat{q}_{1-\alpha_2}(\mathbf{x})]$ ". For this reason, we suggest fixing $\alpha_1 + \alpha_2 = \alpha$ and π at standard level $1 - \alpha = \pi = 0.9$ agreeing with acceptable level of risks.

5 Experiments

We conduct three experiments in this section. The first compares the Sufficient Explanations, Sufficient Rules and Local eXplanatory Importance (LXI) with state-of-art (SOTA) local explanations methods (SHAP, LIME, INVASE) in a simple high-dimensional regression model with small relevant features. Although this model is simple, SOTA (SHAP, LIME) have been shown to poorly detect the important variables of this model [Amoukou, 2021b; Ghalebikesabi, 2021] or see Chapter 3. Then, we analyze the performance of the Sufficient Rules in a real-world regression problem. Finally, we highlight the advantages of the Sufficient Rules in comparison with Anchors in real-world classification datasets. More experiments can be found in Appendix(10).

To effectively compare different explanation methods, we use synthetic data since the ground truth is required. We use the following synthetic model: we have $\mathbf{X} \in \mathbb{R}^p$, $\mathbf{X} \in \mathcal{N}(0, \Sigma)$, $\Sigma = 0.8J_p + 5I_p$ with $p = 100$, I_p is the identity matrix, J_p is all-ones matrix and a piece-wise linear predictor defined as:

$$Y = (X_1 + X_2)\mathbb{1}_{X_5 \leq 0} + (X_3 + X_4)\mathbb{1}_{X_5 > 0}. \quad (4.6)$$

The variables X_i for $i = 6 \dots 100$ are noise variables. We fit a RF with a data set of size $n = 10^4$, $k = 20$ trees and the minimal number of samples by leaf node is set to $\lfloor \sqrt{n} \times \ln(n)^{1.5} / 250 \rfloor$ for the original and the Projected Forest. The $R^2 = 99\%$ on the test set of size 10^4 . The RF is used to compute the explanations of SHAP, LIME. The Projected Forest is also extracted from the RF for the SDP approaches. We choose $\alpha_1 = 0.05, \alpha_2 = 0.95$ and $\pi = 0.90$. For INVASE, we use Neural Networks with 3 hidden layers for the selector model and the predictor model

as in [Yoon, 2018]. Notice that for SHAP, LIME and the SDP approaches, we used the same information (the learned RF) to retrieve the true explanation of the data. The performance of INVASE and the RF is the same, both model perfectly fit the data with a $R^2 = 99\%$.

SDP approaches vs SOTA (SHAP, LIME, INVASE) on regression. Here, we analyze the capacity of each method to discover the local important variables of the model defined in Eq. (4.6). Indeed, Eq. (4.6) shows that if $x_5 \leq 0$, the model uses only the variables x_1, x_2 otherwise it uses the variables x_3, x_4 . Thus, we try to find the top $K = 3$ relevant features for each sample. Note that K is not a required input for SDP and INVASE, but K must be given for SHAP and LIME. We select the top K variables that have the highest absolute values for SHAP and LIME. We use the True Discovery Rate (TDR) (higher is better) and False Discovery Rate (FDR) (lower is better) to measure the performance of the methods on discovery (i.e., discovering which features are relevant). In addition, as one of the objectives of each method is to find the minimal subset \mathbf{x}_S that is relevant for the corresponding target y , we also compute predictive performance metrics that show how well the projected predictor $\mathbb{E}[Y|\mathbf{X}_S = \mathbf{x}_S]$ selected by each method is close to the predictor on the full set of features $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$, under the data distribution. Formally, we denote it as P-MSE = $\mathbb{E}_{\mathbf{Z}} \left[\left(\mathbb{E}[Y|\mathbf{X} = \mathbf{Z}] - \mathbb{E}[Y|\mathbf{X}_S = \mathbf{Z}_{S(\mathbf{Z})}] \right)^2 \right]$ where $\mathbf{Z} \sim P_{\mathbf{X}}$ and $S(\mathbf{Z})$ is the Sufficient Explanation of \mathbf{Z} .

Table 4.1: Discovery metrics of Sufficient Explanation, INVASE, SHAP, LIME.

Methods	TDR	FDR	P-MSE
Sufficient Explanation	100%	2%	0.02
INVASE	99%	87%	0.006
SHAP	73%	27%	0.79
LIME	50%	49%	5.01

In table 4.1, we observe that the Sufficient Explanation succeeds to find the top K relevant variables and outperform the other methods by a significant margin. SHAP and LIME obtain the worst discovery rate. INVASE succeeds in finding relevant variables, but has a high FDR (87%), which means that we cannot distinguish between relevant and irrelevant variables since 87% of the selected variables are irrelevant. We also see that the P-MSE of INVASE is the lowest, which is not surprising as it selects all relevant variables despite its high FDR. Indeed, this metric is not much affected by the FDR. The P-MSE of Sufficient Explanations is also almost zero, and as above, SHAP and LIME perform worse than the other methods. In Table 4.2, we compare the LXI and SHAP values on 1000 observations having $x_5 > 0$. We compare the mean absolute values of SHAP and average LXI on this sub-population. Notice that on this model, these observations have a single Sufficient Explanation which is the variables X_3, X_4, X_5 . Both models give null attributions to the noise variables, but SHAP gives higher importance to the variables X_1, X_2 than the truly important variables X_3, X_4, X_5 . On the other hand, LXI gives non null attributions only on the important variables. We refer to the Appendix (10.2) for an additional comparison with SHAP in a case where there are several Sufficient Explanations.

Table 4.2: Global SHAP values (mean absolute) and average LXI on 1000 observations of the test set having $x_5 > 0$. X_{noises} corresponds to the sum of the attributions of the noises variables (X_i for $i = 6 \dots 100$).

Methods	X_1	X_2	X_3	X_4	X_5	X_{noises}
LXI	0	0	1	1	1	0
SHAP	1.47	1.54	0.56	0.56	0.86	0.005

However, even if the Sufficient Explanation find effectively the top K relevant variables, it cannot provide a complete understanding of the local behavior of the regression model (the SOTA methods can't do it either), i.e., that it's the sign of x_5 that matters. Thus, by extending the Sufficient Explanation into Sufficient Rule we can retrieve the complete story. We choose an observation (\mathbf{x}, y) such that its Sufficient Explanation found is $S = [3, 4, 5]$, with $\mathbf{x}_S = [-3.64, -4.41, 0.68]$. Although the Sufficient Explanation shows that fixing the value \mathbf{x}_S permit to maintain the prediction with high probability, the Sufficient Rule gives the additional information that we can also maintain the prediction within a small radius around y by satisfying the rule $L_S(\mathbf{x}) = \{X_5 > 0 \text{ AND } -4.45 \leq X_4 \leq -4.06 \text{ AND } -3.67 \leq X_3 \leq -3.58\}$. The Sufficient Rule $L_S(\mathbf{x})$ catches perfectly the local behaviour of the model which says that despite the values of x_3, x_4 , it's the sign of x_5 that matters.

SDP approaches on real world regression. We demonstrate the performance and flexibility of the Sufficient Rules (SR) on a real-world regression dataset. Since there are no ground truth explanations for real-world datasets, we use the predictive performance and simplicity (number of variables used) of the SR as an indicator of the effectiveness of the explanations. Indeed, we can build a global model by combining all the Sufficient Rules found for the observations in the training set, and we measure its performance on the test set. We set the output of each rule as the majority class (resp. average values) for classification (resp. regression) of the training observations that satisfy this rule. Note that some rules can overlap and an observation can satisfy multiple rules. To resolve these conflicts, we use the output of the rule with the best precision (AUC or MSE). We called this model Global-SR. We have experimented on Bike Sharing data [Kaggle, 2015] that contains 10886 records and 15 variables about historical usage patterns with weather data in order to forecast bike rental demand in Washington, D.C.

We split the data into train (75%) - test (25%) set and train a RF with $k = 20$ trees and maximal depth = 14. It has mean absolute error $MAE = 25$ and $R^2 = 94\%$ on test set. We use the RF on the test set to generate the Sufficient Rules (SR). Although the Global-SR covers 78% of the test set, we observe that it performs as well as the baseline model with $MAE = 29$, $R^2 = 90\%$, while providing transparency in its decision-making process. Note that the rules of the SR on Bike Sharing Demand are based on 4.5 variables in average. We present examples of the learned rules: $R_1 = \{\text{If Workingday} = \text{True and Hours} \in [5.5, 6.5] \text{ THEN Bike rental demand} = 20\}$, $R_2 = \{\text{If Hours} \in [8.5, 9] \text{ and Year} \leq 2011 \text{ and month} \geq 5 \text{ THEN Bike rental demand} = 192\}$. The number of observations satisfying rules R_1 and R_2 is 134 and 133, respectively, with mean absolute errors of $MAE_{R_1} = 12$ and $MAE_{R_2} = 30$.

Anchors vs Sufficient Rules (SR). To compare our methods with respect to Anchors, we have to consider a classification problem. We use three popular real-world datasets: **Compas** ($n = 6167, p = 14$) [Washington, 2018], **Nhanesi** ($n = 8593, p = 17$) [CDC, 1999-2022], and **Employee Attrition** ($n = 1470, p = 27$) [Kaggle, 2017] which we split into train (75%) - test (25%) set, and train a RF with the parameters of the previous section. We use the RF to generate the local rule-based explanations with Anchors and the SDP approach (Sufficient Rules) to explain the RF’s predictions on the test set. We aim to evaluate the generalization of each explanation across the population. Thus, we measure the following metrics, *Coverage* (higher is better): what fraction of unseen instances fall in the rule and *Accuracy* (higher is better): average number of unseen instances that satisfy the rule and has the same output than the observation that generate the rule, *Sparsity* (lower is better): the mean, variance and maximal size of the rule (number of variables on which it is based).

Table 4.3: Results of the *Accuracy* (Acc), *Coverage* (Cov), and *Sparsity* (Sprs) on **Compas**, **Nhanesi**, **ATTRITION** of the Sufficient Rules (SR) and Anchors. The vector of Sprs (mean, std, max) corresponds to the mean, variance, and max size of the rules.

	COMPAS			NHANESI			ATTRITION		
	Acc	Sprs	Cov	Acc	Sprs	Cov	Acc	Sprs	Cov
SR	0.95	(1.6, 0.96, 7)	0.30	0.97	(1.3, 0.65, 7)	0.41	0.95	(1.15, 0.90, 9)	0.76
Anchors	0.92	(1.83, 1.89, 11)	0.23	0.96	(1.8, 3.91, 16)	0.31	0.95	(0.82, 4.24, 21)	0.74

In table 4.3, we observe that both model have a high accuracy in all datasets, but SR consistently outperforms Anchors on all datasets.

On the other hand, Anchors uses many more features. Indeed, by sampling marginally (i.e. assuming that the features are independent) Anchors succeed to find accurate and high coverage rule, but at the cost of optimality. In fact, we observe in table 4.3 that Anchors tends to give much longer rules. While the observed maximal size of SR is 9 in all dataset, Anchors can provide a rule of size 12 (**Compas**), 16 (**Nhanesi**), 23 (**Attrition**). For instance, the size distribution of Anchors on **Nhanesi** is represented with the following dictionary $\{\mathbf{size} : \mathit{count}\}$: $\{\mathbf{1} : 704, \mathbf{2} : 127, \mathbf{3} : 71, \mathbf{4} : 21, \mathbf{5} : 13, \mathbf{6} : 10, \mathbf{7} : 10, \mathbf{8} : 9, \mathbf{9} : 9, \mathbf{10} : 4, \mathbf{11} : 2, \mathbf{12} : 4, \mathbf{13} : 5, \mathbf{14} : 1, \mathbf{16} : 1\}$, and the corresponding distribution for the SR is $\{\mathbf{1} : 775, \mathbf{2} : 145, \mathbf{3} : 52, \mathbf{4} : 9, \mathbf{5} : 12, \mathbf{7} : 4\}$. Note that this is a significant drawback of Anchors, as simplicity is an essential desideratum for explanation methods. We give an additional experiment confirming these results in the Appendix (10).

Another desirable property of explanation methods is stability, i.e., nearby observations must have the same explanations. Here, we evaluate the stability of the methods with respect to input perturbations. For each observation \mathbf{x} , we compare its rule with the rules of 50 noisy versions of \mathbf{x} obtained by adding random Gaussian noises $\mathcal{N}(0, \sigma^2 \times I)$ to the values of the features with $\sigma^2 = 0.1$. The perturbation is small enough to not change the prediction. For each dataset (**Compas**, **Nhanesi**, **Attrition**), we randomly perturb 100 observations of the test set (50 times), and we observe in average 10 (std=76), 6.83 (std=139), 14 (std=58) different rules for Anchors respectively, while we have 1.5 (std=0.25), 1.1 (std=1.9), 1.13 (std=0.13) for

SR, resp. It shows the large instability of Anchor compared to SR. Indeed, even when $\epsilon = 0$, Anchors gives different rules, e.g., on **Compas** its has 7 (std=70) different rules in average with no perturbations. Results for other values of σ can be found in Appendix (10.3).

These experiments demonstrate that SR provides more interpretable and reliable rules than Anchors. SR exhibits greater stability across perturbations, produces sparser rules, and achieves larger coverage. We also conduct an additional experiment in the Appendix (10.1) to confirm this claim in a setting where we know the ground truth.

6 Conclusion

In this work, we introduce a fast and consistent estimator of the Same Decision Probability and propose a natural generalization of the SDP for regression problems. Then, we introduce the first local rule-based explanations for regression. We give consistent estimates of three local explanation methods: Minimal Sufficient Explanations, Local eXplanatory Importance, and Minimal Sufficient Rules for any data. We prove that these methods considerably improve local variable detection over state-of-the-art algorithms while ensuring minimality, sufficiency, and stability. Our generalization of SDP and Minimal Sufficient Rules are tightly related. They are linked by a Random Forest, which is a computationally and statistically efficient estimator of the SDP and gives the partition that is translated into an interpretable rule. Therefore, our method is principally suitable for datasets where tree-based models work well (e.g., tabular data). In future works, we aim at improving the confidence of the Sufficient Rules by taking into account uncertainty estimates of their predictions.

Chapter 5

Rethinking Counterfactual Explanations as Local and Regional Counterfactual Policies

Abstract

Counterfactual Explanations (CE) face several unresolved challenges, such as ensuring stability, synthesizing multiple CEs, and providing plausibility and sparsity guarantees. From a more practical point of view, recent studies [Pawelczyk, 2022] show that the prescribed counterfactual recourses are often not implemented exactly by individuals and demonstrate that most state-of-the-art CE algorithms are very likely to fail in this noisy environment. To address these issues, we propose a probabilistic framework that gives a sparse local counterfactual rule for each observation, providing rules that give a range of values capable of changing decisions with high probability. These rules serve as a summary of diverse counterfactual explanations and yield robust recourses. We further aggregate these local rules into a regional counterfactual rule, identifying shared recourses for subgroups of the data. Our local and regional rules are derived from the Random Forest algorithm, which offers statistical guarantees and fidelity to data distribution by selecting recourses in high-density regions. Moreover, our rules are sparse as we first select the smallest set of variables having a high probability of changing the decision. We have conducted experiments to validate the effectiveness of our counterfactual rules in comparison to standard CE and recent similar attempts. Our methods are available as a Python package.

Contents

1	Introduction	85
2	Motivation and Related works	86
3	Minimal Counterfactual Rules	88
4	Estimation of the <i>CDP</i> and <i>CRP</i>	91

5	Learning the Counterfactual Rules	93
6	Sampling CE using the CR	95
7	Experiments	96
8	Conclusion	98

1 Introduction

In recent years, many explanations methods have been developed for explaining machine learning models, with a strong focus on local analysis, i.e., generating explanations for individual prediction, see [Molnar, 2022] for a survey. Among this plethora of methods, Counterfactual Explanations [Wachter, 2017] have emerged as one of the most prominent and active techniques. In contrast to popular local attribution methods such as SHAP [Lundberg, 2020b] and LIME [Ribeiro, 2016a], which assign importance scores to each feature, Counterfactuals Explanations (CE) describe the smallest modification to the feature values that changes the prediction to a desired target. These modifications are often called recourses. While CE can be intuitive and user-friendly, providing recourse in certain situations (e.g., loan applications), they have practical limitations. Most CE methods depend on gradient-based algorithms or heuristic approaches [Karimi, 2020b], which can fail to identify the most natural modification and lack guarantees. Most algorithms either do not ensure sparse counterfactuals (changes to the smallest number of features) or fail to generate in-distribution samples (refer to [Verma, 2020; Chou, 2022] for a survey on counterfactual methods). Several studies [Parmentier, 2021; Poyiadzi, 2019; Loovoren, 2019] attempt to address the plausibility and the sparsity issues by incorporating ad-hoc constraints.

In another direction, numerous papers [Mothilal, 2020; Karimi, 2020a; Russell, 2019] encourage the generation of diverse counterfactuals in order to find actionable recourse [Ustun, 2019a]. Actionability is a vital desideratum, as some features may be non-actionable, and generating many counterfactuals increases the chance of getting actionable recourse. However, the diversity of CE compromises the intelligibility of the explanation, and the synthesis of various CE or local explanations, in general, remains an unsolved challenge [Lakkaraju, 2022]. Recently, [Pawelczyk, 2022] highlights a new problem of CE called: *noisy responses to prescribed recourses*. In real-world scenarios, some individuals may not be able to implement exactly the prescribed recourses, and they show that most CE methods fail in this noisy environment. Consequently, we propose to reverse the usual way of explaining with counterfactual by computing *Counterfactual rules*. We introduce a new line of counterfactuals, constructing interpretable policies for changing a decision with a high probability while ensuring the stability of the derived recourse. These policies are sparse, faithful to the data distribution and their computation comes with statistical guarantees. Our proposal is to find a general policy or rule that permits changing the decision while fixing some features instead of generating many counterfactual samples. One of the main challenges is identifying the minimal set of features that provide the directions for changing the decision to the desired output with high probability. Additionally, we show that this method

can be extended to create a common counterfactual policy for subgroups of the data, which aids model debugging and bias detection. Notably, our approach is model-free, meaning it does not need the model to make predictions or calculate other quantities, such as gradients. Instead, it is an inferential approach and relies solely on historical data. As a result, our approach can be applied not only to generate counterfactuals for a specific model but also directly for the data-generating process. An example of the counterfactual rules we introduce is illustrated in Figure 5.1.

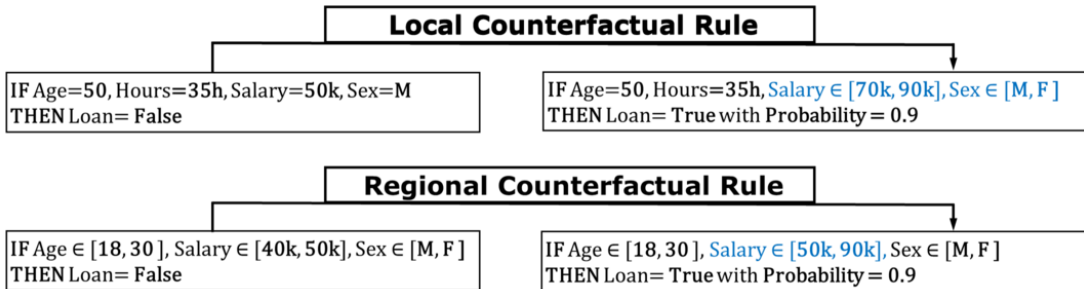


Figure 5.1: Illustration of local and regional Counterfactual Rules for a fictitious dataset with four variables: Age, Salary, Sex, and HoursPerWeek. Local rules change a single instance’s decision, while regional rules apply to a sub-population. Blue indicates the suggested rules for changing decisions.

2 Motivation and Related works

Most Counterfactuals Explanations methods are based on the approach of the seminal work of [Wachter, 2017], where counterfactual samples are generated by cost optimization. This procedure does not account directly for the plausibility of the counterfactual examples, see Table 1 from [Verma, 2020] for a classification of CE methods. Indeed, a major shortcoming is that the action suggested for obtaining the counterfactual sample is not designed to be feasible or representative of the underlying data distribution. Several recent studies have suggested incorporating ad-hoc plausibility constraints into the optimization process. For instance, Local Outlier Factor [Kanamori, 2020], Isolation Forest [Parmentier, 2021], and density-weighted metrics [Poyiadzi, 2019] have been employed to generate realistic samples. Alternatively, [Looveren, 2019] proposes the use of an autoencoder that penalizes out-of-distribution candidates. Instead of relying on ad-hoc constraints, we propose CE that gives plausible explanations by design. Our approach leverages the Random Forest (RF) algorithm, which helps identify high-density regions and ensures counterfactual samples reside within these areas. To ensure sparsity, we begin by identifying the smallest subset S of variables \mathbf{X}_S and associated value ranges for each observation that have the highest probability of changing the prediction. We compute this probability with a consistent estimator of the conditional distribution $Y|\mathbf{X}_S$ obtained from a RF. As a consequence, the sparsity of the counterfactuals is not encouraged indirectly by adding a penalty term (ℓ_0 or ℓ_1) as existing works [Mothilal, 2020]. Our method draws inspiration from the concept of *Same Decision Probability (SDP)* [Chen, 2012], which is used to identify the smallest feature subset that guarantees prediction stability with a high probability. This

minimal subset is called *Sufficient Explanations*. In [Amoukou, 2021a] or Chapter 4, it has been shown that the *SDP* and the *Sufficient Explanations* can be estimated and computed efficiently for identifying important local variables in any classification/regression models using RF. For counterfactuals, we are interested in the dual set. We want the minimal subset of features that allows for a high probability of changing the decision when the other features remain fixed.

Another limitation of the current CE is the multiplicity of the explanations produced. While some papers [Mothilal, 2020; Karimi, 2020a; Russell, 2019] promote the generation of diverse counterfactual samples to ensure actionable recourse, such diverse explanations should be summarized to be intelligible [Lakkaraju, 2022], but the compilation of local explanations is often a very difficult problem. To address this issue, instead of generating counterfactual samples, we construct a rule called *Local Counterfactual Rules* (L-CR) from which counterfactual samples can be derived. In contrast to traditional CE that identify the nearest instances with a desired output, we first determine the most effective rule (or group of similar observations) for each observation that changes the prediction to the intended target. The L-CR can be seen as a summary of the diverse counterfactual samples possible for a given instance. For example, if $\mathbf{x}_0 = \{\text{Age}=20, \text{Salary}=35\text{k}, \text{HoursWeek}=25\text{h}, \text{Sex}=\text{M}, \dots\}$ with $\text{Loan}=\text{False}$, fixing the variables Age and Sex and modifying the Salary and HoursWeek change the decision. Therefore, instead of giving multiples combination of Salary and HoursWeek (e.g., 35k and 40h or 40k and 55h, ...) that result in many samples, the counterfactual rule gives the range of values: $C_0 = [\text{IF HoursWeek} \in [35\text{h}, 50\text{h}], \text{Salary} \in [40\text{k}, 50\text{k}], \text{ and the remaining features are fixed THEN Loan}=\text{True with high probability}]$. One can also have several observations with the same predictions and almost the same counterfactual rules. For example, consider a second observation $\mathbf{x}_1 = \{\text{Age}=25, \text{Salary}=45\text{k}, \text{HoursWeek}=25\text{h}, \text{Sex}=\text{M}, \dots\}$ with $\text{Loan} = \text{False}$, such that $\mathbf{x}_0, \mathbf{x}_1$ are included in the following hyper-rectangle (or rule) $\mathbf{R} = [\text{IF Salary} \in [20\text{k}, 45\text{k}], \text{Age} \in [20, 30] \text{ THEN Loan}=\text{False}]$ which may contain other observations. The local CR of \mathbf{x}_1 is $C_1 := [\text{IF HoursWeek} \in [40\text{h}, 45\text{h}], \text{Salary} \in [48\text{k}, 50\text{k}], \text{ and the remaining features are fixed THEN Loan}=\text{True with high probability}]$. We observe that x_0, x_1 have nearly identical counterfactual rules C_0, C_1 . Hence, the global counterfactual rules enable summarizing such information into a single rule that applies to multiple observations simultaneously. The *Regional Counterfactual Rule* (R-CR) of the rule R could be $C_R := [\text{IF HoursWeek} \in [35\text{h}, 45\text{h}], \text{Salary} \in [40\text{k}, 50\text{k}], \text{ and the remaining rules of } R \text{ are fixed THEN Loan}=\text{True with high probability}]$. It shows that for all observations that are in the hyperrectangle \mathbf{R} , we can apply the same counterfactual rules to change their predictions. These global rules allow us to have a global picture of the model to detect certain patterns that may be used for fairness detection, among other applications. The main difference between a local and a global CR is that the local CR explain a single instance by fixing the remaining feature values (not used in the CR); while a regional CR is defined by keeping the remaining variables in a given interval (not used in the regional CR). Moreover, by giving ranges of values that guarantee a high probability of changing the decision, we partly answer the problem of *noisy responses to prescribed recourses* [Pawelczyk, 2022]. We find that the generated CE remain robust as long as the perturbations remain within the specified ranges.

While the Local Counterfactual Rule is a novel concept, the Regional Counterfactual Rule shares similarities with some recent works. Indeed, [Rawal, 2020] proposed Actionable Recourse Summaries (AReS), a framework that constructs global counterfactual recourses to have a global insight into the model and detect unfair behavior. Despite similarities with the Regional Counterfactual Rule, there are notable differences. Our methods can handle regression problems and work directly with continuous features. AReS requires discretizing continuous features, leading to a trade-off between speed and performance, as observed by [Ley, 2022]. Too few bins yield unrealistic recourse, while too many bins result in excessive computation time. AReS employs a greedy heuristic search approach to find global recourse, which may result in unstable and inaccurate recourse. Our approaches overcome these limitations by leveraging on the informative partitions obtained from a Random Forest, removing the need for an extensive search space, and focusing on high-density regions for plausibility. Additionally, we prioritize changes to the smallest number of features, utilizing a consistent estimator of the conditional distribution.

Another global CE framework has been introduced in [Kanamori, 2022] to ensure transparency. The Counterfactual Explanation Tree (CET) partitions the input space with a decision tree and assigns a suitable action for changing the decision of each subspace, providing a unique recourse for multiple instances. In comparison, our approach offers greater flexibility in counterfactual explanations by providing a range of possible values that guarantee a change with a given probability for each subspace. We also propose a method to derive classic counterfactual samples using the counterfactual rules. We do not make assumptions about the cost of changing the feature or actionability. If such information is available, it can be incorporated as additional post-processing.

3 Minimal Counterfactual Rules

Consider a dataset $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ consisting of i.i.d observations of $(\mathbf{X}, Y) \sim P_{\mathbf{X}}P_{Y|\mathbf{X}}$, where $\mathbf{X} \in \mathcal{X}$ (typically $\mathcal{X} \subseteq \mathbb{R}^p$) and $Y \in \mathcal{Y}$. The output set \mathcal{Y} can be either discrete or continuous. We denote $[p] = \{1, \dots, p\}$, and for a given subset $S \subseteq [p]$, $\mathbf{X}_S = (X_i)_{i \in S}$ represents a subgroup of features, and we write $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_{\bar{S}})$.

For a given observation (\mathbf{x}, y) , we consider a target set $\mathcal{Y}^* \subset \mathcal{Y}$, such that $y \notin \mathcal{Y}^*$. In the case of a classification problem, $\mathcal{Y}^* = \{y^*\}$ is a singleton where $y^* \in \mathcal{Y}$ and $y^* \neq y$. Unlike conventional approaches, our definition of CE also accommodates regression problems by considering $\mathcal{Y}^* = [a, b] \subset \mathbb{R}$, and the definitions and computations remain the same for both classification and regression. The classic CE problem, defined here only for classification, considers a predictor $f : \mathcal{X} \rightarrow \mathcal{Y}$, trained on dataset \mathcal{D}_n and search a function $\mathbf{a} : \mathcal{X} \rightarrow \mathcal{X}$, such that for all observations $\mathbf{x} \in \mathcal{X}$, $f(\mathbf{x}) \neq y^*$, we have $f(\mathbf{a}(\mathbf{x})) = y^*$. The function is defined point-wise by solving an optimisation program. Most often $\mathbf{a}(\cdot)$ is not a single-output function, as $\mathbf{a}(x)$ may be in fact a collection of (random) values $\{\mathbf{x}_1^{CF}, \dots, \mathbf{x}_k^{CF}\}$, which represent the counterfactual samples. A more recent perspective, proposed by [Kanamori, 2022], defines \mathbf{a} as a decision tree, where for each leaf L , a common action is predicted for all instances $\mathbf{x} \in L$ to change their predictions.

Our approach diverges slightly from the traditional model-based definition of CE as we can directly consider observation (\mathbf{X}, Y) rather than model prediction $(\mathbf{X}, f(\mathbf{X}))$. To illustrate the concept, let's consider a binary classification problem, where the input space can be divided into two regions R_0, R_1 . These correspond to the support of the distributions $\mathbf{X}|Y = 0$ and $\mathbf{X}|Y = 1$. These regions may not be disjoint or convex spaces and can be represented as a union of several sets. Given an observation $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_{\bar{S}})$ with label $y = 0$, our method consists of finding the minimal subset of variables $S \subseteq \{1, \dots, p\}$ to move \mathbf{x} by modifying \mathbf{x}_S into a set within the region R_1 . The objective is to move \mathbf{x} to a high-density set with low variance with respect to the target variable Y , while altering as few variables as possible.

Figure 5.2 provides a visual representation of our approach in the binary case. The first step involves learning a tree-based model on our data, enabling us to partition the input space based on the target variable Y . By examining the tree leaves, we can easily identify the optimal direction S to modify the decision and the target region corresponding to the counterfactual rule. Moreover, these leaves can serve not only as rules but also as a means to generate recourses.

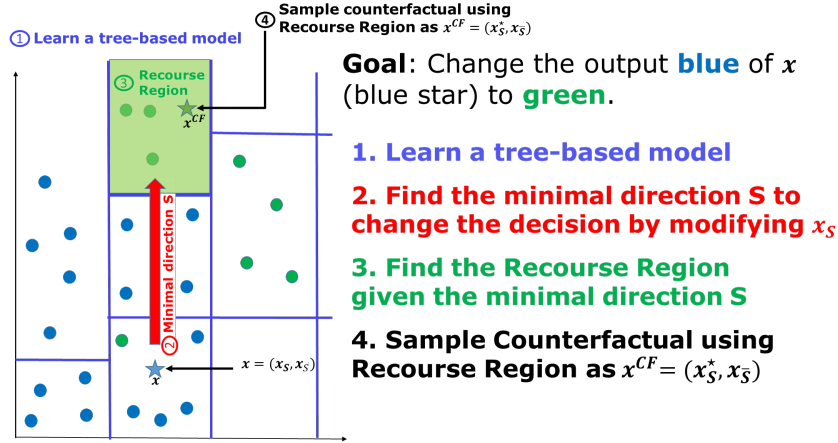


Figure 5.2: Illustration of the 4-stages in our methodology for computing sparse counterfactuals

Our approach is different from the optimization approach for generating recourse, as it is model-free, meaning it does not require the model to generate further predictions or compute gradients. This flexibility allows us to apply our approach to generate recourse either for the predictions of a given model $(\mathbf{X}, f(\mathbf{X}))$ or the data-generating process (\mathbf{X}, Y) . A model-free approach was also proposed by [Black, 2020; De Lara, 2021] under the name of transport-based counterfactuals. It consists of finding a map T between the distribution of $\mathbf{X}|Y = 0$ and $\mathbf{X}|Y = 1$ such that each observation of class $Y = 0$ is linked to the most similar observation of class $Y = 1$. [De Lara, 2021] shows that it coincides with causal counterfactual under appropriate assumptions. In the following discussion, we consider the data (\mathbf{X}, Y) for the presentation of the methods, although they can also be applied to generate recourses for a model prediction $(\mathbf{X}, f(\mathbf{X}))$ as well.

Our approach is hybrid, as we do not suggest a single action for each observation or subspace of \mathcal{X} but provide sets of possible perturbations. A Local Counterfactual Rule (L-CR) for target

\mathcal{Y}^* and observation (\mathbf{x}, y) (with $y \notin \mathcal{Y}^*$) is a rectangle $C_S(\mathbf{x}, \mathcal{Y}^*) = \prod_{i \in S} [a_i, b_i]$, $a_i, b_i \in \overline{\mathbb{R}}$ such that for all counterfactual samples of $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_{\bar{S}})$ obtained as $\mathbf{x}^{CF} = (\mathbf{z}_S, \mathbf{x}_{\bar{S}})$ with $\mathbf{z}_S \in C_S(\mathbf{x}, \mathcal{Y}^*)$ and \mathbf{x}^{CF} an in-distribution sample, then y^{CF} is in \mathcal{Y}^* with a high probability, where y^{CF} is the output of \mathbf{x}^{CF} given by the model f or the data-generating process. Similarly, a Regional Counterfactual Rule (R-CR) $C_S(\mathbf{R}, \mathcal{Y}^*)$ is defined for target \mathcal{Y}^* and a rectangle $\mathbf{R} = \prod_{i=1}^d [a_i, b_i]$, $a_i, b_i \in \overline{\mathbb{R}}$, which represent a subspace of \mathcal{X} of similar observations, if for all observations $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_{\bar{S}}) \in \mathbf{R}$, the countefactual samples obtained as $\mathbf{x}^{CF} = (\mathbf{z}_S, \mathbf{x}_{\bar{S}})$ with $\mathbf{z}_S \in C_S(\mathbf{R}, \mathcal{Y}^*)$ and \mathbf{x}^{CF} an in-distribution sample are such that y^{CF} is in \mathcal{Y}^* with high probability. Our approach constructs such rectangles in a sequential manner. Firstly, we identify the minimal directions $S \subseteq [p]$ that offer the highest probability of changing the decision. Next, we determine the optimal intervals $[a_i, b_i]$ for $i \in S$ that change the decision to the desired target. Additionally, we propose a method to derive traditional Counterfactual Explanations (CE) (i.e., actions that alter the decision) or recourses using our Counterfactual Rules. A central tool in this approach is the Counterfactual Decision Probability presented below.

Definition 3.1. Counterfactual Decision Probability (CDP). The Counterfactual Decision Probability of the subset $S \subseteq \{1, \dots, p\}$, w.r.t $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_{\bar{S}})$, output y and the desired target \mathcal{Y}^* (s.t. $y \notin \mathcal{Y}^*$) is

$$CDP_S(\mathbf{x}, \mathcal{Y}^*) = \mathbb{P}(Y \in \mathcal{Y}^* | \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}}) \quad (5.1)$$

The *CDP* of the subset S is the probability that the decision changes to the desired target \mathcal{Y}^* by sampling the features \mathbf{X}_S given $\mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}}$. It is related to the Same Decision Probability $SDP_S(\mathcal{Y}; \mathbf{x}) = \mathbb{P}(Y \in \mathcal{Y} | \mathbf{X}_S = \mathbf{x}_S)$ used in [Amoukou, 2021a] for solving the dual problem of selecting the most local important variables for obtaining and maintaining the decision $Y \in \mathcal{Y}$, where $\mathcal{Y} \subset \mathcal{Y}$. The set S is called the Minimal Sufficient Explanation. Indeed, we have $CDP_S(\mathbf{x}, \mathcal{Y}^*) = SDP_{\bar{S}}(\mathbf{x}, \mathcal{Y}^*)$. The computation of these probabilities is challenging and discussed in Section 4. Next, we define the minimal subset of features S that allows changing the decision to the target set with a given high probability π .

Definition 3.2. (Minimal Divergent Explanations). Given an instance (\mathbf{x}, y) and a desired target $\mathcal{Y}^* \not\ni y$, S is a Divergent Explanation for probability $\pi > 0$ if

- $CDP_S(\mathbf{x}, \mathcal{Y}^*) \geq \pi$
- no subset Z of S satisfies $CDP_Z(\mathbf{x}, \mathcal{Y}^*) \geq \pi$.

Hence, a Minimal Divergent Explanation is a Divergent Explanation with the smallest size.

The set satisfying these properties is not unique, and we can have several Minimal Divergent Explanations. Note that the probability π represents the minimum level required for a set to be chosen for generating counterfactuals, and its value should be as high as possible and depends on the use case. With these concepts established, we can now define our main criterion for constructing a Local Counterfactual Rule (L-CR).

Definition 3.3. (Local Counterfactual Rule). Given an instance (\mathbf{x}, y) , a desired target $\mathcal{Y}^* \neq y$, a Minimal Divergent Explanation S , the rectangle $C_S(\mathbf{x}, \mathcal{Y}^*) = \prod_{i \in S} [a_i, b_i]$, $a_i, b_i \in \overline{\mathbb{R}}$ is a Local Counterfactual Rule with probability π_C if

- $C_S(\mathbf{x}, \mathcal{Y}^*) = \arg \max_C \mathbb{P}_{P_{\mathbf{X}}}(\mathbf{X}_S \in C \mid \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}})$ such that
- $CRP_S(\mathbf{x}, \mathcal{Y}^*) = \mathbb{P}(Y \in \mathcal{Y}^* \mid \mathbf{X}_S \in C_S(\mathbf{x}, \mathcal{Y}^*), \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}})$ satisfies

$$CRP_S(\mathbf{x}, \mathcal{Y}^*) \geq \pi_C. \quad (5.2)$$

$\mathbb{P}_{P_{\mathbf{X}}}(\mathbf{X}_S \in C_S(\mathbf{x}, \mathcal{Y}^*) \mid \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}})$ represent the plausibility of the rule and by maximizing it, we ensure that the rule lies in a high-density region. CRP_S is the Counterfactual Rule Probability. The higher the probability π_C is, the better the relevance of the rule $C_S(\mathbf{x}, \mathcal{Y}^*)$ is for changing the decision to the desired target.

In practice, we often observe that the Local CR $C_S(\cdot, \mathcal{Y}^*)$ for neighboring observations \mathbf{x} and \mathbf{x}' are quite similar, as the Minimal Divergent Explanations tend to be alike, and the corresponding hyperrectangles frequently overlap. This observation motivates a generalization of these Local CRs to hyperrectangles $\mathbf{R} = \prod_{i=1}^d [a_i, b_i]$, $a_i, b_i \in \overline{\mathbb{R}}$, which group together similar observations. We denote $\text{supp}(\mathbf{R}) = \{i : [a_i, b_i] \neq \overline{\mathbb{R}}\}$ as the support of the rectangle and extend the Local CRs to Regional Counterfactual Rules (R-CR). To achieve this, we denote $\mathbf{R}_{\bar{S}} = \prod_{i \in \bar{S}} [a_i, b_i]$ as the rectangle with intervals of \mathbf{R} in $\text{supp}(\mathbf{R}) \cap \bar{S}$, and define the corresponding Counterfactual Decision Probability (CDP) for rule \mathbf{R} and subset S as $CDP_S(\mathbf{R}, \mathcal{Y}^*) = \mathbb{P}(Y \in \mathcal{Y}^* \mid \mathbf{X}_{\bar{S}} \in \mathbf{R}_{\bar{S}})$. Consequently, we can compute the Minimal Divergent Explanation for rule \mathbf{R} using the corresponding CDP for rules, following Definition (3.2). The Regional Counterfactual Rules (R-CR) correspond to Definition (3.3) with the associated CDP for rules.

4 Estimation of the CDP and CRP

To compute the probabilities CDP_S and CRP_S for any S , we use a dedicated Random Forest (RF) that learns to predict the output of the model or the data-generating process. Indeed, the conditional probabilities CDP_S and CRP_S can be easily computed from a RF by combining the Projected Forest algorithm [Bénard, 2021b] and the Quantile Regression Forest [Meinshausen, 2006]. As a result, we can estimate the probabilities $CDP_S(\mathbf{x}, \mathcal{Y}^*)$ consistently. This method has been previously utilized by [Amoukou, 2021a] for calculating the Same Decision Probability SDP_S .

4.1 Projected Forest and CDP_S

The estimator of the SDP_S is based on the Random Forest [Breiman, 1984] algorithm. Assuming that we have trained a RF $m(\cdot)$ using the dataset \mathcal{D}_n , the model consists of a collection of k randomized trees (for a detailed description of decision trees, see [Loh, 2011]). For each instance \mathbf{x} , the predicted value of the l -th tree is denoted as $m_l(\mathbf{x}; \Theta_l)$, where Θ_l represents

the resampling data mechanism in the j -th tree and the subsequent random splitting directions. The predictions of the individual trees are then averaged to produce the prediction of the forest as $m(\mathbf{x}; \Theta_1, \dots, \Theta_k) = \frac{1}{k} \sum_{l=1}^k m_l(\mathbf{x}; \Theta_l)$. The RF can also be interpreted as an adaptive nearest neighbor predictor [Lin, 2006; Biau, 2010] or kernel methods [Breiman, 2000; Geurts, 2006; Scornet, 2016]. For every instance \mathbf{x} , the observations in \mathcal{D}_n are weighted by $w_{n,i}(\mathbf{x})$, with $i = 1, \dots, n$. As a result, the prediction of the RF can be reformulated as $m(\mathbf{x}; \Theta_1, \dots, \Theta_k) = \sum_{i=1}^n w_{n,i}(\mathbf{x}) Y_i$. This emphasizes the central role played by the weights in the RF’s algorithm. See [Meinshausen, 2006] or Chapter 4 for a detailed description of the weights. Consequently, it naturally gives estimators for other quantities, e.g., cumulative hazard function [Ishwaran, 2008], treatment effect [Wager, 2017; Jocteur, 2023], conditional density [Du, 2021]. For instance, [Meinshausen, 2006] showed that we can use the same weights to estimate the conditional distribution function with the following estimator $\hat{F}(y|\mathbf{X} = \mathbf{x}) = \sum_{i=1}^n w_{n,i}(\mathbf{x}) \mathbb{1}_{Y_i \leq y}$. In another direction, [Bénard, 2021b] introduced the Projected Forest algorithm [Bénard, 2021e; Bénard, 2021b] that aims to estimate $E[Y|\mathbf{X}_S]$ by modifying the RF’s prediction algorithm.

Projected Forest: To estimate $E[Y|\mathbf{X}_S = \mathbf{x}_S]$ instead of $E[Y|\mathbf{X} = \mathbf{x}]$ using a RF, [Bénard, 2021d] suggests to simply ignore the splits based on the variables not contained in S from the tree predictions. More formally, it consists of projecting the partition of each tree of the forest on the subspace spanned by the variables in S . The authors also introduced an algorithmic trick that computes the output of the Projected Forest efficiently without modifying the initial tree structures. It consists of dropping the observations down in the initial trees, ignoring the splits which use a variable not in S : when it encounters a split involving a variable $i \notin S$, the observations are sent both to the left and right children nodes. Therefore, each instance falls in multiple terminal leaves of the tree. To compute the prediction of \mathbf{x}_S , we follow the same procedure, and gather the set of terminal leaves where \mathbf{x}_S falls. Next, we collect the training observations which belong to every terminal leaf of this collection, in other words, we keep only the observations that fall in the intersection of the leaves where \mathbf{x}_S falls. Finally, we average their outputs Y_i to generate the estimation of $E[Y|\mathbf{X}_S = \mathbf{x}_S]$. Notice that the authors show that this algorithm converges asymptotically to the true projected conditional expectation $E[Y|\mathbf{X}_S = \mathbf{x}_S]$ under suitable assumptions. As the RF, the Projected Forest (PRF) assigns a weight to each training observation. The associated PRF is denoted $m^{(S)}(\mathbf{x}_S) = \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) Y_i$. Therefore, as the weights of the original forest was used to estimate the CDF, [Amoukou, 2021a] used the weights of the Projected Forest Algorithm to estimate SDP as $\widehat{SDP}_S(\mathbf{x}, \mathcal{Y}^*) = \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{Y_i \in \mathcal{Y}^*}$. The idea is essentially to replace Y_i by $\mathbb{1}_{Y_i \in \mathcal{Y}^*}$ in the Projected Forest equation defined above. [Amoukou, 2021a] also show that this estimator converges to the true SDP_S under suitable assumptions and works very well in practice. Especially for tabular data, where tree-based models are known to perform well [Grinsztajn, 2022]. Similarly, we can estimate the CDP with statistical guarantees [Amoukou, 2021a] using the following estimator $\widehat{CDP}_S(\mathbf{x}, \mathcal{Y}^*) = \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{Y_i \in \mathcal{Y}^*}$.

Remark: We only give the estimator of CDP_S of an instance \mathbf{x} . The estimator for CDP_S of a rule R will be discussed in the next section, as it is closely related to the estimator of the CRP_S .

4.2 Regional RF and CRP_S

Here, we focus on estimating the $CRP_S(\mathbf{x}, \mathcal{Y}^*) = \mathbb{P}(Y \in \mathcal{Y}^* | \mathbf{X}_S \in C_S(\mathbf{x}, \mathcal{Y}^*), \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}})$ and $CRP_S(\mathbf{R}, \mathcal{Y}^*) = \mathbb{P}(Y \in \mathcal{Y}^* | \mathbf{X}_S \in C_S(\mathbf{R}; \mathcal{Y}^*), \mathbf{X}_{\bar{S}} \in \mathbf{R}_{\bar{S}})$. For ease of reading, we remove the dependency of the rectangles C_S in \mathcal{Y}^* . Based on the previous section, we already know that the estimators using the RF will take the form of $\widehat{CRP}_S(\mathbf{x}, \mathcal{Y}^*) = \sum_{i=1}^n w_{n,i}^R(\mathbf{x}) \mathbb{1}_{Y_i \in \mathcal{Y}^*}$, so we only need to determine the appropriate weighting. The main challenge lies in the fact that we have a condition based on a region, e.g., $\mathbf{X}_S \in C_S(\mathbf{x})$ or $\mathbf{X}_{\bar{S}} \in \mathbf{R}_{\bar{S}}$ (regional-based) instead of a condition of type $\mathbf{X}_S = \mathbf{x}_S$ (fixed value-based) as usual. However, we introduced a natural extension of the RF algorithm to handle predictions when the conditions are both regional-based and fixed value-based. As a result, cases with only regional-based conditions can be naturally derived.

Regional RF to estimate $CRP_S(\mathbf{x}, \mathcal{Y}^*) = \mathbb{P}(Y \in \mathcal{Y}^* | \mathbf{X}_S \in C_S(\mathbf{x}), \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}})$. The algorithm is based on a slight modification of RF and works as follows: we drop the observations in the trees, if a split used variable $i \in \bar{S}$, i.e., fixed value-based condition, we use the classic rules of RF, if $x_i \leq t$, the observations go to the left children, otherwise the right children. However, if a split used variable $i \in S$, i.e., regional-based condition, we use the rectangles $C_S(\mathbf{x}) = \prod_{i=1}^{|S|} [a_i, b_i]$. The observations are sent to the left children if $b_i \leq t$, right children if $a_i > t$ and if $t \in [a_i, b_i]$, the observations are sent both to the left and right children. Consequently, we use the weights of the Regional RF algorithm $w_{n,i}^R(\mathbf{x})$ to estimate CRP_S , the estimator is $\widehat{CRP}_S(\mathbf{x}, \mathcal{Y}^*) = \sum_{i=1}^n w_{n,i}^R(\mathbf{x}) \mathbb{1}_{Y_i \in \mathcal{Y}^*}$. In addition, the number of observations at the leaves is used as an estimate of $\mathbb{P}(\mathbf{X}_S \in C_S(\mathbf{x}) | \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}})$. A more comprehensive description and discussion of the algorithm are provided in the Appendix (13).

To estimate the CDP of a rule $CDP_S(\mathbf{R}, \mathcal{Y}^*) = \mathbb{P}(Y \in \mathcal{Y}^* | \mathbf{X}_{\bar{S}} \in \mathbf{R}_{\bar{S}})$, we just have to apply the Projected Forest algorithm to the Regional RF, i.e., when a split involving a variable outside of \bar{S} is met, the observations are sent both to the left and right children nodes, otherwise we use the Regional RF split rule, i.e., if an interval of $\mathbf{R}_{\bar{S}}$ is below t , the observations go to the left children, otherwise the right children and if t is in the interval, the observations go to the left and right children. The estimator of the $CRP_S(\mathbf{R}, \mathcal{Y}^*) = \mathbb{P}(Y \in \mathcal{Y}^* | \mathbf{X}_S \in C_S(\mathbf{R}; \mathcal{Y}^*), \mathbf{X}_{\bar{S}} \in \mathbf{R}_{\bar{S}})$ for rule \mathbf{R} is also derived from the Regional RF. Indeed, it is a special case of the Regional RF algorithm where there are only regional-based conditions.

5 Learning the Counterfactual Rules

The computation of the Local and Regional CR is performed using the estimators introduced in the previous section. First, we determine the Minimal Divergent Explanation, akin to the Minimal Sufficient Explanation [Amoukou, 2021a], by exploring the subsets obtained using the $K = 10$ most frequently selected variables in the Random Forest estimator. K is a hyperparameter to choose according to the use case and computational power. We can also use any importance measure. An alternative strategy to exhaustively searching through the 2^K possible

subsets would be to sample a sufficient number of subsets, typically a few thousand, that are present in the decision paths of the trees in the forest. By construction, these subsets are likely to contain influential variables. A similar strategy was used in [Basu, 2018; B enard, 2021b].

Given an instance \mathbf{x} or rectangle \mathbf{R} , target set \mathcal{Y}^* and their corresponding Minimal Divergent Explanation S , our objective is to find the maximal rule $C_S(\cdot) = \prod_{i \in S} [a_i, b_i]$ s.t. given $\mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}}$ or $\mathbf{X}_{\bar{S}} \in \mathbf{R}_{\bar{S}}$, and $\mathbf{X}_S \in C_S(\cdot)$, the probability that $Y \in \mathcal{Y}^*$ is high. Formally, we want: $\mathbb{P}(Y \in \mathcal{Y}^* | \mathbf{X}_S \in C_S(\mathbf{x}), \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}})$ or $\mathbb{P}(Y \in \mathcal{Y}^* | \mathbf{X}_S \in C_S(\mathbf{R}), \mathbf{X}_{\bar{S}} \in \mathbf{R}_{\bar{S}})$ above π_C .

The rectangles $C_S(\cdot) = \prod_{i \in S} [a_i, b_i]$ defining the CR are derived from the RF. In fact, these rectangles naturally arise from the partition learned by the RF. ARoS [Rawal, 2020], on the other hand, relies on binned variables to generate candidate rules, testing all possible rules to select the optimal one. By leveraging the partition learned by the RF, we overcome both the computational burden and the challenge of choosing the number of bins. Moreover, by focusing only on the non-empty leaves containing training observations, we significantly reduce the search space. This approach allows identifying high-density regions of the input space to generate plausible counterfactual explanations.

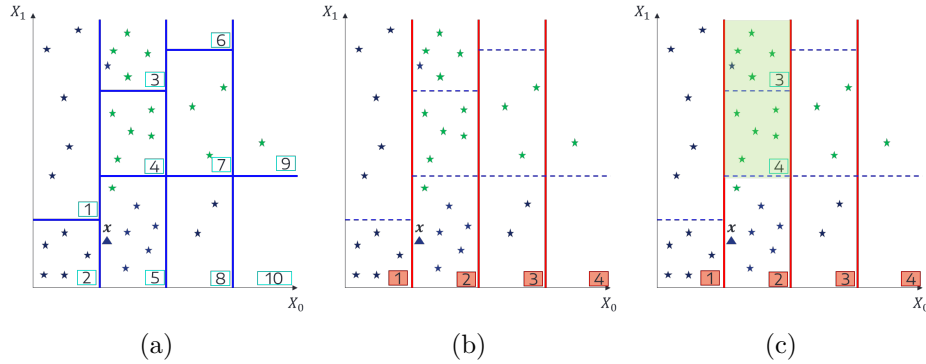


Figure 5.3: (a) Partition of the Random Forest, (b) Partition of the Projected Random Forest when we condition given X_0 , i.e., ignoring the splits on X_1 , (c) The optimal Counterfactual Rule of \mathbf{x} when we condition given $X_0 = x_0$ is the green region.

To illustrate the idea, we use a two-dimensional data (X_0, X_1) with binary label Y represented as green and blue stars in Figure 5.3a. We fit a Random Forest to classify this dataset and show its partition in Figure 5.3a. We consider an instance \mathbf{x} (blue triangle), and our goal is to change its classification from blue to green. From a visual analysis of cells/leaves of the RF, we deduce that the Minimal Divergent Explanation of \mathbf{x} is $S = X_1$. In Figure 5.3b, we observe the leaves of the Projected Forest when not conditioning on $S = X_1$, thus projecting the RF’s partition only on the subspace X_0 . It consists of ignoring all the splits in the other directions (here the X_1 -axis), thus \mathbf{x} falls in the projected leaf 2 (see Figure 5.3b) and its CDP is $CDP_{X_1}(\text{green}; \mathbf{x}) = \frac{10 \text{ green}}{10 \text{ green} + 17 \text{ blue}} = 0.58$. To find the optimal rectangle $C_S(\mathbf{x}) = [a_i, b_i]$ in the direction of X_1 , such that the decision changes, we can utilize the leaves of the RF. By looking at the leaves of the RF (Figure 5.3a) for observations belonging to the projected RF leaf 2 (Figure 5.3b) where \mathbf{x} falls, we observe in Figure 5.3c that the optimal rectangle for changing

the decision, given $X_0 = x_0$ or being in the projected RF leaf 2, is the union of the intervals on X_1 of the leaf 3 and 4 of the RF (see the green region in Figure 5.3c).

Given an instance \mathbf{x} and its Minimal Divergent Explanation S , the first step is to collect observations that belong to the leaf of the Projected Forest given \bar{S} , where \mathbf{x} falls. These observations correspond to those with positive weights in the computation of $CDP_S(\mathbf{x}, \mathcal{Y}^*) = \sum_{i=1}^n w_{n,i}^R(\mathbf{x}_{\bar{S}}) \mathbb{1}_{Y_i \in \mathcal{Y}^*}$, i.e., $\{\mathbf{X}_i : w_{n,i}^R(\mathbf{x}_{\bar{S}}) > 0\}$. Then we use the partition of the original forest to find the possible leaves in the direction S . The possible leaves are among the RF’s leaves of the collected observations $\{\mathbf{X}_i : w_{n,i}^R(\mathbf{x}_{\bar{S}}) > 0\}$. Let denote $L(\mathbf{X}_i)$ the leaf of the observation \mathbf{X}_i with $w_{n,i}^R(\mathbf{x}_{\bar{S}}) > 0$. A possible leaf is a leaf $L(\mathbf{X}_i)$ s.t. $\mathbb{P}(Y \in \mathcal{Y}^* | \mathbf{X}_S \in L(\mathbf{X}_i)_S, \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}}) \geq \pi_C$. Finally, we merge all the possible neighboring leaves to get the largest rectangle, and this maximal rectangle is the counterfactual rule. It is important to note that the union of possible leaves is not necessarily a connected space, which may result in multiple disconnected counterfactual rules.

We apply the same approach to find the regional CR. Given a rule \mathbf{R} and its Minimal Divergent Explanation S , we used the Projection given $\mathbf{X}_{\bar{S}} \in \mathbf{R}_{\bar{S}}$ to identify compatible observations and their leaves. We then combine the possible ones that satisfy $CRP_S(\mathbf{R}, \mathcal{Y}^*) \geq \pi_C$ to obtain the regional CR. For instance, if we consider Leaf 5 of the original forest as a rule (i.e., if $\mathbf{X} \in$ Leaf 5, then predict blue), its Minimal Divergent Explanation is also $S = X_1$. The Regional CR would be the green region in Figure 5.3c. Indeed, satisfying the X_0 condition of Leaf 5 and the X_1 condition of Leaves 3 and 4 would cause the decision to change to green.

6 Sampling CE using the CR

Our approaches cannot be directly compared with traditional CE methods, as they return counterfactual samples, whereas we provide rules (ranges of vector values) that permit changing the decision with high probability. In some applications, users might prefer recourse to CR. Hence, we adapt the CR to generate counterfactual samples using a generative model. For example, given an instance $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_{\bar{S}})$, target set \mathcal{Y}^* and its counterfactual rule $C_S(\mathbf{x}, \mathcal{Y}^*)$, we want to find a sample $\mathbf{x}^{CF} = (\mathbf{z}_S, \mathbf{x}_{\bar{S}})$ with $\mathbf{z}_S \in C_S(\mathbf{x}, \mathcal{Y}^*)$ such that \mathbf{x}^{CF} is a realistic sample and $y^{CF} \in \mathcal{Y}^*$. Instead of using a complex conditional generative model as [Xu, 2019; Patki, 2016], which can be difficult to calibrate, we use an energy-based generative approach [Grathwohl, 2020; Lecun, 2006]. The core idea is to find $\mathbf{z}_S \in C_S(\mathbf{x}, \mathcal{Y}^*)$ such that \mathbf{x}^* maximizes a given energy score, ensuring that \mathbf{x}^* lies in a high-density region. We use the negative outlier score of an Isolation Forest [Liu, 2008] and Simulated Annealing [Guilmeau, 2021] to maximize the negative outlier score using the information of the counterfactual rules $C_S(\mathbf{x}, \mathcal{Y}^*)$. In fact, the range values given by the CR $C_S(\mathbf{x}, \mathcal{Y}^*)$ reduce the search space for \mathbf{z}_S drastically. We used the marginal law of \mathbf{X} given $\mathbf{X}_S \in C_S(\mathbf{x}, \mathcal{Y}^*)$ as the proposal distribution, i.e., we draw a candidate \mathbf{z}_S by independently sampling each variable using the marginal law $\mathbf{z}_S \sim \prod_{i \in S} P_{X_j | \mathbf{X}_S \in C_S(\mathbf{x}, \mathcal{Y}^*)}$ until we find an observation $\mathbf{x}^{CF} = (\mathbf{z}_S, \mathbf{x}_{\bar{S}})$ with high energy. The algorithm works similarly for sampling CE with the Regional CR. The methodology is described below in Algorithm 4.

Algorithm 4: Simulated Annealing to generate counterfactual samples using the Counterfactual Rules

Input : Observation \mathbf{x} , Divergent Explanation S , counterfactual rule $C_S(\mathbf{x}, \mathcal{Y}^*)$, \mathcal{D}_n training data set, number of iterations $maxIter$, temperature T , cooling rate r

Output: Inlier sample \mathbf{x}^{best}

```

1: Set  $\mathbf{x}^{current} \leftarrow \mathbf{x}$ , and  $\mathbf{x}^{best} \leftarrow \mathbf{x}$ 
2: for  $j \in S$  do
3:    $x_j^{current} \leftarrow$  sample uniformly from the set  $\{X_{i,j} : \mathbf{X}_i \in \mathcal{D}_n \text{ and } \mathbf{X}_{i,S} \in C_S(\mathbf{x}, \mathcal{Y}^*)\}$ ;
   /* Generate  $\mathbf{x}^{current} = (z_S, \mathbf{x}_{\bar{S}})$  with  $z_S$  drawn using  $z_S \sim \prod_{i \in S} \hat{P}_{X_j | \mathbf{X}_{\bar{S}} \in C_S(\mathbf{x}, \mathcal{Y}^*)}$ . */
4:    $x_j^{best} \leftarrow x_j^{current}$ ;
   /* Initialize  $x^{best}$  */
5: for  $it$  from 1 to  $maxIter$  do
6:    $\mathbf{x}^{new} \leftarrow \mathbf{x}^{current}$ 
7:    $S' \leftarrow$  sample uniformly from the set  $S$ 
8:   for  $j$  in  $S'$  do
9:      $x_j^{new} \leftarrow$  sample uniformly from the set  $\{X_{i,j} : \mathbf{X}_i \in \mathcal{D}_n \text{ and } \mathbf{X}_{i,S} \in C_{S'}(\mathbf{x}, \mathcal{Y}^*)\}$ 
10:   Compute the Outlier score difference  $\Delta O$  between  $\mathbf{x}^{new}$  and  $\mathbf{x}^{current}$ 
11:   if  $\Delta O < 0$  or  $\exp(-\Delta O/T) > \text{random}(0, 1)$  then
12:     Set  $\mathbf{x}^{current} \leftarrow \mathbf{x}^{new}$ 
13:   if Outlier score of  $\mathbf{x}^{best} <$  Outlier score of  $\mathbf{x}^{current}$  then
14:     Set  $\mathbf{x}^{best} \leftarrow \mathbf{x}^{current}$ 
15:   Decrease  $T$  by  $T = T * r$ 
16: return  $\mathbf{x}^{best}$ 

```

7 Experiments

To demonstrate the performance of our framework, we conduct two experiments on real-world datasets. In the first experiment, we showcase the utility of the Local Counterfactual Rules for explaining a regression model. In the second experiment, we compare our approaches with two baseline methods in the context of classification problems: (1) CET [Kanamori, 2022], which partitions the input space using a decision tree and associates a vector perturbation for each leaf, (2) AReS [Rawal, 2020] performs an exhaustive search for finding global counterfactual rules. We use the implementation of [Kanamori, 2022] that adapts AReS for returning counterfactuals samples instead of rules. We compare the methods only in classification problem as all prior works do not deal with regression problems. In all experiments, we split our dataset into train (75%) - test (25%), and we learn a model f , a LightGBM ($estimators=50$, $nb\ leaves=8$), on the train set, which is the model we want to explain. We learn f 's predictions on the train set with a RF ($estimators=20$, $max\ depth=10$): that will be used to generate the CR with $\pi = 0.9$. The parameters used for AReS, CET are $max\ rules=8$, $bins=10$ and $max\ iterations=1000$, $max\ leaf=8$, $bins=10$ respectively. The other parameters are detailed in Appendix (14).

We evaluate the methods on test set using three metrics. The first, *Accuracy*, measures the average number of instances for which the prescribed action by each method changes the prediction to the desired outcome. The second, *Plausibility*, measures the average number of inliers (predicted by an Isolation Forest) among the generated counterfactual samples. The third, *Sparsity*, measures the average number of features that have been changed. For the global counterfactual

methods (AReS, R-CR), which do not guarantee to cover all instances, we compute the *Coverage*, corresponding to the average number of observations for which they propose a recourse.

Local counterfactual rules for regression. We apply our approach to the California House Price dataset (n=20640, p=8) [Kelley Pace, 1997], which contains information about each district such as income, population, and location, and the goal is to predict the median house value of each district. To demonstrate the effectiveness of our Local CR method, we focus on a subset of the test set consisting of 1566 houses with prices lower than 100k. We aim to find recourse that would increase house prices, bringing them within the target range $\mathcal{Y}^* = [200k, 250k]$. For each instance \mathbf{x} , we compute the Minimal Divergent Explanation S , the Local CR $C_S(\mathbf{x}, [200k, 250k])$, and generate a counterfactual sample using the Simulated Annealing technique described earlier. We succeed in changing the decision for all observations, achieving *Accuracy* = 100%. Furthermore, the majority of counterfactual samples passed the outlier test, with a *Plausibility* score of 0.92. Additionally, our Local CR method achieves high sparsity, with *Sparsity* = 4.45.

For instance, the Local CR for the observation $\mathbf{x} = [\text{Longitude}=-118.2, \text{latitude}=33.8, \text{housing median age}=26, \text{total rooms}=703, \text{total bedrooms}=202, \text{population}=757, \text{households}=212, \text{median income}=2.52]$ is $C_S(\mathbf{x}, [200k, 250k]) = [\text{total room} \in [2132, 3546], \text{total bedrooms} \in [214, 491]]$ with probability 0.97. This means that if the total number of rooms and total bedrooms satisfy the conditions in $C_S(\mathbf{x}, [200k, 250k])$, and the remaining features of \mathbf{x} are fixed, then the probability that the price falls within the target set $\mathcal{Y}^* = [200k, 250k]$ is 0.97.

Comparisons of Local and Regional CR with baselines (AReS, CET). We evaluate our framework on three real-world datasets: Diabetes (n=768, p=8) [Kaggle, 2016] aims to predict whether a patient has diabetes or not, Breast Cancer Wisconsin (BCW, n=569, p=32) [Dua, 2017a] aims to predict whether a tumor is benign or malignant, and Compas (n=6172, p=12) [Washington, 2018] is used to predict criminal recidivism. Our evaluation reveals that AReS and CET are highly sensitive to the number of bins and the maximal number of rules or actions, as previously noted by [Ley, 2022]. Poor parameterization can result in completely useless recourses. Furthermore, these methods require separate models for each target class, while our framework only requires a single RF with good precision.

Table 5.1: Results of the *Accuracy* (Acc), *Plausibility*, and *Sparsity* (Sprs) of the different methods. We compute each metric according to the positive (Pos) and negative (Neg) class.

	COMPAS						BCW						Diabetes					
	Acc		Psb		Sps		Acc		Psb		Sps		Acc		Psb		Sps	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
L-CR	1	0.9	0.87	0.73	2	4	1	1	0.96	1	9	7	0.97	1	0.99	0.8	3	4
R-CR	0.9	0.98	0.74	0.93	2	3	0.89	0.9	0.94	0.93	9	9	0.99	0.99	0.9	0.87	3	4
AReS	0.98	1	0.8	0.61	1	1	0.63	0.34	0.83	0.80	4	3	0.73	0.60	0.77	0.86	1	1
CET	0.85	0.98	0.7	0	2	2	1	0.21	0.6	0.80	8	2	0.84	1	0.60	0.20	6	6

Table 5.1 demonstrates that the Local and Regional CR methods achieve high accuracy in changing decisions on all datasets, surpassing AReS and CET by a significant margin on BCW and Diabetes. Furthermore, the baselines struggle to simultaneously change both the positive and negative classes, e.g., CET has *Acc*=1 in the positive class, and 0.21 for the negative class on

BCW or when they have a good *Acc*, the CE are not plausible. For instance, CET has $Acc=0.98$ and $Psb=0$ on Compas, meaning that all the counterfactual samples are outliers. Regarding the coverage of the global CE, CET covers all the instances as it partitions the input space, but we observe that AReS has a smaller *Coverage*= $\{0.43, 0.44, 0.81\}$ compared to the Regional CR, which has $\{1, 0.7, 1\}$ for BCW, Diabetes, and Compas respectively.

Noisy responses robustness of Local CR: To assess the robustness of our approach against noisy responses, we conduct an experiment inspired by [Pawelczyk, 2022]. We normalized the datasets so that $\mathbf{X} \in [0, 1]^p$ and added small Gaussian noises ϵ to the prescribed recourses, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$, where σ^2 took values of 0.01, 0.025, 0.05. We compute the *Stability*, which is the fraction of unseen instances where the action and perturbed action lead to the same output, for the Compas and Diabetes datasets. We used the simulated annealing approach of Section 6 with the Local CR to generate the actions. The Stability metrics for the different noise levels were 0.98, 0.98, 0.98 for Compas and 0.96, 0.97, 0.96 for Diabetes.

In summary, our CR approach is easier to train, and provides more accurate and plausible rules than the baseline methods. Furthermore, our resulting CE is robust against noisy responses.

8 Conclusion

We propose a novel approach that formulates CE as *Counterfactual Rules*. These rules are simple policies that can change the decision of an individual or sub-population with a high probability. Our method is designed to learn robust, plausible, and sparse adversarial regions that indicate where observations should be moved to satisfy a desired outcome. Random Forests are central to our approach, as they provide consistent estimates of the probabilities of interest and naturally give rise to the counterfactual rules we seek. This also allows us to handle regression problems and continuous features, making our method applicable to a wide range of data sets where tree-based models perform well, such as tabular data [Grinsztajn, 2022]. An interesting avenue to explore would be to incorporate the l_1 cost into our approach. Currently, our method aims to minimize the l_0 distance between the query \mathbf{x}^{obs} and the counterfactual \mathbf{x}^{CF} by altering as few features as possible. However, deriving a counterfactual observation within a counterfactual rule that minimizes the l_1 cost is straightforward with an explicit solution. Given the counterfactual rules (hyperrectangles), represented as a box (\mathbf{l}, \mathbf{r}) , with $\mathbf{l}, \mathbf{r} \in \mathbb{R}^p$, the following optimization problem $\mathbf{x}^{CF} = \underset{\mathbf{x}}{\operatorname{argmin}} d(\mathbf{x}, \mathbf{x}^{obs})$ such that $\mathbf{l} \leq \mathbf{x} \leq \mathbf{r}$ has a closed form solution when the distance is the l_1 or l_2 norm. The solution is $\mathbf{x}^{CF} = \max(\mathbf{l}, \min(\mathbf{r}, \mathbf{x}^{obs}))$ elementwise [Carreira-Perpiñán, 2021]. In future work, we will incorporate the l_1 constraint and assess the effectiveness of our approach in terms of cost relative to other methods.

Chapter 6

Adaptive Conformal Prediction by Reweighting Nonconformity Score

Abstract

Despite attractive theoretical guarantees and practical successes, Predictive Interval (PI) given by Conformal Prediction (CP) may not reflect the uncertainty of a given model. This limitation arises from CP methods using a constant correction for all test points, disregarding their individual epistemic uncertainties, to ensure coverage properties. To address this issue, we propose using a Quantile Regression Forest (QRF) to learn the distribution of nonconformity scores and utilizing the QRF's weights to assign more importance to samples with residuals similar to the test point. This approach results in PI lengths that are more aligned with the model's uncertainty or the epistemic uncertainty. Further, the weights learnt by the QRF provide a partition of the features space, allowing for more efficient computations and improved adaptiveness of the PI through groupwise calibration. Our approach enjoys an assumption-free finite-sample marginal and training-conditional or PAC coverage, and under suitable assumptions, it also ensures asymptotic conditional coverage. Our methods work for any nonconformity score and are available as a [Python package](#). We conduct experiments on simulated and real-world data that demonstrate significant improvements compared to existing methods.

Contents

1	Motivations	100
2	Related works and contributions	102
3	Random Forest Localizer	104
4	Weighted Conformal Prediction	106
5	Asymptotic conditional coverage	111
6	Experiments	112

1 Motivations

Machine learning techniques offer single point predictions, such as mean estimates for regression and class labels for classification, without providing any indication of uncertainty or reliability. This can be a major concern in high-stakes applications where precision is vital.

Consider a training set $\mathcal{D}_m = \{(\mathbf{X}_i, Y_i)\}_{i=1}^m$ with $(\mathbf{X}_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ drawn exchangeably from $P = P_{\mathbf{X}}P_{Y|\mathbf{X}}$, and an algorithm \mathcal{A} that gives an estimate of the conditional mean or quantile $\mathcal{A}(\mathcal{D}_m) = \hat{\mu}(\cdot)$. We consider the problem of constructing a predictive set $\hat{C}(\cdot)$ for the unseen response Y_{n+1} given a new feature \mathbf{X}_{n+1} . Conformal Prediction is a universal framework that constructs a prediction interval $\hat{C}(\mathbf{X}_{n+1})$ that covers Y_{n+1} with finite-sample coverage guarantee without any assumption on P and $\hat{\mu}$. CP methods can be broadly divided into two categories: those that involve retraining the model multiple times, such as full conformal [Vovk, 2005] or jackknife methods [Barber, 2021], and those that use sample splitting, known as split conformal methods [Papadopoulos, 2002; Lei, 2016]. The latter is more computationally feasible at the cost of splitting the data. In this Chapter, we consider the split conformal approach (split-CP).

The foundation of the PI of the CP framework is the nonconformity score $\hat{V}(\mathbf{X}, Y)$ that represents the error of the model $\hat{\mu}$ on (\mathbf{X}, Y) . Given a calibration set $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, training set $\mathcal{D}_m = \{(\mathbf{X}_i, Y_i)\}_{i=1}^m$ all drawn exchangeably from $P = P_{\mathbf{X}}P_{Y|\mathbf{X}}$, and the scores $\hat{V}_i := \hat{V}(\mathbf{X}_i, Y_i)$ for all $i \in \mathcal{D}_n$, the PI of \mathbf{X}_{n+1} at level $1 - \alpha$ given by the split-CP is:

$$\hat{C}(\mathbf{X}_{n+1}) = \left\{ y \in \mathcal{Y} : \hat{V}(\mathbf{X}_{n+1}, y) \leq \mathcal{Q}\left(1 - \alpha; \hat{F}_{n+1}\right) \right\}, \quad (6.1)$$

where $\mathcal{Q}(1 - \alpha; F)$ denotes the $(1 - \alpha)$ -quantile of any cumulative distribution function (c.d.f) F , and $\hat{F}_{n+1}(\cdot)$ is the empirical c.d.f of the samples $\hat{V}_{1:n} \cup \infty$ defined as $\hat{F}_{n+1}(r) = \sum_{i=1}^n \frac{1}{n+1} \mathbb{1}_{\hat{V}_i \leq r} + \frac{1}{n+1} \mathbb{1}_{\infty \leq r}$. By exchangeability of the $n + 1$ data points $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n), (\mathbf{X}_{n+1}, Y_{n+1})$, we have that the PI satisfies marginal coverage, i.e.,

$$\mathbb{P}_{P^{n+1}} \left\{ Y_{n+1} \in \hat{C}(\mathbf{X}_{n+1}) \right\} \geq 1 - \alpha.$$

$\mathbb{P}_{P^{n+1}}$ denotes that the probability is taken with respect to the $n + 1$ data points and $\alpha \in (0, 1)$ is a predefined miscoverage rate. However, despite the marginal guarantees, split-CP cannot represent the variability of the model's uncertainty given \mathbf{X}_{n+1} . Indeed, it constructs the PI of future test points \mathbf{X}_{n+1} through the uniform distribution over the calibration residuals $\hat{F}_{n+1}(\cdot)$ that treat all the calibration residuals as the same regardless of \mathbf{X}_{n+1} . To better illustrate the issue, consider a simple example where the true distribution of Y is homoskedastic, meaning that $Y = \mu(\mathbf{X}) + \epsilon$, where \mathbf{X} and ϵ are independent. In this case, the true residuals of the calibration samples $V_i := V(\mathbf{X}_i, Y_i) = |Y_i - \mu(\mathbf{X}_i)| = |\epsilon|$ are independent of \mathbf{X}_i and $V_i \sim |\epsilon|$ for $i \in \mathcal{D}_n$. Hence, we have $F_V(\cdot) = F_{V|\mathbf{X}=x}(\cdot)$. However, in practice, we only have the estimated residuals, $\hat{V}_i := \hat{V}(\mathbf{X}_i, Y_i) = |Y_i - \hat{\mu}(\mathbf{X}_i)| = |\mu(\mathbf{X}_i) - \hat{\mu}(\mathbf{X}_i) + \epsilon|$, which do depend on \mathbf{X}_i as the

accuracy of $\hat{\mu}$ can vary for different \mathbf{X}_i . For example, if \mathbf{X}_i is in a high density region with a large amount of data, $\hat{\mu}$ is likely to be more accurate, while in a low density region with a small amount of data, $\hat{\mu}$ is likely to be less accurate. In contrast of the true residual, the conditional law of the estimated residuals $\hat{V}|\mathbf{X} = \mathbf{x}$ is not equal to the marginal law of \hat{V} , thus using the latter $F_{\hat{V}}(\cdot)$ as in split-CP to construct the PI of a given observation \mathbf{x} may produce under/over coverage PI as $\mathcal{Q}(1 - \alpha; F_{\hat{V}})$ may be greater or lower than $\mathcal{Q}(1 - \alpha; F_{\hat{V}|\mathbf{X}=\mathbf{x}})$.

Our goal is to construct Prediction Intervals (PIs) with valid coverage for the model of interest $\hat{\mu}$, while adjusting the width of the intervals to help visualize and represent the uncertainty of the model $\hat{\mu}$. In fact, the split-CP uses a constant correction term $\mathcal{Q}(1 - \alpha; \hat{F}_{n+1})$ for all test samples, while we aim to have an adaptive correction term that depends on the specific test observation \mathbf{X}_{n+1} . To achieve this, we propose to directly estimate the conditional distribution of the nonconformity score given \mathbf{X}_{n+1} by re-weighting the distribution $\hat{F}_{n+1}(\cdot)$ in order to favor the residuals $\{\hat{V}_i\}_{i \in \mathcal{D}_n}$ closer to the residual of \mathbf{X}_{n+1} . In Figure 6.1, we show the correction terms of split-CP, our method, and the true error of the model $\hat{\mu}$ computed on California house price dataset [Kelley Pace, 1997]. It shows that our corrections are more aligned with the true error of the model.

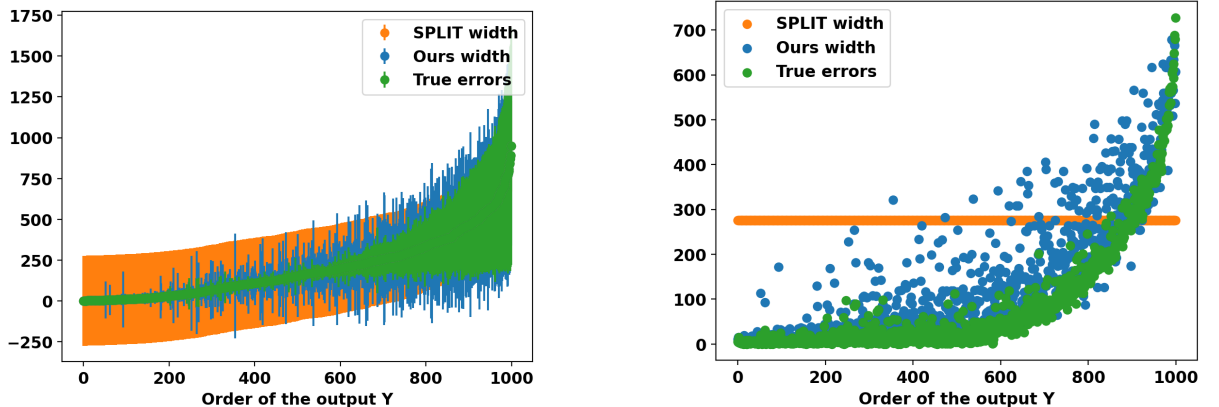


Figure 6.1: Correction terms of SPLIT, Ours, and the true error on the California house price dataset

We also aim to give PI with stronger coverage guarantee. Indeed, in practical applications, what is of interest is the coverage rate on future test points based on a given calibration set. However, the marginal coverage does not address this concern. It only bounds the coverage rate on average over all possible sets of calibration and test observations. In contrast, the training-conditional coverage ensures that with probability $1 - \delta$ over the calibration samples \mathcal{D}_n , the resulting coverage on future test observation is still above $1 - \alpha$. Formally,

$$\mathbb{P}_{P^n} \left\{ \mathbb{P}_P \left\{ Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{D}_n \right\} \geq 1 - \alpha \right\} \geq 1 - \delta.$$

This style of guarantee is also known as ‘‘Probably Approximately Correct’’ (PAC) predictive interval [Valiant, 1984]. The roots of this type of guarantee can be traced back to the earlier works of [Wilks, 1941; Wald, 1943]. Despite the importance of training-conditional coverage

in practice, only a few methods have been proven to achieve it. [Vovk, 2012] was the first to establish this result for split conformal methods, and recently [Bian, 2022] has shown that the K-fold CV+ method also achieves it. However, no analogous results are currently known for other CP methods, such as jackknife+ [Barber, 2021] and full-conformal [Vovk, 2005]. Therefore, we propose a further calibration step such that our proposed adaptive PI also achieves training-conditional or PAC coverage.

There is another area of research that focuses on developing CP procedures for conditional coverage $\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1}) \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1}) \geq 1 - \alpha$. It is well known that obtaining nontrivial distribution-free conditional coverage is impossible with a finite sample size [Lei, 2014b; Vovk, 2012; Barber, 2019b]. Consequently, we prove under suitable assumptions that our methods also achieve asymptotic conditional coverage.

2 Related works and contributions

For the sake of simplicity, we use the absolute residual as the nonconformity score $\widehat{V}_i := \widehat{V}(\mathbf{X}_i, Y_i) = |Y_i - \widehat{\mu}(\mathbf{X}_i)|$, without loss of generality. As a result, the best (symmetric) PI that can be constructed with $\widehat{\mu}(\cdot)$ and the score $\widehat{V}(\cdot)$ is $C^*(\mathbf{X}_{n+1}) = [\widehat{\mu}(\mathbf{X}_{n+1}) \pm q_{1-\alpha}^*(\mathbf{X}_{n+1})]$ where $q_{1-\alpha}^*(\mathbf{X}_{n+1})$ is the $(1 - \alpha)$ -quantile of $F_{\widehat{V}_{n+1} \mid \mathbf{X}_{n+1}}$. To construct adaptive PIs, we propose focusing on the estimated residuals of the calibration samples $\{\widehat{V}_i\}_{i \in \mathcal{D}_n}$, and approximate the distribution of $\widehat{V} \mid \mathbf{X} = \mathbf{x}$ or identify the stable regions A where $\mathbb{V}_P(\widehat{V}(\mathbf{X}, Y) \mid \mathbf{X} \in A) \approx 0$, which would allow us to isolate the regions where there is high/low uncertainty of the model.

Recently, [Guan, 2022] proposed Localized Conformal Prediction (LCP) and [Han, 2022] inspired by [Lin, 2021] proposed Split Localized Conformal Prediction (SLCP) which uses kernel-based weights $w_h(\mathbf{x}, \mathbf{X}_i)$ or Nadaraya-Watson (NW) estimator [Nadaraya, 1964] to approximate the conditional c.d.f of $\widehat{V} \mid \mathbf{X} = \mathbf{x}$. Both methods differ in how they learn the NW estimator, SLCP uses the training data \mathcal{D}_m to learn the estimator $\widehat{F}_h^{(S)}(r \mid \mathbf{X} = \mathbf{x}) = \sum_{i \in \mathcal{D}_m} w_h(\mathbf{x}, \mathbf{X}_i) \mathbb{1}_{\widehat{V}_i \leq r}$, while LCP uses the calibration data \mathcal{D}_n to learn the estimator $\widehat{F}_h^{(L)}(r \mid \mathbf{X} = \mathbf{x}) = \sum_{i \in \mathcal{D}_n} w_h(\mathbf{x}, \mathbf{X}_i) \mathbb{1}_{\widehat{V}_i \leq r}$. The calibration step of these two methods is also different. The PI of \mathbf{X}_{n+1} given by SLCP is:

$$C^S(\mathbf{X}_{n+1}) = \left[\widehat{\mu}(\mathbf{X}_{n+1}) \pm \mathcal{Q}\left(1 - \alpha; \widehat{F}_h^{(S)}(\cdot \mid \mathbf{X} = \mathbf{X}_{n+1})\right) + \widehat{Q} \right]$$

where \widehat{Q} is the split-CP correction term to achieve marginal coverage. In contrast, LCP does not use split-CP but instead adapts the threshold $\tilde{\alpha} = 1 - \alpha$ in $\mathcal{Q}\left(1 - \alpha; \widehat{F}_h^{(L)}(\cdot \mid \mathbf{X}_{n+1})\right)$ to achieve the marginal coverage. LCP constructs the predictive interval for a new point \mathbf{X}_{n+1} as follows:

$$C^L(\mathbf{X}_{n+1}) = \left[\widehat{\mu}(\mathbf{X}_{n+1}) \pm \mathcal{Q}\left(\tilde{\alpha}; \widehat{F}_h^{(L)}(\cdot \mid \mathbf{X} = \mathbf{X}_{n+1})\right) \right]$$

where $\tilde{\alpha}$ is chosen to achieve the marginal coverage. However, while both LCP and SLCP address the problem and guarantee marginal coverage, they have some limitations. A main limitation is that they are based on kernel methods, which are known to be limited in high dimensions due

to the curse of dimensionality. Additionally, choosing the appropriate kernel bandwidth can be challenging and it can be difficult to define kernels that handle both categorical and continuous variables. Another limitation of SLCP is that it learns $\hat{F}_h^{(S)}(\cdot|\mathbf{X} = \mathbf{x})$ on the training data \mathcal{D}_m , which may result in overfitting and thus the calibration step using split-CP may produce large intervals to attain the marginal coverage. In contrast, LCP learns $\hat{F}_h^{(L)}(\cdot|\mathbf{X} = \mathbf{x})$ on the calibration data \mathcal{D}_n , but the calibration step that consists of finding the adaptive $\tilde{\alpha}$ is computationally costly.

In this work, we propose to replace the Nadaraya-Watson (NW) estimator with the Quantile Regression Forest (QRF) algorithm [Meinshausen, 2006] to estimate the distribution $\hat{V}|\mathbf{X} = \mathbf{x}$ and use the LCP approach to calibrate the PI. The QRF algorithm is an adaptation of the Random Forest (RF) algorithm [Breiman, 1984], which can be seen as an adaptive neighborhood procedure [Lin, 2006]. It estimates the conditional c.d.f of $\hat{V}|\mathbf{X} = \mathbf{x}$ as $\hat{F}(r|\mathbf{X} = \mathbf{x}) = \sum_i w_n(\mathbf{x}, \mathbf{X}_i) \mathbf{1}_{\hat{V}_i \leq r}$ where the weights correspond to the average number of times where \mathbf{X}_i falls in the same leaves of the RF as the observation \mathbf{x} . Unlike kernel-based methods, the weights given by the RF depend on both feature input \mathbf{X}_i and the residual \hat{V}_i due to the splits. We called this approach LCP-RF. This estimator has several advantages over the NW estimator. First, it is known to perform well in practice, even in high dimensions. It can handle both categorical and continuous variables. Additionally, it has interesting theoretical properties in high dimensions; under certain assumptions, it can be shown to be consistent and to adapt to the intrinsic dimension [Klusowski, 2021; Scornet, 2015]. To illustrate, we compute the PI of these methods using a random forest fitted on a toy model with input $\mathbf{X} \in [0, 7]^{21}$, and the target defined as $Y = \sin(X_1)^2 + 0.1 + 0.6 \times \epsilon \times \sin(2X_1)$, where $\epsilon \sim \mathcal{N}(0, 1)$, and $X_i \sim \mathcal{U}(0, 7)$ for all $i \in [21]$. As seen in Figure 6.2, the competitors LCP and SLCP fail to perform well even on this very simple example with 1 active and 20 noise features, while our method benefits from the power of the Random Forest algorithm on tabular data [Grinsztajn, 2022].

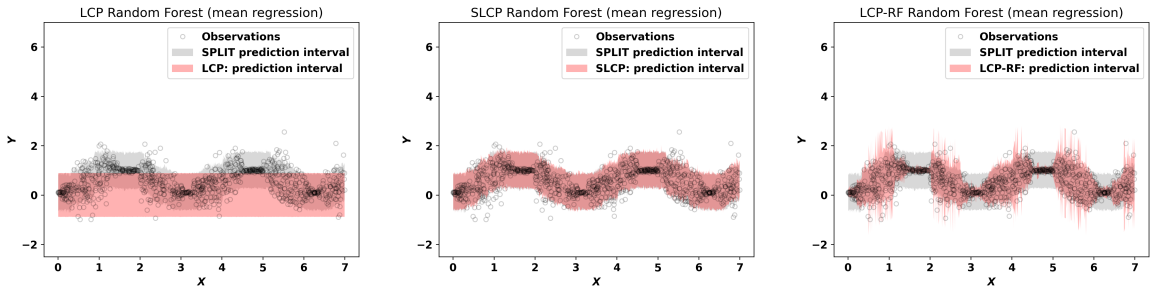


Figure 6.2: Predictive interval at level $1 - \alpha$ of SLCP, LCP and LCP-RF of a random Forest fitted on toy model (\mathbf{X}, Y) , $\mathbf{X} \in [0, 7]^{21}$ and the target is defined as $Y = \sin(X_1)^2 + 0.1 + 0.6 \times \epsilon \times \sin(2X_1)$ with $\epsilon \sim \mathcal{N}(0, 1)$, and $X_i \sim \mathcal{U}(0, 7)$ for all $i \in [21]$.

Additionally, we show that the learned weights of the RF can be used to create a relevant partition or groups/clusters of the input space. This allows for a more efficient computation of the LCP calibration and also allows for groupwise calibration to give a more adaptive PI. In practice, it is often desirable to have a stronger coverage guarantee than marginal coverage. Consequently,

we propose a further calibration step such that our PI satisfies training-conditional coverage. We also show that it achieves conditional coverage guarantee under suitable assumptions.

An active area of research involves using a better nonconformity score to provide an adaptive prediction interval considering the variability of $Y|\mathbf{X} = \mathbf{x}$. Several methods have been proposed such as Conformal Quantile Regression (CQR) [Romano, 2019], which uses score functions based on estimated quantiles, Locally Adaptive Split Conformal methods [Lei, 2016] which use a scaled residual, and [Izbicki, 2020] proposed using the estimated conditional density as the conformity score. These methods incorporate different nonconformity scores $\widehat{V}(\cdot)$ that are better suited for handling the variability of Y . However, the extracted residuals \widehat{V}_i of these nonconformity scores still depend on and vary according to the input, and the split-CP makes a constant correction for all observations. For instance, consider the CQR interval of the form: $[\hat{q}_{Y|\mathbf{X}}(\alpha_{lo}; \mathbf{x}) - \hat{Q}; \hat{q}_{Y|\mathbf{X}}(\alpha_{hi}; \mathbf{x}) + \hat{Q}]$, where $\hat{q}_{Y|\mathbf{X}}(\cdot; \mathbf{x})$ is an estimator of $\mathcal{Q}(\cdot; F_{Y|\mathbf{X}=\mathbf{x}})$ and \hat{Q} is the split-CP correction term computed using calibration data. The estimators $\hat{q}_{Y|\mathbf{X}}(\alpha_{lo}; \mathbf{x}), \hat{q}_{Y|\mathbf{X}}(\alpha_{hi}; \mathbf{x})$ may exhibit varying precision across different regions of the input space. However, CQR doesn't account for this by inflating the interval using a constant, non-adaptive, \hat{Q} for all points. Our aim is to address the epistemic uncertainty of the estimators used in CQR or any Split-CP approach by proposing an adaptive correction: $[\hat{q}_{Y|\mathbf{X}}(\alpha_{lo}; \mathbf{x}) - \hat{Q}(\mathbf{x}); \hat{q}_{Y|\mathbf{X}}(\alpha_{hi}; \mathbf{x}) + \hat{Q}(\mathbf{x})]$ to adapt to the estimation error of the quantile regressors or any estimator used in split-CP at each input. Methods that give adaptive PI are not competing with the LCP-RF approach as it can be applied to them to improve their PIs.

The main contributions of this Chapter are: (1) Developing an adaptive PI that better represents the uncertainty of a given model $\hat{\mu}$ by using QRF to learn the conditional distribution of the residuals $\widehat{V}(\mathbf{X}, Y)|\mathbf{X} = \mathbf{x}$, and utilizing the LCP framework to calibrate the resulting PI for marginal coverage, (2) Introducing a calibration step to achieve training-conditional or PAC coverage, (3) Exploiting the structure of the weights of the QRF to create groups for more adaptive PI and efficient computation through groupwise calibration, (4) Showing that our methods achieve asymptotic conditional coverage under suitable conditions, (5) Demonstrating through simulations and real-world datasets that our methods outperform competitors LCP and SLCP, and providing a [Python package](#) for the methods.

3 Random Forest Localizer

In this section, we present the RF Localizer for constructing adaptive PI that depends on the test point \mathbf{X}_{n+1} . The approach uses the learned weights of the RF and assigns higher weights to calibration samples that have residuals \widehat{V}_i similar to \widehat{V}_{n+1} . This is based on the RF algorithm's ability to partition the input space by recursively splitting the data, resulting in similar observations with respect to the target variable (here, residuals) within each leaf node of the trees. The basic idea of the trees of the RF is to partition the input space into cells such that $\mathbb{V}_P(\widehat{V}(\mathbf{X}, Y) | \mathbf{X} \in A) \approx 0$ in each cell A . The weight of each calibration sample for \mathbf{X}_{n+1} is determined by the number of times it appears in the leaves of the trees where \mathbf{X}_{n+1} falls.

The Random Forest (RF) is an ensemble learning method that utilizes the bagging principle [Breiman, 1996] to combine k randomized trees derived from the CART algorithm [Breiman, 1984]. Each tree is constructed using a random sample of the training data with replacement, and the best split at every node is identified by optimizing the CART-criterion among a random subset of variables. The predictions from all trees are then averaged to produce the final output of the forest. The Random Forest estimator can also be seen as an adaptive neighborhood procedure [Lin, 2006; Biau, 2010]. Let assume we have trained the RF on \mathcal{D}_n , then for every instance \mathbf{x} , the observations in \mathcal{D}_n are weighted by $w_n^{RF}(\mathbf{x}, \mathbf{X}_i)$, $i = 1, \dots, n$. Therefore, the prediction of Random Forests and the weights can be rewritten as $m(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n) = \sum_{i=1}^n w_n^{RF}(\mathbf{x}, \mathbf{X}_i) Y_i$ and

$$w_n^{RF}(\mathbf{x}, \mathbf{X}_i) = \frac{1}{k} \sum_{l=1}^k \frac{B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}; \Theta_l)}}{N_n(A_n(\mathbf{x}; \Theta_l))}, \quad (6.2)$$

where $\Theta_l, l = 1, \dots, k$ are independent random vectors that represent the observations that are used to build each tree, i.e., the bootstrap samples, and the random subset of splitting candidate variables used in each node. $A_n(\mathbf{x}; \Theta_l)$ is the tree leaf (cell) containing \mathbf{x} , $N_n(A_n(\mathbf{x}; \Theta_l))$ is the number of bootstrap elements of \mathcal{D}_n that fall into $A_n(\mathbf{x}; \Theta_l)$, $B_n(\mathbf{X}_i; \Theta_l)$ is the number of times \mathbf{X}_i has been chosen from the training data, and $w_n^{RF}(\mathbf{x}, \mathbf{X}_i)$ represents the average number of times \mathbf{X}_i appears in the same leaves as \mathbf{x} .

Random Forests can be used to estimate more complex quantities, such as cumulative hazard function [Ishwaran, 2008], treatment effect [Wager, 2017; Jocteur, 2023], and conditional density [Du, 2021]. Quantile Regression Forests proposed by [Meinshausen, 2006] use the same weights $w_n^{RF}(\mathbf{x}, \mathbf{X}_i)$ as Random Forests to approximate the c.d.f $F(y|\mathbf{X} = \mathbf{x})$ as $\sum_{i=1}^n w_n^{RF}(\mathbf{x}, \mathbf{X}_i) \mathbb{1}_{Y_i \leq y}$.

Random Forest Localizer. To approximate the estimated residuals $\widehat{V}|\mathbf{X} = \mathbf{x}$, we propose to fit a Quantile Regression Forest $\widehat{F}(\cdot|\mathbf{x})$ on the calibration data residuals $\widehat{\mathcal{D}}_n = \{(\mathbf{X}_i, \widehat{V}_i)\}_{i=1}^n$, and the estimator is defined as

$$\widehat{F}(r|\mathbf{x}) = \sum_{i=1}^{n+1} w_n(\mathbf{x}, \mathbf{X}_i) \mathbb{1}_{\widehat{V}_i \leq r}, \quad w_n(\mathbf{x}, \mathbf{X}_i) = \frac{1}{k} \sum_{l=1}^k \frac{B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}; \Theta_l)}}{N_{n+1}(A_n(\mathbf{x}; \Theta_l))} \quad (6.3)$$

where $\widehat{V}_{n+1} = +\infty$ unless specified and $N_{n+1}(A_n(\mathbf{x}; \Theta_l)) = \sum_{i=1}^{n+1} B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{[\mathbf{X}_i \in A_n(\mathbf{x}; \Theta_l)]}$ with $B_n(\mathbf{X}_{n+1}; \Theta_l) = 1$ so that $\sum_{i=1}^{n+1} w_n(\mathbf{x}, \mathbf{X}_i) = 1$. It's worth noting that this estimator (6.3) is slightly different from (6.2), as it includes the observation \mathbf{X}_{n+1} in the weighted sum, and for any \mathbf{x} , $w_n(\mathbf{x}, \mathbf{X}_i)$ is computed using $\{\mathbf{X}_i\}_{i=1}^n$ and the test observation \mathbf{X}_{n+1} . We will see later that this addition would be essential to prove the marginal coverage property of our method. Using this estimator, a natural PI for \widehat{V}_{n+1} is:

$$\widehat{C}_V(\mathbf{X}_{n+1}) = \left\{ v : v \leq \mathcal{Q} \left(1 - \alpha; \widehat{F}(\cdot|\mathbf{X}_{n+1}) \right) \right\}. \quad (6.4)$$

Recall that we obtain the prediction interval $\widehat{C}(\mathbf{X}_{n+1})$ for Y_{n+1} by inverting the nonconformity score $\widehat{V}(\mathbf{X}_{n+1}, \cdot)$ using $\widehat{C}_V(\mathbf{X}_{n+1})$ as in Equation (6.1). Thus, the real quantity of interest is $\widehat{C}_V(\mathbf{X}_{n+1})$. The question at hand is whether the PI $\widehat{C}_V(\mathbf{X}_{n+1})$ defined in (6.4) satisfies the marginal coverage. If $w_n(\mathbf{X}_{n+1}, \mathbf{X}_i) = \frac{1}{n+1}$, we have $\mathcal{Q}\left(1 - \alpha; \widehat{F}(\cdot | \mathbf{X}_{n+1})\right) = \widehat{V}_{(\lceil(1-\alpha)(n+1)\rceil)}$ and thanks to the quantile lemma (Chapter 1, Lemma 2.3) and exchangeability of the \widehat{V}_i , we have the marginal coverage. However, if $\widehat{F}(\cdot | \mathbf{X}_{n+1})$ gives non-equal weights to the calibration samples, it is no longer the case. Recent methods have been proposed by [Tibshirani, 2019] and [Barber, 2022] that achieve marginal coverage when using reweighting. However, these methods cannot be applied to calibrate our PI, as they work under different assumptions. The method introduced by [Barber, 2022] assumes that the weights do not depend on the data, while the method proposed by [Tibshirani, 2019] handles data-dependent weights but assumes a covariate shift, where the training and test data have different input distributions but the same conditional distribution $P_{Y|\mathbf{X}}$.

To calibrate our PI, we use the Localized Conformal Prediction (LCP) framework [Guan, 2022] to select an appropriate level $\tilde{\alpha}$ of the quantile used in the PI (6.4) to ensure marginal coverage at level $1 - \alpha$. Hence, the PI becomes

$$\widehat{C}_V(\mathbf{X}_{n+1}) = \left\{v : v \leq \mathcal{Q}\left(\tilde{\alpha}; \widehat{F}(\cdot | \mathbf{X}_{n+1})\right)\right\}. \quad (6.5)$$

4 Weighted Conformal Prediction

In this section, we give a comprehensive overview of the LCP framework of [Guan, 2022] with the Random Forest Localizer for completeness. Additionally, we describe our calibration approach that guarantees training-conditional or PAC coverage, and how we leverage the weights of the RF to improve the LCP calibration process and produce more adaptive prediction intervals. For ease of reading, we follow [Guan, 2022] and introduce $\mathcal{F}_i = \widehat{F}(\cdot | \mathbf{X}_i) = \sum_{j=1}^n w_n(\mathbf{X}_i, \mathbf{X}_j) \mathbb{1}_{\widehat{V}_j \leq \cdot} + w_n(\mathbf{X}_i, \mathbf{X}_{n+1}) \mathbb{1}_{\widehat{V}_{n+1} \leq \cdot}$ as the estimated c.d.f of \widehat{V} given \mathbf{X}_i by the RF Localizer. As \widehat{V}_{n+1} is not observed and we need to consider the possible values of \widehat{V}_{n+1} for constructing the PI, we introduce the additional notations \mathcal{F}_i^v for the estimated c.d.f \mathcal{F}_i when $\widehat{V}_{n+1} = v$ if v is finite, and \mathcal{F}_{n+1}^∞ if $\widehat{V}_{n+1} = +\infty$.

4.1 Localized Conformal Prediction [Guan, 2022]

The following lemma is the cornerstone of the LCP framework. It shows how to achieve marginal coverage by properly selecting the level $\tilde{\alpha}$ of the quantile of the localizer.

Lemma 4.1. *Let $\tilde{\alpha}$ be the smallest value in $\Gamma = \left\{\sum_{j=1}^k w_n(\mathbf{X}_i, \mathbf{X}_j) : i, k \in \{1, \dots, n+1\}\right\}$ such that*

$$\sum_{i=1}^{n+1} \frac{1}{n+1} \mathbb{1}_{\widehat{V}_i \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_i)} \geq 1 - \alpha, \quad (6.6)$$

then $\mathbb{P}_{P^{n+1}} \left\{\widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1})\right\} \geq 1 - \alpha$, or equivalently $\mathbb{P}_{P^{n+1}} \left\{\widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}^\infty)\right\} \geq 1 - \alpha$.

Proof. Recall that both $\tilde{\alpha}$ and \mathcal{F}_{n+1} depend on $\widehat{\mathcal{D}}_n = \{(\mathbf{X}_i, \widehat{V}_i)\}_{i=1}^n$ and $(\mathbf{X}_{n+1}, \widehat{V}_{n+1})$, but we won't specify them for clarity. Let us define the event $E_{n+1} = \{\widehat{Z}_1 = \widehat{z}_1, \dots, \widehat{Z}_{n+1} = \widehat{z}_{n+1}\}$ where $\widehat{Z}_i = (\mathbf{X}_i, \widehat{V}_i)$ and $\widehat{z}_i = (\mathbf{x}_i, \widehat{v}_i) \in \mathcal{X} \times \mathcal{Y}$. The exchangeability of the residuals implies that $\widehat{V}_{n+1}|E_{n+1}$ is uniform on the set $\{\widehat{v}_1, \dots, \widehat{v}_{n+1}\}$, and

$$\begin{aligned} \mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}) \mid E_{n+1} \right\} &= \sum_{i=1}^{n+1} \mathbb{P}_{P^{n+1}}(\widehat{V}_{n+1} = \widehat{v}_i \mid E_{n+1}) \mathbb{1}_{\widehat{v}_i \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_i)} \\ &= \sum_{i=1}^{n+1} \frac{1}{n+1} \mathbb{1}_{\widehat{v}_i \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_i)} \geq 1 - \alpha \quad (\text{By Equation 6.6}) \end{aligned}$$

The formulation $\widehat{V}_{n+1}|E_{n+1}$ aims to provide another way to represent the uniformity of ranks when variables are exchangeable. It corresponds to a scenario where we had observed an unordered set of variables $E_{n+1} = \{\widehat{Z}_1 = \widehat{z}_1, \dots, \widehat{Z}_{n+1} = \widehat{z}_{n+1}\}$ and have forgotten which value v_i each random variable V_j is associated with. By leveraging the uniformity of ranks of exchangeable random variables (see Chapter 1, Lemma 2.2), we establish that $P(V_j = v_i | E_{n+1}) = \frac{1}{n+1}$.

By marginalizing over the event E_{n+1} , we have $\mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}) \right\} \geq 1 - \alpha$. In addition, we can remove the dependence on the unknown residuals \widehat{V}_{n+1} using the well-known fact that $\widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}) \iff \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}^\infty)$ (Chapter 1, proof of Lemma 2.3). Thus, we also have $\mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}^\infty) \right\} \geq 1 - \alpha$. \square

Now, we can use Lemma 4.1 to test $H_0 : \widehat{V}_{n+1} = v$ for each $v \in \overline{\mathbb{R}}$ under exchangeability, then invert the test to construct the PI. $\widehat{C}_V(\mathbf{X}_{n+1})$ consists of all values v that are not rejected by this test. The resulting PI has marginal coverage as shown in the following theorem.

Theorem 4.2. *Given $\widehat{V}_{n+1} = v$, let define $\tilde{\alpha}(v)$ that depends on $\widehat{\mathcal{D}}_n$ and (\mathbf{X}_{n+1}, v) to be the smallest value $\tilde{\alpha}(v) \in \Gamma = \left\{ \sum_{j=1}^k w_n(\mathbf{X}_i, \mathbf{X}_j) : i, k \in \{1, \dots, n+1\} \right\}$ such that*

$$\sum_{i=1}^{n+1} \frac{1}{n+1} \mathbb{1}_{\widehat{v}_i \leq \mathcal{Q}(\tilde{\alpha}(v); \mathcal{F}_i^v)} \geq 1 - \alpha. \quad (6.7)$$

Set $\widehat{C}_V(\mathbf{X}_{n+1}) = \{v : v \leq \mathcal{Q}(\tilde{\alpha}(v); \mathcal{F}_{n+1}^\infty)\}$, $\widehat{C}(\mathbf{X}_{n+1}) = \{y : v \leq \mathcal{Q}(\tilde{\alpha}(v); \mathcal{F}_{n+1}^\infty), v = \widehat{V}(\mathbf{X}_{n+1}, y)\}$, then by construction, Lemma 4.1 gives

$$\mathbb{P}_{P^{n+1}} \left\{ Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1}) \right\} = \mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \in \widehat{C}_V(\mathbf{X}_{n+1}) \right\} = \mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}^\infty) \right\} \geq 1 - \alpha.$$

At this point, the LCP method is not practical as it requires computing $\tilde{\alpha}(v)$ for every possible value of $v \in [0, \infty]$ in order to construct the prediction interval. This process can be extremely time-consuming and computationally intensive. However, [Guan, 2022] shows that the computation of $\widehat{C}_V(\mathbf{X}_{n+1})$ can be done efficiently thanks to its interesting properties. Specifically, if v is accepted in $\widehat{C}_V(\mathbf{X}_{n+1})$, all $v' \leq v$ are also accepted, as $\mathcal{Q}(\tilde{\alpha}; \mathcal{F}_i^v)$ is non-decreasing in both $\tilde{\alpha}$ and v . Hence, it is sufficient to find the largest accepted value v^* . Additionally, as

$\mathcal{Q}(\tilde{\alpha}(v); \mathcal{F}_i^v)$ is monotone and piece-wise constant in v , with value changes only occurring at different $v = \hat{V}_i, i \in [n+1]$, it can be proven that the largest value is attained by one of the residuals \hat{V}_{k^*} with $k^* \in [n+1]$. Therefore, the closure $\bar{C}_V(\mathbf{X}_{n+1})$ of $\hat{C}_V(\mathbf{X}_{n+1})$ is given by $\bar{C}_V(\mathbf{X}_{n+1}) = \{v : v \leq \hat{V}_{k^*}\}$ for some $k^* \in [n+1]$. The following Lemma shows how to find V_{k^*} .

Lemma 4.3. *We denote $\hat{V}_{(1)}, \dots, \hat{V}_{(n)}$ the order statistics of the nonconformity score of the calibration samples, set $\hat{V}_{(0)} = -\infty$, and $\hat{V}_{(n+1)} = +\infty$, and $\tilde{\theta}_k = \sum_{i=1}^n w_n(\mathbf{X}_{n+1}, \mathbf{X}_i) \mathbb{1}_{\hat{V}_i < \hat{V}_{(k)}}$. Let $k^* \in \{1, \dots, n+1\}$ the largest index such that*

$$S(k) := \sum_{i=1}^n \frac{1}{n+1} \mathbb{1}_{\hat{V}_i \leq \mathcal{Q}(\tilde{\theta}_k; \mathcal{F}_i^{\hat{V}_{(k-1)}})} < \alpha. \quad (6.8)$$

Then, $\bar{C}_V(\mathbf{X}_{n+1}) = \{v : v \leq \hat{V}_{(k^*)}\}$ is the closure of $\hat{C}_V(\mathbf{X}_{n+1})$.

[Guan, 2022] also proposed an algorithm that computed $S(k)$ in $\mathcal{O}(n \log(n))$ time. The description of the algorithm can be found in the original paper.

4.2 Training-Conditional coverage for LCP-RF

In this section, we consider training-conditional coverage or PAC predictive interval guarantees for the LCP-RF. We define the coverage rate given a calibration set \mathcal{D}_n as $\text{cov}(\mathcal{D}_n) = \mathbb{P}_P \left\{ \hat{V}_{n+1} \in \hat{C}_V(\mathbf{X}_{n+1}) \mid \mathcal{D}_n \right\}$ where the probability is taken with respect to the test observation $(\mathbf{X}_{n+1}, \hat{V}_{n+1})$. The PAC predictive interval ensures that for most draws of the calibration samples $\mathcal{D}_n \sim P^n$, we have $\text{cov}(\mathcal{D}_n) \geq 1 - \alpha$. Formally, $\exists \delta$ such that

$$\mathbb{P}_{P^n} \{ \text{cov}(\mathcal{D}_n) \geq 1 - \alpha \} \geq 1 - \delta.$$

We use a two-step approach to ensure training-conditional coverage for the LCP-RF. First, we use a portion of the calibration samples to ensure marginal coverage by applying the LCP approach. Next, we use a separate portion of the calibration samples to learn a correction term, which is then added to the LCP-RF approach to ensure training-conditional coverage. This approach is similar to the one used in [Kivaranovic, 2020]. We split the calibration set $\hat{\mathcal{D}}_n$ into two sets $\hat{\mathcal{D}}_{n_i}^i = \{(\mathbf{X}_1^i, \hat{V}_1^i), \dots, (\mathbf{X}_{n_i}^i, \hat{V}_{n_i}^i)\}$ for $i = 1, 2$ with $n_1 + n_2 = n$. We train the Quantile Regression Forest on $\hat{\mathcal{D}}_{n_1}^1$, and compute PI for the observations in the second set $\hat{\mathcal{D}}_{n_2}^2$ using the LCP-RF. The PI of each $i \in \hat{\mathcal{D}}_{n_2}^2$ is $\hat{C}_V(\mathbf{X}_i^2) = \{v : v \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_i^2); \mathcal{F}_i^{2,\infty})\}$, where $\tilde{\alpha}(\mathbf{X}_i^2)$ is the adapted level $\tilde{\alpha}$ to have marginal coverage if \mathbf{X}_i^2 is the test point and $\mathcal{F}_i^{2,\infty} = \sum_{j=1}^{n_1+1} w_n(\mathbf{X}_i^2, \mathbf{X}_j^1) \mathbb{1}_{\hat{V}_j^1 \leq \cdot}$ is the estimated residual distribution learn on $\mathcal{D}_{n_1}^1$ evaluated on \mathbf{X}_i^2 where we set $\mathbf{X}_{n_1+1}^1 = \mathbf{X}_i^2$ and $\hat{V}_{n_1+1}^1 = +\infty$. In this context, for a given test point of interest \mathbf{X}_{n+1} , $\mathcal{F}_{n+1}^\infty = \sum_{j=1}^{n_1+1} w_n(\mathbf{X}_{n+1}, \mathbf{X}_j^1) \mathbb{1}_{\hat{V}_j^1 \leq \cdot}$ with $\mathbf{X}_{n_1+1}^1 = \mathbf{X}_{n+1}$ and $\hat{V}_{n_1+1}^1 = +\infty$ as the QRF is trained on $\hat{\mathcal{D}}_{n_1}^1$.

The following lemma shows how we can correct the corresponding $\tilde{\alpha}(\mathbf{X}_{n+1})$ by adding a correction term $\hat{\alpha}$ to ensure PAC coverage.

Theorem 4.4. *Suppose that all observations are i.i.d. drawn from the distribution P . For any given $\epsilon > 0$ and $\alpha - \epsilon > 0$, let $\hat{\alpha}$ be the smallest value in the uniform grid $T = \{\alpha_1 = \frac{1}{K}, \dots, \alpha_K = 1\}$ of size K such that*

$$\sum_{i=1}^{n_2} \frac{1}{n_2} \mathbf{1}_{\hat{V}_i^2 \leq \mathcal{Q}(1 \wedge (\tilde{\alpha}(\mathbf{X}_i^2) + \hat{\alpha}); \mathcal{F}_i^{2,\infty})} \geq 1 - \alpha. \quad (6.9)$$

Then, we have

$$\mathbb{P}_{P^{n_1}} \{cov(\mathcal{D}_{n_1}) \geq 1 - \alpha - \epsilon\} \geq 1 - \delta, \quad (6.10)$$

with $\delta = K \exp(-2n_2\epsilon^2)$ and $cov(\mathcal{D}_{n_1}) = \mathbb{P}_P \left\{ \hat{V}_{n+1} \leq \mathcal{Q}(1 \wedge (\tilde{\alpha}(\mathbf{X}_{n+1}) + \hat{\alpha}); \mathcal{F}_{n+1}^\infty) \mid \mathcal{D}_{n_1} \right\}$.

Remark. This result is valid under the i.i.d assumption and not under exchangeability as the previous results of this chapter. We suggest choosing a grid $T \subset [0, \alpha]$ as we have observed in most practical scenarios that $\tilde{\alpha}(\mathbf{X}_{n+1}) \approx 1 - \alpha$. In our experiments, a grid of size $K=10$ was effective. However, the central idea remains unaltered - to select a grid that enables transitioning from $\tilde{\alpha}(\mathbf{X}_{n+1})$ to 1. Additionally, as $\tilde{\alpha}(\mathbf{X}_{n+1}) + \hat{\alpha}$ may be above 1, we use $1 \wedge (\tilde{\alpha}(\mathbf{X}_{n+1}) + \hat{\alpha})$ to ensure that it does not exceed 1.

4.3 Clustering using the weights of LCP-RF

In this section, we analyze the weights of the Random Forest Localizer and show that it offers several benefits compared to traditional kernel-based localizer. These benefits include faster computation and more adaptive PIs. One key difference between the RF localizer and kernel-based localizer is that the RF localizer's weights are sparse, i.e., many weights being zero. For a given test point \mathbf{X}_{n+1} , if $w_n(\mathbf{X}_{n+1}, \mathbf{X}_i) = 0$, then the estimated c.d.f \mathcal{F}_i does not depend on the value of \hat{V}_{n+1} . Thus, it may not be necessary to use \mathcal{F}_i in the LCP's marginal calibration (Eq. (6.8) in Lemma 4.3).

The weights defined by the Random Forest Localizer have a structure that can be utilized to group similar observations together before applying the calibration steps. Indeed, we can view the weights of the RF on the calibration set as a transition matrix or a weighted adjacency matrix G where $G_{i,j} = w_n(\mathbf{X}_i, \mathbf{X}_j)$, and $\forall j \in [n]$, we have $\sum_{i=1}^n w_n(\mathbf{X}_j, \mathbf{X}_i) = \sum_{i=1}^n w_n(\mathbf{X}_i, \mathbf{X}_j) = 1$.

To exploit this structure, we propose to group observations that are connected to each other and separate observations that are not connected. This can be done by considering the connected components of the graph represented by the matrix G . Assume that G has L connected components represented by the disjoint sets of vertices G_1, \dots, G_L , defined such that for any $\mathbf{X}_i, \mathbf{X}_j \in G_l$, there is a path from \mathbf{X}_i to \mathbf{X}_j , and they are connected to no other vertices outside the vertices in G_l . This leads to the existence of a partition of the input space $\cup_{i=1}^L R_i = \mathcal{X}$, where $\forall k, l \in [L], R_l \cap R_k = \emptyset$, and for all $\mathbf{X}_i \in R_p, \mathbf{X}_j \in R_q$, we have $w_n(\mathbf{X}_i, \mathbf{X}_j) = 0$. The regions R_i is defined as $R_i = \{\mathbf{x} \in \mathcal{X} : \exists \mathbf{X} \in G_i, w_n(\mathbf{x}, \mathbf{X}) > 0 \text{ and } \forall \mathbf{X}' \in G_k, k \neq i, w_n(\mathbf{x}, \mathbf{X}') = 0\}$. By definition of the weights, we can also define R_i using the leaves of the RF as $R_i = \cup_{\mathbf{X}_i \in G_i} \left[\cup_{l=1}^k A_n(\mathbf{X}_i, \Theta_l) \right]$. This shows that the R_i are connected space. Hence, we can apply the calibration steps separately on each group and use only the obser-

vations that are connected to the test point. By using the calibration by group, we reduce the computation of $S(k)$ in Lemma 4.3 needed for the computation of the PI from $\mathcal{O}(n \log(n))$ to $\mathcal{O}(|\mathbf{R}(\mathbf{X}_{n+1})| \log |\mathbf{R}(\mathbf{X}_{n+1})|)$, where $R(\mathbf{X}_{n+1})$ represents the region containing \mathbf{X}_{n+1} , and $|\mathbf{R}(\mathbf{X}_{n+1})|$ denotes the number of observations within $\mathbf{R}(\mathbf{X}_{n+1})$. Indeed, we only need to use the observations in the region where \mathbf{X}_{n+1} belongs in the calibration step. This results in a more accurate and efficient PI. In addition, no coverage guarantees are lost as the R_i forms a partition. We prove the marginal coverage of the group-wise LCP-RF in the Appendix (18).

In some cases, the graph may have a single connected component. Consequently, we propose to regroup calibration observations by (non-overlapping) communities using the weights of the RF. This involves grouping the nodes (calibration samples) of the graph into communities such that nodes within the same community are strongly connected to each other and weakly connected to nodes in other groups. Various methods exist for detecting communities in graphs, such as hierarchical clustering, spectral clustering, random walk, label propagation, and modularity maximization. A comprehensive overview of these methods can be found in [Schaeffer, 2007]. Nonetheless, it is challenging to determine the most suitable approach as the selection depends on the particular problem and characteristics of the graph. In our experiments, we found that the popular Louvain-Leiden [Traag, 2019] method coupled with Markov Stability [Delvenne, 2010] is effective in detecting communities of the learned weights of the Random Forest. However, any clustering method can be used depending on the specific application and dataset.

Let's assume a graph-clustering algorithm that returns L disjoint clusters $C(\mathcal{D}_n) = \{C_1, \dots, C_L\}$. Note that contrary to connected components, we can have $\mathbf{X} \in C_i$, $\mathbf{X}' \in C_j$ and $w_n(\mathbf{X}, \mathbf{X}') \neq 0$, therefore it's more difficult to define the associated regions R_1, \dots, R_L that form a partition of \mathcal{X} s.t. for any $\mathbf{X} \in C_i$, then $\mathbf{X} \in R_i$. We define R_i as the set of points \mathbf{x} that assigns the highest weights to the observations in cluster C_i . As w can be interpreted as a transition matrix, we define R_i as the set of \mathbf{X} such that $\mathbb{P}_G(\mathbf{X} \in C_i) > \mathbb{P}_G(\mathbf{X} \in C_k)$ for all $k \neq i$, where the probability is computed using the weights of the forest represented by the graph G . Formally, $\mathbb{P}_G(\mathbf{X} \in C_i) = \sum_{\mathbf{X}_j \in C_i} w_n(\mathbf{X}, \mathbf{X}_j)$ and R_i can be represented as

$$R_i = \left\{ \mathbf{x} \in \mathcal{X} : \sum_{\mathbf{X}_j \in C_i} w_n(\mathbf{x}, \mathbf{X}_j) > \sum_{\mathbf{X}_j \in C_k} w_n(\mathbf{x}, \mathbf{X}_j), k \neq i \right\}.$$

However, we also need to define another set for observations that are "undecidable", i.e., belong to several groups at the same time. We define this set as

$$\bar{R} = \left\{ \mathbf{x} \in \mathcal{X} : \exists k, l \in [L], \sum_{\mathbf{X}_j \in C_l} w_n(\mathbf{x}, \mathbf{X}_j) = \sum_{\mathbf{X}_j \in C_k} w_n(\mathbf{x}, \mathbf{X}_j) \right\}.$$

As the groups R_1, \dots, R_L and \bar{R} induced by the clusters form a partition of the input space, we get marginal/PAC coverage similar to the connected components case by applying the calibration step by group.

5 Asymptotic conditional coverage

Here, we study the conditional coverage of LCP-RF. It is widely recognized that obtaining meaningful distribution-free conditional coverage is impossible with a finite sample size [Lei, 2014b; Vovk, 2012]. Below, we demonstrate the asymptotic conditional coverage of LCP-RF while making weaker assumptions than the original LCP based on kernel [Guan, 2022].

Assumption 5.1. For all $r \in \mathbb{R}$, the c.d.f $\mathbf{x} \mapsto F(r|\mathbf{X} = \mathbf{x})$ is continuous.

Assumption 5.1 is necessary to get uniform convergence of the RF estimator.

Assumption 5.2. For any $l \in [k]$, the variation of the conditional cumulative distribution function within any cell goes to 0, i.e., $\forall \mathbf{x} \in \mathbb{R}^d, \forall r \in \mathbb{R}, \sup_{\mathbf{z} \in A_n(\mathbf{x}; \Theta_l)} |F(r|\mathbf{z}) - F(r|\mathbf{x})| \xrightarrow[n \rightarrow +\infty]{a.s.} 0$.

Assumption 5.2 allows to control the approximation error of the estimator. If for all y , $F(y|\cdot)$ is continuous, Assumption 4.2 is satisfied provided that the diameter of the cell goes to zero. Note that the vanishing of the diameter of the cell is a common condition used to prove the consistency of general partitioning estimator (see chapter 4 in [Györfi, 2002]). [Scornet, 2015] show that this is true when the data come from additive regression models [Stone, 1985b], and [Elie-Dit-Cosaque, 2022] show that it holds for a more general class, such as product functions or sums of product functions. The result is also valid for all regression functions, with a slightly modified version of RF, where each child node contains at least a small fraction of the observations in the parent node, and the probability that each variable $j = 1, \dots, p$ is chosen for splitting is positive for every node. Under these small modifications, Lemma 2 from [Meinshausen, 2006] gives that the diameter of each leaf node vanishes. Therefore, we do not need to assume that for all r , $F(r|\cdot)$ is Lipschitz, as required in LCP [Guan, 2022], which is a much stronger assumption.

Assumption 5.3. Let k and the number of bootstrap observations in a leaf node $N_n(A_n(\mathbf{x}; \Theta_l))$, s.t. there exists $k = \mathcal{O}(n^\alpha)$, with $\alpha > 0$, and $\forall \mathbf{x} \in \mathbb{R}^d, N_n(A_n(\mathbf{x}; \Theta_l)) = \Omega^1(\sqrt{n} \ln(n)^\beta)$, with $\beta > 1$ a.s.

Assumption 5.3 allows us to control the estimation error and means that the cells should contain a sufficiently large number of points so that averaging among the observations is effective. It can be enforced by adjusting the hyperparameters of the RF.

Under these assumptions, we prove that the selected $\tilde{\alpha}(v)$ when $\widehat{V}_{n+1} = v$ given by the LCP-RF converges to $1 - \alpha$, and the resulting PI achieves the target level $1 - \alpha$.

Theorem 5.4. *Suppose that all observations are i.i.d. drawn from the distribution P and let $\tilde{\alpha}(v)$ and $\widehat{C}_V(\mathbf{X}_{n+1})$ define as in Theorem 4.2. Under assumptions 5.1-5.3, we have for all $\epsilon > 0$ and any nonatomic points \mathbf{x}_{n+1} of $P_{\mathbf{X}}$,*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \widehat{V}_{n+1} \in \widehat{C}_V(\mathbf{X}_{n+1}) \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1} \right\} = 1 - \alpha \quad \text{and}$$

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \max_v |\tilde{\alpha}(v) - (1 - \alpha)| < \epsilon \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1} \right\} = 1.$$

¹ $f(n) = \Omega(g(n)) \iff \exists c > 0, \exists n_0 > 0 \mid \forall n \geq n_0, |f(n)| \geq c|g(n+1)|.$

The proof of Theorem 5.4 can be found in Appendix (19).

6 Experiments

We evaluate the performance of our proposed methods: LCP-RF (Random Forest Localizer with marginal and training-conditional calibration), LCP-RF-G (LCP-RF with groupwise calibration) and QRF-TC (Random Forest Localizer with only training-conditional calibration) against their competitors SPLIT (split-CP), SLCP and LCP. We used the original implementation of SLCP and LCP and tuned the kernel widths as described in their respective papers. We test the methods on simulated data with heterogeneous output and 3 real-world datasets from UCI [Dua, 2017a]: bike sharing demand (bike, $n = 10886, p = 12$), California house price (cali, $n = 20640, p = 8$), and community crime (commu, $n = 1993, p = 128$). The datasets are divided into three sets, namely the training set (40%), calibration set (40%), and the test set (20%). To ensure that the model’s error is not constant across all observations, we created a hole in the data by removing all observations from the training set whose output exceeds the 0.7-quantile of the training outputs. The PI is computed on the test sets at a level of $1 - \alpha = 0.9$.

We consider two nonconformity scores: mean score $\widehat{V}(\mathbf{X}, Y) = |Y - \widehat{\mu}(\mathbf{X})|$ where $\widehat{\mu}$ is mean estimate, and quantile score $\widehat{V}^Q(\mathbf{X}, Y) = \max\{\widehat{q}_{\alpha/2}(\mathbf{X}) - Y, Y - \widehat{q}_{1-\alpha/2}(\mathbf{X})\}$ where $\{\widehat{q}_{\alpha/2}, \widehat{q}_{1-\alpha/2}\}$ are quantile estimates at level $\alpha/2$ and $1 - \alpha/2$ respectively. We use XGBoost [Chen, 2016] of scikit-learn [Pedregosa, 2011] with default parameters as the mean estimate $\widehat{\mu}$ in our experiments. We leave the analysis of different models and the quantile score for the Appendix (20). We denote $C^m(\mathbf{X}_{n+1}) = [\widehat{\mu}(\mathbf{X}_{n+1}) \pm q^m(\mathbf{X}_{n+1})]$ the PI of each method m , and the oracle PI as $C^*(\mathbf{X}_{n+1}) = [\widehat{\mu}(\mathbf{X}_{n+1}) \pm q^*(\mathbf{X}_{n+1})]$ where $q^*(\mathbf{X}_{n+1}) = \mathcal{Q}(1 - \alpha; F_{\widehat{V}_{n+1}|\mathbf{X}_{n+1}})$.

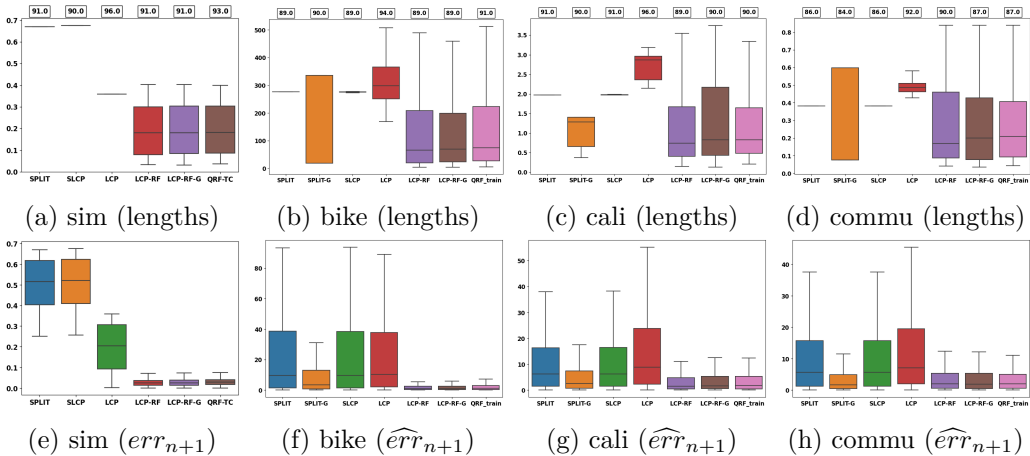


Figure 6.3: PI lengths and errors of the different methods with the mean score. The training-conditional coverages are at the top of the figure.

The simulated data (sim) is defined as: $\mathbf{X} \in [0, 1]^{50}$, $\mathbf{X}_i \sim \mathcal{U}([0, 1])$ for all $i \in [50]$ and $Y = \mathbf{X}_1 + \epsilon \times \mathbf{X}_1 / (1 + \mathbf{X}_1)$ where $\epsilon \sim \mathcal{N}(0, 1)$. In Figure 6.3e, we compute the absolute relative distance between the PI of each method and the oracle PI as $err_{n+1} = |q^m(\mathbf{X}_{n+1}) - q^*(\mathbf{X}_{n+1})| / q^*(\mathbf{X}_{n+1})$ showing that our methods are much closer to the oracle PI than its competitors. SLCP and

SPLIT are close, but they are less accurate than LCP. Figure 6.3a shows that most methods provide training-conditional coverage or empirical coverage over the test points at nearly 90%. Figure 6.3a also shows that our methods give varied intervals while the others have almost constant intervals.

The analysis of real-world data is more challenging because we don't have the oracle PI. To evaluate the effectiveness of the methods, we compare the length of the PI $q^m(\mathbf{X}_{n+1})$ to the true error of the model $\widehat{V}_{n+1} = \widehat{V}(\mathbf{X}_{n+1}, Y_{n+1})$. Indeed, a larger error of the model should result in a larger PI. Note that if $Y_{n+1}|\mathbf{X}_{n+1}$ does not vary too much then $\widehat{V}_{n+1} \approx q^*(\mathbf{X}_{n+1})$. We denote $\widehat{err}_{n+1} = |q^m(\mathbf{X}_{n+1}) - \widehat{V}_{n+1}|/\widehat{V}_{n+1}$ as the model's fidelity errors. We also introduce a new method (SPLIT-G), corresponding to groupwise split-CP using the groups defined by the RF's weights.

Figure 6.3 summarizes the results on the 3 real-world datasets. Starting with average coverage (top of the figure), most methods have empirical coverage at nearly exact nominal levels for all datasets. Our methods are slightly lower, which could be explained by the sample splitting used for the PAC interval calibration. Indeed, the bound in Theorem 4.4 depends on the size of the data and as we split the calibration set in two, we lose a bit in statistical efficiency.

The top figures (6.3b-6.3d) display the distribution of the lengths of the PI, while the four figures at the bottom (6.3f-6.3h) show the distribution of fidelity errors of the model \widehat{err}_{n+1} . Overall, our methods significantly outperform the others in terms of the model's uncertainty fidelity and adaptiveness of lengths. SLCP fails to provide any significant improvement over the standard split-CP. This could be due to the fact that it learns the localizer on the residuals of the training set, which may not represent the residuals of the calibration data, thereby leading to overfitting. While LCP-RF-G and QRF-TC are faster than LCP-RF, their performance are similar. In these datasets, we suspect that the RF localizer is so accurate that it is difficult to distinguish between the groupwise LCP-RF and the LCP-RF. However, we observe that by using the groups defined by the RF with the split-CP (SPLIT-G), we were able to improve the PI of split-CP. This demonstrates how our methods can improve the performance of any CP approach by just utilizing the groups established by the RF.

7 Conclusion

In this work, we have significantly enhanced the applicability of the Localized Conformal Prediction framework, which previously only worked on simple models with fewer than five variables, by adapting it for high-dimensional scenarios, accommodating categorical variables, and providing a PAC coverage guarantee. Our reweighting strategy based on the Random Forest algorithm can improve the PI computed using any nonconformity score. This results in more adaptive PI with marginal, training-conditional, and conditional coverage, making Conformal Predictive Intervals more similar to those produced by traditional statistics. This may ease their interpretation in terms of risks and give a clearer relationship between the length of the PI and the uncertainties of a given model $\widehat{\mu}$, thereby allowing for a better understanding of the limitations of a model $\widehat{\mu}$.

Chapter 7

Future works

Abstract

In this chapter, we present two ongoing works. The first explores the application of conformal prediction to enable models to abstain from making predictions in situations of high uncertainty. The second also leverages conformal prediction to generate plausible counterfactual explanations.

Contents

1	Prediction with reject option using conformal p-value	115
2	Conformal Protection Layers for Counterfactual Explanations	116

1 Prediction with reject option using conformal p-value

Machine learning model provides prediction, even when it is likely to be inaccurate. This behavior should be avoided in many decision support applications, where mistakes can have severe consequences. In this context, it may be beneficial to allow the model to abstain from predicting if the model is least confident.

Consider a training set $\mathcal{D}_m = \{(\mathbf{X}_i, Y_i)\}_{i=1}^m$ with $(\mathbf{X}_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ drawn exchangeably from $P = P_{\mathbf{X}}P_{Y|\mathbf{X}}$, and an algorithm \mathcal{A} that gives a predictive model $\hat{\mu}(\cdot) = \mathcal{A}(\mathcal{D}_m)$. We consider the problem of prediction with reject option where one is allowed to abstain from predicting if the error of the model is too important for a given observation. In the following discussion, we consider both regression and classification tasks. We define $\sigma(\mathbf{X}, Y)$ as the residual of the model, which is a measure of how closely $\hat{\mu}(\mathbf{X})$ aligns with the true output Y . Given a maximum tolerable residual score, denoted as σ^* , our objective is to approximate the oracle predictor with reject option $\Gamma_{\hat{\mu}}^*$ defined as:

$$\Gamma_{\hat{\mu}}^*(\mathbf{x}) = \begin{cases} \hat{\mu}(\mathbf{X}) & \text{if } \sigma(\mathbf{X}, Y) \leq \sigma^* \\ \emptyset & \text{otherwise.} \end{cases} \quad (7.1)$$

The primary obstacle is that we need the target variable Y to compute $\sigma(\mathbf{X}, Y)$. As a result, we lack the value of $\sigma(\mathbf{X}, Y)$ for any new observations \mathbf{X} or for the test set $\mathcal{D}^{test} = \{(\mathbf{X}_{m+j}, Y_{m+j})\}_{j=1}^n$. Our goal is to propose an approach to approximate $\Gamma_{\hat{\mu}}^*$ on the test set by utilizing an estimator $\hat{\sigma}$ of σ , while simultaneously ensuring finite-sample guarantees over the rejection set $\mathcal{R} = \{j \in \{1, \dots, n\} : \sigma(\mathbf{X}_{m+j}, Y_{m+j}) > \sigma^*\}$.

The starting point of our solution is the interpretation of approximating $\Gamma_{\hat{\mu}}^*$ as a multiple hypothesis testing problem. Indeed, the prediction with reject option on a given test set $\mathcal{D}^{test} = \{(\mathbf{X}_{m+j}, Y_{m+j})\}_{j=1}^n$ defined by Equation (7.1) is equivalent to testing for the following set of null hypotheses:

$$\mathcal{H}_{0,j} : \sigma(\mathbf{X}_{m+j}, Y_{m+j}) \leq \sigma^* \quad \text{for all } j \in \{1, \dots, n\}. \quad (7.2)$$

Our objective now is to determine a valid p-value for each hypothesis and establish a procedure that maintains control over the False Discovery Rate (FDR) while maximizing statistical power. Let $\mathcal{S} \subseteq \{1, \dots, n\}$ represent the set of rejected observations from a given procedure, the FDR is defined as the expected value of the False Discovery Proportion (FDP):

$$\text{FDR} = \mathbb{E}_{P^n}[\text{FDP}], \quad \text{FDP} = \frac{\sum_{j=1}^n \mathbb{1}\{j \in \mathcal{S}, \sigma(\mathbf{X}_{m+j}, Y_{m+j}) \leq \sigma^*\}}{1 \vee |\mathcal{S}|}.$$

The statistical power (Power), on the other hand, represents the number of valid rejections we

have correctly identified, and is defined as:

$$\text{Power} = \mathbb{E}_{P^n} \left[\frac{\sum_{j=1}^n \mathbb{1}\{j \in \mathcal{S}, \sigma(\mathbf{X}_{m+j}, Y_{m+j}) > \sigma^*\}}{\sum_{j=1}^n \mathbb{1}\{\sigma(\mathbf{X}_{m+j}, Y_{m+j}) > \sigma^*\}} \right].$$

Despite the lack of knowledge regarding the value of residual $\sigma(\mathbf{X}_{m+j}, Y_{m+j})$ on the test set, we can generate valid p-values for these null hypotheses without any assumptions about the data distribution or the model $\hat{\mu}$. This can be achieved by using an estimator $\hat{\sigma}$ of σ and leveraging the conformal inference framework. Our proposal is inspired by [Balasubramanian, 2014; Bates, 2021; Bates, 2023]. It involves many-to-one comparisons of the estimated residual $\hat{\sigma}$ of the individual test point against the estimated residuals of a control sample containing exclusively accepted observations, i.e., satisfying $\sigma(\mathbf{X}_i, Y_i) \leq \sigma^*$. Given a calibration data $\mathcal{D}_l^{\text{cal}} = \{(\mathbf{X}_{m+n+k}, Y_{m+n+k}) : \sigma(\mathbf{X}_{m+n+k}, Y_{m+n+k}) \leq \sigma^*, k = 1, \dots, l\}$, we opt to reject each test point $(\mathbf{X}_{m+j}, Y_{m+j}), j = 1, \dots, n$ if its estimated residual $\hat{\sigma}(\mathbf{X}_{m+j})$ significantly exceeds the estimated residuals of the calibration data $\hat{\sigma}(\mathbf{X}_{m+n+k})$ for $k = 1, \dots, l$. The rank of $\hat{\sigma}(\mathbf{X}_{m+j})$ among the estimated residuals of the calibration data $\hat{\sigma}(\mathbf{X}_{m+n+k})$ for $k = 1, \dots, l$ is used to generate a valid p-value for each hypothesis $\mathcal{H}_{0,j} : \sigma(\mathbf{X}_{m+j}, Y_{m+j}) \leq \sigma^*, j = 1, \dots, n$.

In addition, we are interested in ensuring conditional control of the False Discovery Proportion (FDP) instead of its marginal control, i.e., control of the FDR. The control we currently have using a method such as Benjamini-Hochberg (BH) [Benjamini, 1995] is $\text{FDR} = \mathbb{E}_{P^{l+n}}[\text{FDP}] \leq \alpha$, where the expectation is taken under the calibration and test sets. Our objective is to develop a procedure that allows for control conditionally on a specific calibration set. More formally, we want PAC-type guarantees, i.e., given calibration data $\mathcal{D}_l^{\text{cal}}$, we aim that with probability $1 - \delta, \delta \in (0, 1)$, our procedure will ensure $\text{FDP} \leq \alpha$.

2 Conformal Protection Layers for Counterfactual Explanations

The main challenge in generating counterfactual explanations is finding plausible modifications to the original observation. Several techniques attempt to assure this plausibility by integrating constraint based on outlier scores into the underlying optimization problem. For instance, Local Outlier Factor [Kanamori, 2020], Isolation Forest [Parmentier, 2021], and density-weighted metrics [Poyiadzi, 2019] have been employed to generate realistic samples. However, these techniques don't necessarily ensure that the resulting counterfactual examples are actually inliers. Therefore, we are presently exploring the use of conformal prediction as a means to statistically validate that the generated observations are not outliers. This approach mirrors the one we developed above; it involves comparing the outlier scores of the newly generated observations with the outlier scores of the original observations. If these scores significantly deviate from the outlier scores of the inlier observations, we reject the observation. Thanks to the exchangeability or approximate exchangeability [Tibshirani, 2019] property of the observations, we can effectively control the Type I error. This methodology was recently employed for outlier detection in [Bates, 2023; Jin, 2022; Marandon, 2022].

Conclusion

The first contribution of this thesis is providing a detailed and theoretical analysis demonstrating the limitations of the most used method for analyzing predictions of machine learning models, namely Shapley Values. In Chapter 2, we show that the TreeSHAP algorithm, used to compute Shapley Values for tree-based models, is biased when variables are dependent. We also show that the commonly used approach to compute Shapley Values in the presence of categorical variables is incorrect. As a consequence, we propose a better estimation of Shapley values and highlight how to compute these values in the presence of categorical variables. Chapter 3 extends this analysis by demonstrating that, in addition to these estimation problems, Local Shapley Values (SHAP) and the LIME method are not reliable in detecting local important variables.

Subsequently, our goal is to propose better alternatives for local explanations of models. Beyond capturing the local behavior of the model accurately, we aim for methods that satisfy certain desirable properties. In all the methods we propose, we strive to be model-free. This means that we do not need access to the prediction function $f(\cdot)$, which is particularly useful when the model is private or unavailable, or when the cost of a prediction is high. Thus, we can explain the observed values $\{(\mathbf{X}_i, f(\mathbf{X}_i))\}_{i=1}^n$ without making new predictions using $f(\cdot)$ or directly explaining the observed data outputs $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$. Additionally, our methods do not require generating new observations, avoiding the problem of extrapolation found in most approaches that evaluate the model on impossible observations that do not respect the data distribution, thereby compromising reliability. Our approaches solely utilize the observed data. Furthermore, the quantities sought by the proposed methods are clearly specified and are accompanied by consistency results.

In this thesis, we present different forms of explanations, such as importance measures, R-LOCO (Chapter 3), LXI (Chapter 4), selection of important local variables, Sufficient Explanations (Chapter 4), Decision Rules (Chapter 4), Counterfactual actions (Chapter 5), and Counterfactual Rules (Chapter 5). These diverse explanations address a variety of questions and allow for personalizing the discourse based on the intended recipient of the explanation. The various stakeholders in AI, such as the client, the auditor, the data scientist, and the business, have different needs, and certain explanations may resonate more with each of them. For example, for a client wanting to understand why his credit application was not accepted, counterfactual actions would be more relevant. For an auditor interested in determining whether the model

uses protected attributes to make decisions, the selection of local variables, decision rules or counterfactual rules provide a more explicit summary of the model’s behavior. As our approach does not require new predictions, and we can directly use historical data to generate explanations, a data scientist might utilize importance measures directly on residuals $Y_i - f(\mathbf{X}_i)$ to attempt model improvement.

In the second part of the thesis, we focus on constructing predictive intervals, which enable handling the uncertainty associated with predictions. This tool proves to be powerful for decision support or the automation of machine learning models. We specifically investigate conformal prediction, which provides non-asymptotic coverage guarantees with minimal assumptions, i.e., without assuming anything about the model and assuming data exchangeability. We identify a limitation in the construction of these predictive intervals, as they are inherently non-adaptive to the considered observation. Although it is possible to obtain adaptive intervals by adjusting the nonconformity score, the calculated correction to calibrate these intervals does not depend on the considered observation and is constant. Our solution is to learn the nonconformity score using a Random Forest, which introduces weighting on the calibration set and allows for an adaptive correction. While the often-highlighted guarantee in the conformal prediction literature is the marginal coverage rate, in practice, it is useful to ensure the coverage rate conditionally on a given calibration set. Thus, we propose a method to ensure that our approach satisfies this conditional coverage property with respect to the given calibration set. Finally, our approach is general and improves the predictive intervals returned by conformal prediction, regardless of the nonconformity score used.

In our future perspectives, we plan to merge the two aspects of this thesis. Specifically, we aim to use conformal prediction to provide non-asymptotic guarantees regarding the explanations we propose. Additionally, we intend to utilize explanation techniques to discern the variables that influence the predictive intervals. We have started exploration of these concepts in Chapter 7. There, we present an approach that employs conformal prediction as a filter to identify counterfactual examples that are implausible with statistical guarantees. Moreover, we propose a method for automating the handling of uncertainty associated with predictions, known as prediction with rejection. This approach allows the model to abstain from making a prediction if the uncertainty associated with that prediction exceeds a threshold defined by the user, utilizing conformal prediction to provide statistical guarantees.

Overall, this thesis sheds light on the limitations of existing methods, proposes novel and various techniques for local explanations with theoretical guarantees, and paves the way for future research to integrate explanation and uncertainty management in machine learning models.

Bibliographie

- [Aas, 2020] Kjersti Aas, Martin Jullum, and Anders Løland. *Explaining individual predictions when features are dependent: More accurate approximations to Shapley values*. 2020. arXiv: [1903.10464](https://arxiv.org/abs/1903.10464) [stat.ML] (cit. on pp. 23, 42, 43).
- [Aas, 2021] Kjersti Aas, Thomas Nagler, Martin Jullum, and Anders Løland. “Explaining predictive models using Shapley values and non-parametric vine copulas”. *arXiv preprint arXiv:2102.06416* (2021) (cit. on pp. 23, 42, 43).
- [Abo-Alsabeh, 2023] Rewayda Razaq Abo-Alsabeh, Hajem Ati Daham, and Abdellah Salhi. “On the maximum empty hyper-rectangle problem”. *Journal of Algorithms & Computational Technology* 17 (2023), p. 17483026221151197 (cit. on p. 77).
- [ACPR, 2022] Banque de France ACPR. *Techsprint sur l’explicabilité*. 2022 (cit. on p. 10).
- [Adadi, 2018] Amina Adadi and Mohammed Berrada. “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)”. *IEEE access* 6 (2018), pp. 52138–52160 (cit. on p. 5).
- [Agarwal, 2022] Abhineet Agarwal, Yan Shuo Tan, Omer Ronen, Chandan Singh, and Bin Yu. “Hierarchical Shrinkage: Improving the accuracy and interpretability of tree-based models.” *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 111–135 (cit. on p. 19).
- [Aggarwal, 1991] Alok Aggarwal, Hiroshi Imai, Naoki Katoh, and Subhash Suri. “Finding k points with minimum diameter and related problems”. *Journal of Algorithms* 12.1 (1991), pp. 38–56 (cit. on p. 77).
- [Aggarwal, 1987] Alok Aggarwal and Subhash Suri. “Fast algorithms for computing the largest empty rectangle”. *Proceedings of the third annual symposium on Computational geometry*. 1987, pp. 278–290 (cit. on p. 77).
- [Alvarez-Melis, 2018] David Alvarez-Melis and Tommi S Jaakkola. “On the robustness of interpretability methods”. *arXiv preprint arXiv:1806.08049* (2018) (cit. on pp. 15, 24, 43, 57).
- [Amoukou, 2021a] Salim I Amoukou and Nicolas JB Brunel. “Consistent Sufficient Explanations and Minimal Local Rules for explaining regression and classification models”. *arXiv preprint arXiv:2111.04658* (2021) (cit. on pp. 8, 87, 90–93).
- [Amoukou, 2022a] Salim I Amoukou and Nicolas JB Brunel. “Rethinking Counterfactual Explanations as Local and Regional Counterfactual Policies”. *arXiv preprint arXiv:2209.14568* (2022) (cit. on p. 8).
- [Amoukou, 2023] Salim I Amoukou and Nicolas JB Brunel. “Adaptive Conformal Prediction by Reweighting Nonconformity Score”. *arXiv preprint arXiv:2303.12695* (2023) (cit. on pp. 8, 35).
- [Amoukou, 2021b] Salim I Amoukou, Nicolas JB Brunel, and Tangi Salaün. “Accurate and robust Shapley Values for explaining predictions and focusing on local important variables”. *arXiv preprint arXiv:2106.03820* (2021) (cit. on pp. 23, 79).
- [Amoukou, 2021c] Salim I Amoukou, Nicolas JB Brunel, and Tangi Salaün. “The shapley value of coalition of variables provides better explanations”. *arXiv preprint arXiv:2103.13342* (2021) (cit. on p. 47).
- [Amoukou, 2022b] Salim I. Amoukou, Tangi Salaün, and Nicolas Brunel. “Accurate Shapley Values for explaining tree-based models”. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, 2022, pp. 2448–2465 (cit. on p. 8).

-
- [Angelopoulos, 2020] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. “Uncertainty sets for image classifiers using conformal prediction”. *arXiv preprint arXiv:2009.14193* (2020) (cit. on p. 33).
- [Angelopoulos, 2021a] Anastasios N Angelopoulos and Stephen Bates. “A gentle introduction to conformal prediction and distribution-free uncertainty quantification”. *arXiv preprint arXiv:2107.07511* (2021) (cit. on pp. 28, 33).
- [Angelopoulos, 2021b] Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. “Learn then test: Calibrating predictive algorithms to achieve risk control”. *arXiv preprint arXiv:2110.01052* (2021) (cit. on p. 35).
- [Apley, 2020] Daniel W Apley and Jingyu Zhu. “Visualizing the effects of predictor variables in black box supervised learning models”. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.4 (2020), pp. 1059–1086 (cit. on pp. 20, 21).
- [Arenal-Gutiérrez, 1996] Eusebio Arenal-Gutiérrez, Carlos Matrán, and Juan Antonio Cuesta-Albertos. “Unconditional Glivenko-Cantelli-type theorems and weak laws of large numbers for bootstrap”. *Statistics & Probability Letters* 26 (1996), pp. 365–375 (cit. on p. 152).
- [Athey, 2019] Susan Athey, Julie Tibshirani, and Stefan Wager. “Generalized random forests” (2019) (cit. on p. 64).
- [Bach, 2008] Francis R Bach. “Bolasso: model consistent lasso estimation through the bootstrap”. *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 33–40 (cit. on p. 16).
- [Backer, 2009] Jonathan Backer and J Mark Keil. “The Bichromatic Rectangle Problem in High Dimensions.” *CCCG*. 2009, pp. 157–160 (cit. on p. 77).
- [Balasubramanian, 2014] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014 (cit. on p. 116).
- [Barber, 2020] Rina Foygel Barber. “Is distribution-free inference possible for binary regression?” *Electronic Journal of Statistics* 14.2 (2020), pp. 3487–3524 (cit. on p. 27).
- [Barber, 2019a] Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. “Predictive inference with the jackknife+”. *arXiv preprint arXiv:1905.02928* (2019) (cit. on p. 33).
- [Barber, 2021] Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. “Predictive inference with the jackknife+”. *The Annals of Statistics* 49.1 (2021), pp. 486–507 (cit. on pp. 28, 34, 100, 102).
- [Barber, 2022] Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. “Conformal prediction beyond exchangeability”. *arXiv preprint arXiv:2202.13415* (2022) (cit. on pp. 28, 35, 106).
- [Barber, 2019b] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. “The limits of distribution-free conditional predictive inference”. *Information and Inference* 10 (2 2019), pp. 455–482 (cit. on pp. 34, 35, 102).
- [Barocas, 2020] Solon Barocas, Andrew D Selbst, and Manish Raghavan. “The hidden assumptions behind counterfactual explanations and principal reasons”. *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 80–89 (cit. on p. 26).
- [Basu, 2018] Sumanta Basu, Karl Kumbier, James B Brown, and Bin Yu. “Iterative random forests to discover predictive and stable high-order interactions”. *Proceedings of the National Academy of Sciences* 115.8 (2018), pp. 1943–1948 (cit. on pp. 76, 94).
- [Bates, 2021] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. “Distribution-free, risk-controlling prediction sets”. *Journal of the ACM (JACM)* 68.6 (2021), pp. 1–34 (cit. on pp. 35, 116).
- [Bates, 2023] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. “Testing for outliers with conformal p-values”. *The Annals of Statistics* 51.1 (2023), pp. 149–178 (cit. on pp. 13, 35, 116).
- [Beery, 2018] Sara Beery, Grant Van Horn, and Pietro Perona. “Recognition in terra incognita”. *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 456–473 (cit. on p. 4).
- [Bellucci, 2021] Matthieu Bellucci, Nicolas Delestre, Nicolas Malandain, and Cecilia Zanni-Merk. “Towards a terminology for a fully contextualized XAI”. *Procedia Computer Science* 192 (2021), pp. 241–250 (cit. on p. 5).
- [Bénard, 2021a] Clément Bénard. “Forêts aléatoires et interprétabilité des algorithmes d’apprentissage”. PhD thesis. Sorbonne université, 2021 (cit. on p. 15).
- [Bénard, 2021b] Clément Bénard, Gérard Biau, Sébastien Da Veiga, and Erwan Scornet. “SHAFF: Fast and consistent SHApely eFFect estimates via random Forests”. *arXiv preprint arXiv:2105.11724* (2021) (cit. on pp. 23, 47, 63, 72, 74, 76, 91, 92, 94).

-
- [Bénard, 2021c] Clément Bénard, Gérard Biau, Sébastien Da Veiga, and Erwan Scornet. “Sirus: Stable and interpretable rule set for classification”. *Electronic Journal of Statistics* 15.1 (2021), pp. 427–505 (cit. on pp. 19, 140).
- [Bénard, 2021d] Clément Bénard, Gérard Biau, Sébastien Veiga, and Erwan Scornet. “Interpretable random forests via rule extraction”. *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 937–945 (cit. on pp. 43, 76, 92).
- [Bénard, 2021e] Clément Bénard, Sébastien Da Veiga, and Erwan Scornet. “MDA for random forests: inconsistency, and a practical solution via the Sobol-MDA”. *arXiv preprint arXiv:2102.13347* (2021) (cit. on pp. 47, 72, 74, 76, 92).
- [Benjamini, 1995] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300 (cit. on p. 116).
- [Bian, 2022] Michael Bian and Rina Foygel Barber. “Training-conditional coverage for distribution-free predictive inference”. *arXiv preprint arXiv:2205.03647* (2022) (cit. on pp. 34, 102).
- [Biau, 2010] Gérard Biau and Luc Devroye. “On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification”. *Journal of Multivariate Analysis* 101.10 (2010), pp. 2499–2518 (cit. on pp. 73, 92, 105).
- [Black, 2020] Emily Black, Samuel Yeom, and Matt Fredrikson. “Fliptest: fairness testing via optimal transport”. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 111–121 (cit. on pp. 26, 89).
- [Bloniarz, 2016] Adam Bloniarz, Ameet Talwalkar, Bin Yu, and Christopher Wu. “Supervised Neighborhoods for Distributed Nonparametric Regression”. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Ed. by Arthur Gretton and Christian C. Robert. Vol. 51. Proceedings of Machine Learning Research. Cadiz, Spain: PMLR, 2016, pp. 1450–1459 (cit. on p. 19).
- [Bordt, 2023] Sebastian Bordt and Ulrike von Luxburg. “From Shapley values to generalized additive models and back”. *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 709–745 (cit. on p. 52).
- [Bousquet, 2002] Olivier Bousquet and André Elisseeff. “Stability and generalization”. *The Journal of Machine Learning Research* 2 (2002), pp. 499–526 (cit. on p. 15).
- [Breiman, 1996] L Breiman. “Bagging predictors”, *Machine Learning* 24, 123–140”. *Google Scholar Google Scholar Digital Library Digital Library* (1996) (cit. on p. 105).
- [Breiman, 1976] L. Breiman and William S. Meisel. “General Estimates of the Intrinsic Variability of Data in Nonlinear Regression Models”. *Journal of the American Statistical Association* 71 (1976), pp. 301–307 (cit. on p. 19).
- [Breiman, 2000] Leo Breiman. *Some infinity theory for predictor ensembles*. Tech. rep. Citeseer, 2000 (cit. on pp. 73, 92).
- [Breiman, 2001] Leo Breiman. “Random forests”. *Machine learning* 45.1 (2001), pp. 5–32 (cit. on pp. 6, 10, 11, 37, 52).
- [Breiman, 1984] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. “Classification and regression trees. Wadsworth Int”. *Group* 37.15 (1984), pp. 237–251 (cit. on pp. 17, 73, 91, 103, 105).
- [Buja, 1989] Andreas Buja, Trevor Hastie, and Robert Tibshirani. “Linear smoothers and additive models”. *The Annals of Statistics* (1989), pp. 453–510 (cit. on p. 17).
- [Candes, 2023] Emmanuel Candes, Lihua Lei, and Zhimei Ren. “Conformalized survival analysis”. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 85.1 (2023), pp. 24–45 (cit. on p. 35).
- [Carreira-Perpiñán, 2021] Miguel Á Carreira-Perpiñán and Suryabhan Singh Hada. “Counterfactual explanations for oblique decision trees: Exact, efficient algorithms”. *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 8. 2021, pp. 6903–6911 (cit. on p. 98).
- [CDC, 1999-2022] CDC. *National Health and Nutrition Examination Survey*. 1999-2022 (cit. on pp. 82, 163).
- [Chan, 2023] Timothy M Chan. “Faster algorithms for largest empty rectangles and boxes”. *Discrete & Computational Geometry* (2023), pp. 1–21 (cit. on p. 77).
- [Chan, 2021] Timothy M Chan and Sarel Har-Peled. “Smallest k-enclosing rectangle revisited”. *Discrete & Computational Geometry* 66.2 (2021), pp. 769–791 (cit. on p. 77).

-
- [Chastaing, 2012] Gaëlle Chastaing, Fabrice Gamboa, and Clémentine Prieur. “Generalized hoeffding-sobol decomposition for dependent variables-application to sensitivity analysis” (2012) (cit. on pp. 22, 53).
- [Chen, 2020] Hugh Chen, Joseph D Janizek, Scott Lundberg, and Su-In Lee. “True to the Model or True to the Data?” *arXiv preprint arXiv:2006.16234* (2020) (cit. on pp. 22, 37, 53).
- [Chen, 2018] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. “Learning to Explain: An Information-Theoretic Perspective on Model Interpretation”. *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 882–891 (cit. on p. 69).
- [Chen, 2022] Kuan-Lin Chen, Harinath Garudadri, and Bhaskar D Rao. “Improved Bounds on Neural Complexity for Representing Piecewise Linear Functions”. *arXiv preprint arXiv:2210.07236* (2022) (cit. on pp. 54, 55).
- [Chen, 2012] S. Chen, Arthur Choi, and Adnan Darwiche. “The Same-Decision Probability: A New Tool for Decision Making”. 2012 (cit. on pp. 70, 86).
- [Chen, 2013] Suming Chen, Arthur Choi, and Adnan Darwiche. “An Exact Algorithm for Computing the Same-Decision Probability”. *IJCAI ’13*. Beijing, China: AAAI Press, 2013, pp. 2525–2531 (cit. on p. 72).
- [Chen, 2016] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794 (cit. on pp. 10, 112).
- [Chernozhukov, 2010] Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. “Quantile and probability curves without crossing”. *Econometrica* 78.3 (2010), pp. 1093–1125 (cit. on p. 175).
- [Choi, 2020] YooJung Choi, Antonio Vergari, and Guy Van den Broeck. *Probabilistic circuits: A unifying framework for tractable probabilistic models*. Tech. rep. Technical report, 2020 (cit. on p. 72).
- [Chou, 2022] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. “Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications”. *Information Fusion* 81 (2022), pp. 59–83 (cit. on p. 85).
- [Chua, 1988] Leon O Chua and A-C Deng. “Canonical piecewise-linear representation”. *IEEE Transactions on Circuits and Systems* 35.1 (1988), pp. 101–111 (cit. on p. 54).
- [Cohen, 1995] William W Cohen. “Fast effective rule induction”. *Machine learning proceedings 1995*. Elsevier, 1995, pp. 115–123 (cit. on p. 17).
- [Cohen, 1999] William W Cohen and Yoram Singer. “A simple, fast, and effective rule learner”. *AAAI/IAAI* 99.335-342 (1999), p. 3 (cit. on p. 18).
- [Corbett-Davies, 2018] Sam Corbett-Davies and Sharad Goel. “The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning”. *ArXiv abs/1808.00023* (2018) (cit. on p. 4).
- [Covert, 2020a] Ian Covert and Su-In Lee. “Improving KernelSHAP: Practical Shapley Value Estimation via Linear Regression”. *CoRR abs/2012.01536* (2020). arXiv: [2012.01536](https://arxiv.org/abs/2012.01536) (cit. on p. 42).
- [Covert, 2020b] Ian Covert, Scott Lundberg, and Su-In Lee. “Explaining by Removing: A Unified Framework for Model Explanation”. *arXiv preprint arXiv:2011.14878* (2020) (cit. on p. 37).
- [Covert, 2020c] Ian Covert, Scott Lundberg, and Su-In Lee. “Understanding Global Feature Contributions Through Additive Importance Measures”. *CoRR abs/2004.00668* (2020). arXiv: [2004.00668](https://arxiv.org/abs/2004.00668) (cit. on pp. 23, 39, 52).
- [Covert, 2020d] Ian Covert, Scott M Lundberg, and Su-In Lee. “Understanding global feature contributions with additive importance measures”. *Advances in Neural Information Processing Systems* 33 (2020), pp. 17212–17223 (cit. on p. 22).
- [Covert, 2021] Ian C Covert, Scott Lundberg, and Su-In Lee. “Explaining by removing: A unified framework for model explanation”. *The Journal of Machine Learning Research* 22.1 (2021), pp. 9477–9566 (cit. on p. 6).
- [Da Veiga, 2021] Sébastien Da Veiga, Fabrice Gamboa, Bertrand Iooss, and Clémentine Prieur. *Basics and trends in sensitivity analysis: theory and practice in R*. SIAM, 2021 (cit. on p. 5).
- [Darwiche, 2020] Adnan Darwiche and Auguste Hirth. “On the reasons behind decisions”. *arXiv preprint arXiv:2002.09284* (2020) (cit. on pp. 68–70).
- [Dastin, 2018] Jeffrey Dastin. “Amazon scraps secret AI recruiting tool that showed bias against women”. *Ethics of data and analytics*. Auerbach Publications, 2018, pp. 296–299 (cit. on p. 1).

-
- [Datta, 1995] A. Datta, H.P. Lenhof, C. Schwarz, and M. Smid. “Static and Dynamic Algorithms for k-Point Clustering Problems”. *Journal of Algorithms* 19.3 (1995), pp. 474–503 (cit. on p. 77).
- [Datta, 2000] Amitava Datta and Subbiah Soundaralakshmi. “An efficient algorithm for computing the maximum empty rectangle in three dimensions”. *Information Sciences* 128.1-2 (2000), pp. 43–65 (cit. on p. 77).
- [Dawid, 1982] A Philip Dawid. “The well-calibrated Bayesian”. *Journal of the American Statistical Association* 77.379 (1982), pp. 605–610 (cit. on p. 7).
- [De Lara, 2021] Lucas De Lara, Alberto González-Sanz, Nicholas Asher, and Jean-Michel Loubes. “Transport-based counterfactual models”. *arXiv preprint arXiv:2108.13025* (2021) (cit. on pp. 26, 89).
- [DeGrave, 2021] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. “AI for radiographic COVID-19 detection selects shortcuts over signal”. *Nature Machine Intelligence* 3.7 (2021), pp. 610–619 (cit. on p. 4).
- [Delvenne, 2010] J-C Delvenne, Sophia N Yaliraki, and Mauricio Barahona. “Stability of graph communities across time scales”. *Proceedings of the national academy of sciences* 107.29 (2010), pp. 12755–12760 (cit. on p. 110).
- [Deng, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database”. *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on p. 33).
- [Dhurandhar, 2018] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, et al. “Explanations based on the missing: Towards contrastive explanations with pertinent negatives”. *Advances in neural information processing systems* 31 (2018) (cit. on p. 69).
- [Doshi-Velez, 2017] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. *arXiv preprint arXiv:1702.08608* (2017) (cit. on p. 5).
- [Du, 2021] Qiming Du, Gérard Biau, François Petit, and Raphaël Porcher. “Wasserstein Random Forests and Applications in Heterogeneous Treatment Effects”. *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1729–1737 (cit. on pp. 73, 92, 105).
- [Dua, 2017a] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017 (cit. on pp. 49, 57, 97, 112, 140, 160, 175).
- [Dua, 2017b] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017 (cit. on p. 140).
- [Dumitrescu, 2013] Adrian Dumitrescu and Minghui Jiang. “On the largest empty axis-parallel box amidst n points”. *Algorithmica* 66.2 (2013), pp. 225–248 (cit. on p. 77).
- [EBA, 2020] EBA. *Guidelines on Loan Origination and Monitoring*. 2020 (cit. on p. 3).
- [Eckstein, 2002] Jonathan Eckstein, Peter L Hammer, Ying Liu, Mikhail Nediak, and Bruno Simeone. “The maximum box problem and its application to data analysis”. *Computational Optimization and Applications* 23.3 (2002), pp. 285–298 (cit. on p. 77).
- [Efron, 1981] Bradley Efron and Charles Stein. “The jackknife estimate of variance”. *The Annals of Statistics* (1981), pp. 586–596 (cit. on p. 53).
- [Elie-Dit-Cosaque, 2022] Kevin Elie-Dit-Cosaque and Véronique Maume-Deschamps. “Random forest estimation of conditional distribution functions and conditional quantiles”. *Electronic Journal of Statistics* 16.2 (2022), pp. 6553–6583 (cit. on pp. 45, 74, 75, 111, 145, 146, 148, 168).
- [Eppstein, 1994] David Eppstein and Jeff Erickson. “Iterated nearest neighbors and finding minimal polytopes”. *Discrete & Computational Geometry* 11.1 (1994), pp. 321–350 (cit. on p. 77).
- [Feldman, 2005] Barry E Feldman. “Relative importance and value”. *Available at SSRN 2255827* (2005) (cit. on pp. 22, 54).
- [FICO, 2018] FICO. *FICO. Explainable machine learning challenge*. 2018 (cit. on p. 163).
- [Flaxman, 2016] Seth Flaxman, Dino Sejdinovic, John P Cunningham, and Sarah Filippi. “Bayesian learning of kernel embeddings”. *arXiv preprint arXiv:1603.02160* (2016) (cit. on p. 57).
- [Fraser, 2011] Donald AS Fraser. “Is Bayes posterior just quick and dirty confidence?” (2011) (cit. on p. 7).
- [Frey, 2007] Brendan J Frey and Delbert Dueck. “Clustering by passing messages between data points”. *science* 315.5814 (2007), pp. 972–976 (cit. on p. 61).
- [Friedberg, 2020] Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. “Local linear forests”. *Journal of Computational and Graphical Statistics* 30.2 (2020), pp. 503–517 (cit. on pp. 19, 64).
- [Friedman, 2001a] Jerome H Friedman. “Greedy function approximation: a gradient boosting machine”. *Annals of statistics* (2001), pp. 1189–1232 (cit. on p. 20).

-
- [Friedman, 2008] Jerome H Friedman and Bogdan E Popescu. “Predictive learning via rule ensembles”. *The annals of applied statistics* (2008), pp. 916–954 (cit. on p. 18).
- [Friedman, 2001b] Jerome H. Friedman. “Greedy function approximation: A gradient boosting machine.” *Ann. Statist.* 29.5 (2001), pp. 1189–1232 (cit. on p. 37).
- [Frye, 2020] Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. “Shapley explainability on the data manifold”. *arXiv preprint arXiv:2006.01272* (2020) (cit. on p. 37).
- [Fukumizu, 2009] Kenji Fukumizu, Arthur Gretton, Gert Lanckriet, Bernhard Schölkopf, and Bharath K Sriperumbudur. “Kernel choice and classifiability for RKHS embeddings of probability distributions”. *Advances in neural information processing systems* 22 (2009) (cit. on p. 57).
- [Fürnkranz, 2015] Johannes Fürnkranz and Tomáš Kliegr. “A brief overview of rule learning”. *Rule Technologies: Foundations, Tools, and Applications: 9th International Symposium, RuleML 2015, Berlin, Germany, August 2-5, 2015, Proceedings 9*. Springer. 2015, pp. 54–69 (cit. on p. 18).
- [Gammerman, 1998] A. Gammerman, V. Vovk, and V. Vapnik. “Learning by Transduction”. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. UAI’98. Madison, Wisconsin: Morgan Kaufmann Publishers Inc., 1998, pp. 148–155 (cit. on p. 28).
- [Gan, 2022] Luqin Gan, Lili Zheng, and Genevera I Allen. “Inference for Interpretable Machine Learning: Fast, Model-Agnostic Confidence Intervals for Feature Importance”. *arXiv preprint arXiv:2206.02088* (2022) (cit. on p. 52).
- [Garreau, 2017] Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. “Large sample analysis of the median heuristic”. *arXiv preprint arXiv:1707.07269* (2017) (cit. on p. 57).
- [Garreau, 2020] Damien Garreau and Ulrike Luxburg. “Explaining the explainer: A first theoretical analysis of LIME”. *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1287–1296 (cit. on pp. 23, 52).
- [Geurts, 2006] Pierre Geurts, Damien Ernst, and Louis Wehenkel. “Extremely randomized trees”. *Machine learning* 63 (2006), pp. 3–42 (cit. on pp. 73, 92).
- [Ghalebikesabi, 2021] Sahra Ghalebikesabi, Lucile Ter-Minassian, Karla Diaz-Ordaz, and Chris Holmes. “On Locality of Local Explanation Models”. *arXiv preprint arXiv:2106.14648* (2021) (cit. on pp. 68, 79).
- [Gibbs, 2021] Isaac Gibbs and Emmanuel Candes. “Adaptive conformal inference under distribution shift”. *Advances in Neural Information Processing Systems* 34 (2021), pp. 1660–1672 (cit. on pp. 28, 35).
- [Gibbs, 2023] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. “Conformal Prediction With Conditional Guarantees”. *arXiv preprint arXiv:2305.12616* (2023) (cit. on p. 35).
- [Goehry, 2020] Benjamin Goehry. “Random forests for time-dependent processes”. *ESAIM: Probability and Statistics* 24 (2020), pp. 801–826 (cit. on pp. 74, 168).
- [Goldstein, 2015] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation”. *Journal of Computational and Graphical Statistics* 24.1 (2015), pp. 44–65 (cit. on p. 37).
- [Goodfellow, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. *arXiv preprint arXiv:1412.6572* (2014) (cit. on p. 24).
- [Goodman, 2017] Bryce Goodman and Seth Flaxman. “European Union regulations on algorithmic decision-making and a “right to explanation””. *AI magazine* 38.3 (2017), pp. 50–57 (cit. on p. 3).
- [Gosiewska, 2019] Alicja Gosiewska and P. Biecek. “Do Not Trust Additive Explanations”. *arXiv: Learning* (2019) (cit. on p. 68).
- [Gou, 2021] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. “Knowledge distillation: A survey”. *International Journal of Computer Vision* 129 (2021), pp. 1789–1819 (cit. on p. 15).
- [Grathwohl, 2020] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. “Your classifier is secretly an energy based model and you should treat it like one”. *International Conference on Learning Representations*. 2020 (cit. on p. 95).
- [Grinsztajn, 2022] Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. “Why do tree-based models still outperform deep learning on typical tabular data?” *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2022 (cit. on pp. 10, 92, 98, 103).
- [Grömping, 2007] Ulrike Grömping. “Estimators of relative importance in linear regression based on variance decomposition”. *The American Statistician* 61.2 (2007), pp. 139–147 (cit. on p. 54).

-
- [Grömping, 2020] Ulrike Grömping. “Model-agnostic effects plots for interpreting machine learning models”. *Reports in Mathematics, Physics and Chemistry, Department II, Beuth University of Applied Sciences Berlin Report 1* (2020), p. 2020 (cit. on p. 21).
- [Grunewalder, 2018] Steffen Grunewalder. “Plug-in estimators for conditional expectations and probabilities”. *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 1513–1521 (cit. on p. 45).
- [Guan, 2022] Leying Guan. “Localized conformal prediction: a generalized inference framework for conformal prediction”. *Biometrika* (2022). eprint: <https://academic.oup.com/biomet/advance-article-pdf/doi/10.1093/biomet/asac040/45911782/asac040.pdf> (cit. on pp. 35, 102, 106–108, 111, 166, 167, 171).
- [Gui, 2023] Yu Gui, Rina Foygel Barber, and Cong Ma. “Conformalized matrix completion”. *arXiv preprint arXiv:2305.10637* (2023) (cit. on p. 35).
- [Guilmeau, 2021] Thomas Guilmeau, Emilie Chouzenoux, and Víctor Elvira. “Simulated Annealing: a Review and a New Scheme”. 2021, pp. 101–105 (cit. on p. 95).
- [Györfi, 2002] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Vol. 1. Springer, 2002 (cit. on pp. 45, 75, 111, 145).
- [Hama, 2022] Naofumi Hama, Masayoshi Mase, and Art B Owen. “Model free Shapley values for high dimensional data”. *arXiv preprint arXiv:2211.08414* (2022) (cit. on p. 58).
- [Han, 2022] Xing Han, Ziyang Tang, Joydeep Ghosh, and Qiang Liu. “Split Localized Conformal Prediction”. *arXiv preprint arXiv:2206.13092* (2022) (cit. on p. 102).
- [Hanin, 2019a] Boris Hanin and David Rolnick. “Complexity of linear regions in deep networks”. *International Conference on Machine Learning*. PMLR. 2019, pp. 2596–2604 (cit. on p. 55).
- [Hanin, 2019b] Boris Hanin and David Rolnick. “Deep relu networks have surprisingly few activation patterns”. *Advances in neural information processing systems* 32 (2019) (cit. on p. 55).
- [Hastie, 1987] Trevor Hastie and Robert Tibshirani. “Generalized additive models: some applications”. *Journal of the American Statistical Association* 82.398 (1987), pp. 371–386 (cit. on p. 16).
- [Hastie, 2015] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015 (cit. on p. 16).
- [He, 1997] Xuming He. “Quantile curves without crossing”. *The American Statistician* 51.2 (1997), pp. 186–192 (cit. on p. 175).
- [Hebiri, 2012] Mohamed Hebiri and Johannes Lederer. “How correlations influence lasso prediction”. *IEEE Transactions on Information Theory* 59.3 (2012), pp. 1846–1854 (cit. on p. 16).
- [Herin, 2022] Margot Herin, Marouane El Idrissi, Vincent Chabridon, and Bertrand Iooss. “Proportional marginal effects for global sensitivity analysis”. *arXiv preprint arXiv:2210.13065* (2022) (cit. on p. 22).
- [Heskes, 2020] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. “Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models”. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. 2020 (cit. on pp. 22, 37, 53).
- [Hinton, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network”. *arXiv preprint arXiv:1503.02531* (2015) (cit. on p. 15).
- [Hoeffding, 1948] Wassily Hoeffding. “A Class of Statistics with Asymptotically Normal Distribution”. *Annals of Mathematical Statistics* 19 (1948), pp. 308–334 (cit. on pp. 22, 53).
- [Hooker, 2007] Giles Hooker. “Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables”. *Journal of Computational and Graphical Statistics* 16 (2007), pp. 709–732 (cit. on p. 53).
- [Hooker, 2019] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. “A benchmark for interpretability methods in deep neural networks”. *Advances in neural information processing systems* 32 (2019) (cit. on p. 58).
- [Hu, 2020] Xiaoyu Hu and Jing Lei. “A distribution-free test of covariate shift using conformal prediction”. *arXiv preprint arXiv:2010.07147* (2020) (cit. on p. 35).
- [Hu, 2019] Xiyang Hu, Cynthia Rudin, and Margo Seltzer. “Optimal sparse decision trees”. *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 18).
- [Ignatiev, 2019] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. “On validating, repairing and refining heuristic ML explanations”. *arXiv preprint arXiv:1907.02509* (2019) (cit. on p. 69).

-
- [Iooss, 2015] Bertrand Iooss and Paul Lemaitre. “A review on global sensitivity analysis methods”. *Uncertainty management in simulation-optimization of complex systems: algorithms and applications* (2015), pp. 101–122 (cit. on p. 52).
- [Ishwaran, 2008] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. “Random survival forests”. *The annals of applied statistics* 2.3 (2008), pp. 841–860 (cit. on pp. 73, 92, 105).
- [Izbicki, 2020] Rafael Izbicki, Gilson Shimizu, and Rafael Stern. “Flexible distribution-free conditional predictive bands using density estimators”. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3068–3077 (cit. on p. 104).
- [Janzing, 2020] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. “Feature relevance quantification in explainable AI: A causal problem”. *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2907–2916 (cit. on pp. 22, 37, 53).
- [Jethani, 2021] Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. “Have We Learned to Explain?: How Interpretability Methods Can Learn to Encode Predictions in their Interpretations.” *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1459–1467 (cit. on p. 69).
- [Jin, 2022] Ying Jin and Emmanuel J Candès. “Selection by Prediction with Conformal p-values”. *arXiv preprint arXiv:2210.01408* (2022) (cit. on pp. 13, 35, 116).
- [Jocteur, 2023] Bérénice-Alexia Jocteur, Véronique Maume-Deschamps, and Pierre Ribereau. “Heterogeneous Treatment Effect based Random Forest: HTERF” (2023) (cit. on pp. 73, 92, 105).
- [Johnson, 2004] Jeff W Johnson and James M LeBreton. “History and use of relative importance indices in organizational research”. *Organizational research methods* 7.3 (2004), pp. 238–257 (cit. on p. 54).
- [Joshi, 2019] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. “Towards realistic individual recourse and actionable explanations in black-box decision making systems”. *arXiv preprint arXiv:1907.09615* (2019) (cit. on p. 25).
- [Kaggle, 2015] Kaggle. *Bike Sharing Demand*. 2015 (cit. on pp. 20, 81).
- [Kaggle, 2016] Kaggle. *Pima Indians Diabetes Database*. 2016 (cit. on pp. 97, 160).
- [Kaggle, 2017] Kaggle. *IBM HR Analytics Employee Attrition Performance*. 2017 (cit. on p. 82).
- [Kanamori, 2020] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. “DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization”. *IJCAI*. 2020 (cit. on pp. 25, 86, 116).
- [Kanamori, 2022] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Yuichi Ike. “Counterfactual Explanation Trees: Transparent and Consistent Actionable Recourse with Decision Trees”. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, PMLR 151:1846-1870. 2022 (cit. on pp. 88, 96).
- [Kaplan, 2019] Haim Kaplan, Sasanka Roy, and Micha Sharir. “Finding axis-parallel rectangles of fixed perimeter or area containing the largest number of points”. *Computational Geometry* 81 (2019), pp. 1–11 (cit. on p. 77).
- [Karimi, 2020a] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. “Model-Agnostic Counterfactual Explanations for Consequential Decisions”. *ArXiv abs/1905.11190* (2020) (cit. on pp. 85, 87).
- [Karimi, 2020b] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. “A survey of algorithmic recourse: definitions, formulations, solutions, and prospects”. *CoRR abs/2010.04050* (2020). arXiv: [2010.04050](https://arxiv.org/abs/2010.04050) (cit. on pp. 25, 85).
- [Karimi, 2021] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. “Algorithmic recourse: from counterfactual explanations to interventions”. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 353–362 (cit. on pp. 25, 26).
- [Kelley Pace, 1997] R. Kelley Pace and Ronald Barry. “Sparse spatial autoregressions”. *Statistics, Probability Letters* 33.3 (1997), pp. 291–297 (cit. on pp. 64, 97, 101).
- [Kivaranovic, 2020] Danijel Kivaranovic, Kory D Johnson, and Hannes Leeb. “Adaptive, distribution-free prediction intervals for deep networks”. *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 4346–4356 (cit. on p. 108).
- [Klusowski, 2020] Jason Klusowski. “Sparse learning with CART”. *Advances in Neural Information Processing Systems* 33 (2020), pp. 11612–11622 (cit. on p. 76).

-
- [Klusowski, 2021] Jason M Klusowski. “Universal consistency of decision trees in high dimensions”. *arXiv preprint arXiv:2104.13881* (2021) (cit. on p. 103).
- [Koenker, 2001] Roger Koenker and Kevin F Hallock. “Quantile regression”. *Journal of economic perspectives* 15.4 (2001), pp. 143–156 (cit. on p. 7).
- [König, 2021] Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. “A causal perspective on meaningful and robust algorithmic recourse”. *arXiv preprint arXiv:2107.07853* (2021) (cit. on p. 26).
- [König, 2023] Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. “Improvement-focused causal recourse (ICR)”. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 10. 2023, pp. 11847–11855 (cit. on p. 26).
- [Kovács, 2022] László Kovács. “Feature selection algorithms in generalized additive models under concurrency”. *Computational Statistics* (2022), pp. 1–33 (cit. on p. 17).
- [Kuchibhotla, 2020] Arun Kumar Kuchibhotla. “Exchangeability, conformal prediction, and rank tests”. *arXiv preprint arXiv:2005.06095* (2020) (cit. on p. 28).
- [Künzel, 2022] Sören R Künzel, Theo F Saarinen, Edward W Liu, and Jasjeet S Sekhon. “Linear aggregation in tree-based estimators”. *Journal of Computational and Graphical Statistics* 31.3 (2022), pp. 917–934 (cit. on pp. 19, 64).
- [Kusner, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. “Counterfactual fairness”. *Advances in neural information processing systems* 30 (2017) (cit. on p. 25).
- [Lakkaraju, 2022] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. “Rethinking Explainability as a Dialogue: A Practitioner’s Perspective”. *CoRR* abs/2202.01875 (2022). arXiv: [2202.01875](https://arxiv.org/abs/2202.01875) (cit. on pp. 25, 85, 87).
- [Laurent, 1976] Hyafil Laurent and Ronald L Rivest. “Constructing optimal binary decision trees is NP-complete”. *Information processing letters* 5.1 (1976), pp. 15–17 (cit. on p. 18).
- [Lecun, 2006] Yann Lecun, Sumit Chopra, and Raia Hadsell. “A tutorial on energy-based learning”. 2006 (cit. on p. 95).
- [Lee, 2021] Yonghoon Lee and Rina Barber. “Distribution-free inference for regression: discrete, continuous, and in between”. *Advances in Neural Information Processing Systems* 34 (2021), pp. 7448–7459 (cit. on p. 27).
- [Lei, 2016] Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry A. Wasserman. “Distribution-Free Predictive Inference for Regression”. *Journal of the American Statistical Association* 113 (2016), pp. 1094–1111 (cit. on pp. 7, 28, 30–32, 52, 58, 100, 104).
- [Lei, 2011] Jing Lei, James Robins, and Larry Wasserman. “Efficient nonparametric conformal prediction regions”. *arXiv preprint arXiv:1111.1418* (2011) (cit. on p. 28).
- [Lei, 2013] Jing Lei, James Robins, and Larry Wasserman. “Distribution-free prediction sets”. *Journal of the American Statistical Association* 108.501 (2013), pp. 278–287 (cit. on p. 28).
- [Lei, 2014a] Jing Lei and Larry Wasserman. “Distribution-free prediction bands for non-parametric regression”. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* (2014), pp. 71–96 (cit. on p. 35).
- [Lei, 2014b] Jing Lei and Larry A. Wasserman. “Distribution-free prediction bands for non-parametric regression”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (2014) (cit. on pp. 34, 35, 102, 111).
- [Lewis, 1973] David Lewis. “Causation”. *The journal of philosophy* 70.17 (1973), pp. 556–567 (cit. on p. 25).
- [Ley, 2022] Dan Ley, Saumitra Mishra, and Daniele Magazzeni. *Global Counterfactual Explanations: Investigations, Implementations and Improvements*. 2022 (cit. on pp. 88, 97).
- [Lim, 2016] Chinghay Lim and Bin Yu. “Estimation stability with cross-validation (ESCV)”. *Journal of Computational and Graphical Statistics* 25.2 (2016), pp. 464–492 (cit. on p. 16).
- [Lin, 2020] Jimmy Lin, Chudi Zhong, Diane Hu, Cynthia Rudin, and Margo Seltzer. “Generalized and scalable optimal sparse decision trees”. *International Conference on Machine Learning*. PMLR. 2020, pp. 6150–6160 (cit. on p. 18).
- [Lin, 2006] Yi Lin and Yongho Jeon. “Random forests and adaptive nearest neighbors”. *Journal of the American Statistical Association* 101.474 (2006), pp. 578–590 (cit. on pp. 73, 92, 103, 105).
- [Lin, 2018] Yin-Ting Lin and Jing-Sin Liu. “Revisit of minimum-area enclosing rectangle of a convex polygon”. *2018 5th international conference on control, decision and information technologies (CoDIT)*. IEEE. 2018, pp. 1051–1056 (cit. on p. 77).

-
- [Lin, 2021] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. “Locally valid and discriminative prediction intervals for deep learning models”. *Advances in Neural Information Processing Systems* 34 (2021), pp. 8378–8391 (cit. on p. 102).
- [Liu, 2008] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation forest”. *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422 (cit. on p. 95).
- [Loh, 2011] Wei-Yin Loh. “Classification and regression trees”. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (2011) (cit. on p. 91).
- [Looveren, 2019] Arnaud Van Looveren and Janis Klaise. “Interpretable Counterfactual Explanations Guided by Prototypes”. *CoRR* abs/1907.02584 (2019). arXiv: [1907.02584](https://arxiv.org/abs/1907.02584) (cit. on pp. 85, 86).
- [Lopardo, 2023] Gianluigi Lopardo, Frederic Precioso, and Damien Garreau. “A Sea of Words: An In-Depth Analysis of Anchors for Text Data”. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. PMLR, 2023, pp. 4848–4879 (cit. on p. 24).
- [Lou, 2013] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. “Accurate intelligible models with pairwise interactions”. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013, pp. 623–631 (cit. on p. 16).
- [Lundberg, 2020a] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, et al. “From local explanations to global understanding with explainable AI for trees”. *Nature machine intelligence* 2.1 (2020), pp. 56–67 (cit. on pp. 53, 69).
- [Lundberg, 2018] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. “Consistent individualized feature attribution for tree ensembles”. *arXiv preprint arXiv:1802.03888* (2018) (cit. on pp. 43, 138).
- [Lundberg, 2017a] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. 2017, pp. 4765–4774 (cit. on pp. 8, 9, 21, 22, 37, 42, 53, 68).
- [Lundberg, 2017b] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. *Advances in neural information processing systems* 30 (2017) (cit. on pp. 52, 68).
- [Lundberg, 2020b] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, et al. “From local explanations to global understanding with explainable AI for trees”. *Nature Machine Intelligence* 2.1 (2020), pp. 2522–5839 (cit. on pp. 10, 23, 37, 38, 42, 43, 47, 85, 138).
- [Mahajan, 2019] Divyat Mahajan, Chenhao Tan, and Amit Sharma. “Preserving causal constraints in counterfactual explanations for machine learning classifiers”. *arXiv preprint arXiv:1912.03277* (2019) (cit. on p. 25).
- [Malle, 2006] Bertram F Malle. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press, 2006 (cit. on p. 5).
- [Marandon, 2022] Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. “Machine learning meets false discovery rate”. *arXiv preprint arXiv:2208.06685* (2022) (cit. on p. 116).
- [Margot, 2021] Vincent Margot, Jean-Patrick Baudry, Frederic Guilloux, and Olivier Wintenberger. “Consistent regression using data-dependent coverings”. *Electronic Journal of Statistics* 15.1 (2021), pp. 1743–1782 (cit. on p. 45).
- [Massart, 1990] Pascal Massart. “The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality”. *The annals of Probability* (1990), pp. 1269–1283 (cit. on p. 173).
- [Meinshausen, 2010a] Nicolai Meinshausen. “Node harvest”. *The Annals of Applied Statistics* (2010), pp. 2049–2072 (cit. on p. 18).
- [Meinshausen, 2010b] Nicolai Meinshausen and Peter Bühlmann. “Stability selection”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010), pp. 417–473 (cit. on p. 16).
- [Meinshausen, 2006] Nicolai Meinshausen and Greg Ridgeway. “Quantile regression forests.” *Journal of Machine Learning Research* 7.6 (2006) (cit. on pp. 45, 72, 74, 75, 91, 92, 103, 105, 111, 145).
- [Miller, 2019] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. *Artificial intelligence* 267 (2019), pp. 1–38 (cit. on pp. 5, 6, 25).
- [Molnar, 2022] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. 2022 (cit. on pp. 15, 85).
- [Mothilal, 2020] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. “Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations”. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 607–617 (cit. on pp. 85–87).

-
- [Murdoch, 2019] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. “Interpretable machine learning: definitions, methods, and applications”. *arXiv preprint arXiv:1901.04592* (2019) (cit. on p. 15).
- [Naamad, 1984] Amnon Naamad, DT Lee, and W-L Hsu. “On the maximum empty rectangle problem”. *Discrete Applied Mathematics* 8.3 (1984), pp. 267–277 (cit. on p. 77).
- [Nadaraya, 1964] Elizbar A Nadaraya. “On estimating regression”. *Theory of Probability & Its Applications* 9.1 (1964), pp. 141–142 (cit. on pp. 43, 102).
- [Nori, 2019] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. “InterpretML: A Unified Framework for Machine Learning Interpretability”. *arXiv preprint arXiv:1909.09223* (2019) (cit. on pp. 9, 16).
- [OpenAI, 2023] OpenAI. “GPT-4 Technical Report”. *ArXiv abs/2303.08774* (2023) (cit. on p. 4).
- [Osborne, 1994] Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994 (cit. on p. 21).
- [Ovchinnikov, 2000] Sergei Ovchinnikov. “Max-min representation of piecewise linear functions”. *arXiv preprint math/0009026* (2000) (cit. on p. 54).
- [Owen, 2014] Art B Owen. “Sobol’ indices and Shapley value”. *SIAM/ASA Journal on Uncertainty Quantification* 2.1 (2014), pp. 245–251 (cit. on pp. 22, 53).
- [Owen, 2017] Art B Owen and Clémentine Prieur. “On Shapley value for measuring importance of dependent inputs”. *SIAM/ASA Journal on Uncertainty Quantification* 5.1 (2017), pp. 986–1002 (cit. on pp. 22, 39, 53).
- [Papadopoulos, 2008] Harris Papadopoulos, Alex Gammerman, and Volodya Vovk. “Normalized nonconformity measures for regression conformal prediction”. *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*. 2008, pp. 64–69 (cit. on pp. 31, 32).
- [Papadopoulos, 2002] Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alexander Gammerman. “Inductive Confidence Machines for Regression”. *European Conference on Machine Learning*. 2002 (cit. on pp. 28, 100).
- [Parmentier, 2021] Axel Parmentier and Thibaut Vidal. “Optimal Counterfactual Explanations in Tree Ensembles”. *CoRR abs/2106.06631* (2021). arXiv: [2106.06631](https://arxiv.org/abs/2106.06631) (cit. on pp. 25, 85, 86, 116).
- [Pascanu, 2013] Razvan Pascanu, Guido Montufar, and Yoshua Bengio. “On the number of response regions of deep feed forward networks with piece-wise linear activations”. *arXiv preprint arXiv:1312.6098* (2013) (cit. on p. 55).
- [Patki, 2016] N. Patki, R. Wedge, and K. Veeramachaneni. “The Synthetic Data Vault”. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2016, pp. 399–410 (cit. on p. 95).
- [Pawelczyk, 2022] Martin Pawelczyk, Teresa Datta, Johannes van-den-Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. *Algorithmic Recourse in the Face of Noisy Human Responses*. 2022 (cit. on pp. 26, 84, 85, 87, 98).
- [Pearl, 1994] Judea Pearl. “A probabilistic calculus of actions”. *Uncertainty Proceedings 1994*. Elsevier, 1994, pp. 454–462 (cit. on p. 26).
- [Pearl, 2000] Judea Pearl et al. “Models, reasoning and inference”. *Cambridge, UK: CambridgeUniversityPress* 19.2 (2000), p. 3 (cit. on p. 26).
- [Pearl, 2009] Judea Pearl. *Causality*. Cambridge university press, 2009 (cit. on p. 25).
- [Pedregosa, 2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. “Scikit-learn: Machine Learning in Python”. *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 61, 112, 161).
- [Pessach, 2022] Dana Pessach and Erez Shmueli. “A Review on Fairness in Machine Learning”. *ACM Computing Surveys (CSUR)* 55 (2022), pp. 1–44 (cit. on p. 4).
- [Plassier, 2023] Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. *Conformal Prediction for Federated Uncertainty Quantification Under Label Shift*. 2023. arXiv: [2306.05131](https://arxiv.org/abs/2306.05131) [stat.ML] (cit. on p. 35).
- [Póczos, 2013] Barnabás Póczos, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman. “Distribution-free distribution regression”. *artificial intelligence and statistics*. PMLR. 2013, pp. 507–515 (cit. on p. 28).
- [Podkopaev, 2021] Aleksandr Podkopaev and Aaditya Ramdas. “Distribution-free uncertainty quantification for classification under label shift”. *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 844–853 (cit. on p. 35).
- [Poyiadzi, 2019] Rafael Poyiadzi, Kacper Sokol, Raúl Santos-Rodríguez, Tijl De Bie, and Peter A. Flach. “FACE: Feasible and Actionable Counterfactual Explanations”. *CoRR abs/1909.09369* (2019). arXiv: [1909.09369](https://arxiv.org/abs/1909.09369) (cit. on pp. 25, 85, 86, 116).

-
- [Quinlan, 2014] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014 (cit. on p. 17).
- [Quinlan, 1986] J. Ross Quinlan. “Induction of decision trees”. *Machine learning* 1 (1986), pp. 81–106 (cit. on p. 17).
- [Ramsay, 2003] Timothy O Ramsay, Richard T Burnett, and Daniel Krewski. “The effect of concavity in generalized additive models linking mortality to ambient particulate matter”. *Epidemiology* 14.1 (2003), pp. 18–23 (cit. on p. 17).
- [Rawal, 2020] Kaivalya Rawal and Himabindu Lakkaraju. “Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses”. *Advances in Neural Information Processing Systems* 33 (2020), pp. 12187–12198 (cit. on pp. 88, 94, 96).
- [Razavi, 2021] Saman Razavi, Anthony Jakeman, Andrea Saltelli, Clémentine Prieur, Bertrand Iooss, Emanuele Borgonovo, et al. “The future of sensitivity analysis: An essential discipline for systems modeling and policy support”. *Environmental Modelling & Software* 137 (2021), p. 104954 (cit. on p. 5).
- [Ribeiro, 2016a] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ““Why should i trust you?” Explaining the predictions of any classifier”. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144 (cit. on pp. 8, 10, 23, 37, 52, 56, 68, 69, 85).
- [Ribeiro, 2016b] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ““Why Should I Trust You?": Explaining the Predictions of Any Classifier”. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 1135–1144 (cit. on p. 9).
- [Ribeiro, 2018] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Anchors: High-precision model-agnostic explanations”. *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018 (cit. on pp. 24, 68, 69, 71).
- [Rinaldo, 2019] Alessandro Rinaldo, Larry Wasserman, and Max G’Sell. “Bootstrapping and sample splitting for high-dimensional, assumption-lean inference” (2019) (cit. on p. 52).
- [Rivest, 1987] Ronald L Rivest. “Learning decision lists”. *Machine learning* 2 (1987), pp. 229–246 (cit. on p. 18).
- [Romano, 2020a] Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. “With malice toward none: Assessing uncertainty via equalized coverage”. *Harvard Data Science Review* 2.2 (2020), p. 4 (cit. on p. 34).
- [Romano, 2019] Yaniv Romano, Evan Patterson, and Emmanuel Candes. “Conformalized quantile regression”. *Advances in neural information processing systems* 32 (2019) (cit. on pp. 31, 32, 104, 175).
- [Romano, 2020b] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. “Classification with valid and adaptive coverage”. *Advances in Neural Information Processing Systems* 33 (2020), pp. 3581–3591 (cit. on p. 33).
- [Russell, 2019] Chris Russell. “Efficient Search for Diverse Coherent Explanations”. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 20–28 (cit. on pp. 85, 87).
- [Sadinle, 2019] Mauricio Sadinle, Jing Lei, and Larry Wasserman. “Least ambiguous set-valued classifiers with bounded error levels”. *Journal of the American Statistical Association* 114.525 (2019), pp. 223–234 (cit. on p. 33).
- [Saleh, 2021] Resve A Saleh and AK Saleh. “Solution to the Non-Monotonicity and Crossing Problems in Quantile Regression”. *arXiv preprint arXiv:2111.04805* (2021) (cit. on p. 175).
- [Salmon, 2006] Wesley C Salmon. *Four decades of scientific explanation*. University of Pittsburgh press, 2006 (cit. on p. 5).
- [Saltelli, 2008] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, et al. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008 (cit. on p. 5).
- [Saunders, 1999] Craig Saunders, Alexander Gammerman, and Volodya Vovk. “Transduction with confidence and credibility” (1999) (cit. on p. 28).
- [Schaeffer, 2007] Satu Elisa Schaeffer. “Survey Graph clustering”. 2007 (cit. on pp. 59, 110).
- [Scornet, 2016] Erwan Scornet. “Random forests and kernel methods”. *IEEE Transactions on Information Theory* 62.3 (2016), pp. 1485–1500 (cit. on pp. 73, 92).
- [Scornet, 2015] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. “Consistency of random forests”. *The Annals of Statistics* 43.4 (2015), pp. 1716–1741 (cit. on pp. 45, 74–76, 103, 111, 145, 168).
- [Shah, 2018] Rajen Dinesh Shah and J. Peters. “The hardness of conditional independence testing and the generalised covariance measure”. *The Annals of Statistics* (2018) (cit. on p. 27).

-
- [Shapley, 1953] Lloyd S Shapley. “Greedy function approximation: A gradient boosting machine.” *Contribution to the Theory of Games 2* (1953), pp. 307–317 (cit. on pp. 21, 22, 37).
- [Shih, 2018] Andy Shih, Arthur Choi, and Adnan Darwiche. “A symbolic approach to explaining bayesian network classifiers”. *arXiv preprint arXiv:1805.03364* (2018) (cit. on pp. 68, 69).
- [Siems, 2023] Julien Siems, Konstantin Ditschuneit, Winfried Ripken, Alma Lindborg, Maximilian Schambach, Johannes S Otterbach, et al. “Curve Your Enthusiasm: Concurvity Regularization in Differentiable Generalized Additive Models”. *arXiv preprint arXiv:2305.11475* (2023) (cit. on p. 17).
- [Singh, 2021] Chandan Singh, Keyan Nasser, Yan Shuo Tan, Tiffany Tang, and Bin Yu. “imodels: a python package for fitting interpretable models”. *Journal of Open Source Software* 6.61 (2021), p. 3192 (cit. on pp. 9, 17, 19, 161).
- [Slack, 2020] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. “Fooling lime and shap: Adversarial attacks on post hoc explanation methods”. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 180–186 (cit. on p. 69).
- [Sobol, 1990] Il’ya Meerovich Sobol’. “On sensitivity estimation for nonlinear mathematical models”. *Matematicheskoe modelirovanie* 2.1 (1990), pp. 112–118 (cit. on pp. 22, 53).
- [Song, 2016] Eunhye Song, Barry L Nelson, and Jeremy Staum. “Shapley effects for global sensitivity analysis: Theory and computation”. *SIAM/ASA Journal on Uncertainty Quantification* 4.1 (2016), pp. 1060–1083 (cit. on pp. 22, 23).
- [Spooner, 2021] Thomas Spooner, Danial Dervovic, Jason Long, Jon Shepard, Jiahao Chen, and Daniele Magazzeni. “Counterfactual Explanations for Arbitrary Regression Models”. *ArXiv abs/2106.15212* (2021) (cit. on p. 25).
- [Stone, 1985a] Charles J Stone. “Additive regression and other nonparametric models”. *The annals of Statistics* 13.2 (1985), pp. 689–705 (cit. on p. 16).
- [Stone, 1985b] Charles J. Stone. “Additive Regression and Other Nonparametric Models”. *The Annals of Statistics* 13.2 (1985), pp. 689–705 (cit. on pp. 75, 111, 145).
- [Strobl, 2007] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. “Bias in random forest variable importance measures: Illustrations, sources and a solution”. *BMC bioinformatics* 8.1 (2007), pp. 1–21 (cit. on p. 76).
- [Strumbelj, 2010] Erik Strumbelj and Igor Kononenko. “An Efficient Explanation of Individual Classifications using Game Theory”. *Journal of Machine Learning Research* 11 (2010), pp. 1–18 (cit. on p. 37).
- [Tan, 2022] Yan Shuo Tan, Chandan Singh, Keyan Nasser, Abhineet Agarwal, and Bin Yu. “Fast interpretable greedy-tree sums (FIGS)”. *arXiv preprint arXiv:2201.11931* (2022) (cit. on p. 19).
- [Tibshirani, 1996] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288 (cit. on pp. 16, 19).
- [Tibshirani, 2019] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. “Conformal prediction under covariate shift”. *Advances in neural information processing systems* 32 (2019) (cit. on pp. 28, 30, 35, 106, 116).
- [Traag, 2019] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. “From Louvain to Leiden: guaranteeing well-connected communities”. *Scientific reports* 9.1 (2019), pp. 1–12 (cit. on p. 110).
- [Ustun, 2019a] Berk Ustun, Alexander Spangher, and Yang Liu. “Actionable Recourse in Linear Classification”. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019) (cit. on p. 85).
- [Ustun, 2019b] Berk Ustun, Alexander Spangher, and Yang Liu. “Actionable recourse in linear classification”. *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 10–19 (cit. on p. 26).
- [Valiant, 1984] Leslie G Valiant. “A theory of the learnable”. *Communications of the ACM* 27.11 (1984), pp. 1134–1142 (cit. on pp. 12, 34, 101).
- [Vapnik, 1971] Vladimir Naumovich Vapnik. “Chervonenkis: On the uniform convergence of relative frequencies of events to their probabilities”. 1971 (cit. on pp. 149, 153).
- [Verdinelli, 2023] Isabella Verdinelli and Larry Wasserman. “Feature Importance: A Closer Look at Shapley Values and LOCO”. *arXiv preprint arXiv:2303.05981* (2023) (cit. on pp. 22, 54, 58, 66).
- [Verma, 2020] Sahil Verma, John P. Dickerson, and Keegan Hines. “Counterfactual Explanations for Machine Learning: A Review”. *CoRR abs/2010.10596* (2020). arXiv: 2010.10596 (cit. on pp. 85, 86).
- [Visani, 2020] Giorgio Visani, Enrico Bagli, and Federico Chesani. “OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms”. *ArXiv abs/2006.05714* (2020) (cit. on p. 57).

-
- [Vovk, 2012] Vladimir Vovk. “Conditional validity of inductive conformal predictors”. *Asian conference on machine learning*. PMLR, 2012, pp. 475–490 (cit. on pp. 34, 35, 102, 111).
- [Vovk, 2005] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005 (cit. on pp. 7, 28, 30, 31, 33, 34, 100, 102).
- [Vovk, 1999] Volodya Vovk, Alexander Gammerman, and Craig Saunders. “Machine-learning applications of algorithmic randomness” (1999) (cit. on p. 28).
- [Wachter, 2017] Sandra Wachter, Brent Daniel Mittelstadt, and Chris Russell. “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR”. *Cybersecurity* (2017) (cit. on pp. 24, 26, 68, 85, 86).
- [Wager, 2017] Stefan Wager and Susan Athey. *Estimation and Inference of Heterogeneous Treatment Effects using Random Forests*. 2017. arXiv: 1510.04342 [stat.ME] (cit. on pp. 73, 74, 92, 105, 146, 168).
- [Wahba, 1990] Grace Wahba. *Spline models for observational data*. SIAM, 1990 (cit. on p. 16).
- [Wald, 1943] Abraham Wald. “An extension of Wilks’ method for setting tolerance limits”. *The Annals of Mathematical Statistics* 14.1 (1943), pp. 45–55 (cit. on pp. 34, 101).
- [Wang, 2020] Eric Wang, Pasha Khosravi, and Guy Van den Broeck. “Towards Probabilistic Sufficient Explanations”. *Extending Explainable AI Beyond Deep Models and Classifiers Workshop at ICML (XXAI)*. 2020 (cit. on pp. 68, 69, 72).
- [Washington, 2018] Anne L Washington. “How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate”. *Colo. Tech. LJ* 17 (2018), p. 131 (cit. on pp. 4, 82, 97).
- [Wei, 2015] Pengfei Wei, Zhenzhou Lu, and Jingwen Song. “Variable importance analysis: a comprehensive review”. *Reliability Engineering & System Safety* 142 (2015), pp. 399–432 (cit. on p. 5).
- [Wilks, 1941] Samuel S Wilks. “Determination of sample sizes for setting tolerance limits”. *The Annals of Mathematical Statistics* 12.1 (1941), pp. 91–96 (cit. on pp. 34, 101).
- [Williamson, 2021] Brian D Williamson, Peter B Gilbert, Noah R Simon, and Marco Carone. “A general framework for inference on algorithm-agnostic variable importance”. *Journal of the American Statistical Association* (2021), pp. 1–14 (cit. on pp. 58, 59).
- [Williamson, 2020] Brian D. Williamson and Jean Feng. *Efficient nonparametric statistical inference on population feature importance using Shapley values*. 2020. arXiv: 2006.09481 [stat.ME] (cit. on pp. 23, 42, 52).
- [Xu, 2019] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. “Modeling Tabular data using Conditional GAN”. *NeurIPS*. 2019 (cit. on p. 95).
- [Yoon, 2018] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. “INVASE: Instance-wise variable selection using neural networks”. *International Conference on Learning Representations*. 2018 (cit. on pp. 69, 80).
- [Yu, 2013] Bin Yu. “Stability” (2013) (cit. on p. 15).
- [Yu, 2002] Chong Ho Yu. “Resampling methods: concepts, applications, and justification”. *Practical Assessment, Research, and Evaluation* 8.1 (2002), p. 19 (cit. on p. 7).
- [Zafar, 2019] Muhammad Rehman Zafar and Naimul Mefraz Khan. “DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems”. *ArXiv abs/1906.10263* (2019) (cit. on p. 57).
- [Zaffran, 2023] Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, and Yaniv Romano. “Conformal Prediction with Missing Values”. *arXiv preprint arXiv:2306.02732* (2023) (cit. on p. 35).
- [Zaffran, 2022a] Margaux Zaffran, Olivier Feron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. “Adaptive Conformal Predictions for Time Series”. *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 25834–25866 (cit. on p. 35).
- [Zaffran, 2022b] Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. “Adaptive conformal predictions for time series”. *International Conference on Machine Learning*. PMLR. 2022, pp. 25834–25866 (cit. on p. 28).
- [Zhang, 2022] Rui Zhang, Rui Xin, Margo Seltzer, and Cynthia Rudin. “Optimal Sparse Regression Trees”. *arXiv preprint arXiv:2211.14980* (2022) (cit. on p. 18).
- [Zhang, 2019] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. “Why Should You Trust My Explanation?” Understanding Uncertainty in LIME Explanations”. *arXiv preprint arXiv:1904.12991* (2019) (cit. on p. 57).

-
- [Zhou, 2023] Yichen Zhou, Zhengze Zhou, and Giles Hooker. “Approximation trees: statistical reproducibility in model distillation”. *Data Mining and Knowledge Discovery* (2023), pp. 1–39 (cit. on p. 15).
- [Zhou, 2021] Zhengze Zhou, Giles Hooker, and Fei Wang. “S-LIME: Stabilized-LIME for Model Explanation”. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021) (cit. on p. 57).
- [Zou, 2005] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320 (cit. on p. 16).

Appendix for Chapter 2

1 Proof of SV invariance for transformed continuous variables

Proposition 1.1. *Let f and $\tilde{f} = f \circ \varphi^{(-1)}$ its reparametrization, then we have for all $i \in [p]$, and $\mathbf{u} = \varphi(\mathbf{x})$:*

$$\phi_{x_i}(f) = \phi_{u_i}(\tilde{f}).$$

Proof. It is a direct application of the change of variables formula. If $g(\mathbf{x})$ is the joint density of X_1, \dots, X_p , the transformed variable $\mathbf{U} = \varphi(\mathbf{X}) = (\varphi_1(X_1), \dots, \varphi_p(X_p))$ has density $\tilde{g}(\mathbf{u}) = g \circ \varphi^{(-1)}(\mathbf{u}) \times \prod_{i=1}^p |J(\varphi_i^{(-1)})(u_i)|$ where $|J(\varphi_i^{(-1)})(u_i)|$ represents the determinant of the Jacobian of $\varphi_i^{(-1)}$ evaluated at u_i . We have

$$\tilde{g}(u_{\bar{S}}|u_S) = \frac{\tilde{g}(u_{\bar{S}}, u_S)}{\tilde{g}_S(u_S)} = g\left(\varphi_{\bar{S}}^{(-1)}(u_{\bar{S}})|\varphi_S^{(-1)}(u_S)\right) \times \prod_{i \in \bar{S}} |J(\varphi_i^{(-1)})(u_i)|.$$

The computation of the reduced predictor is straightforward

$$\begin{aligned} \mathbb{E}[f(\mathbf{X})|\mathbf{X}_S = \mathbf{x}_S] &= \int f(\mathbf{x}_S, \mathbf{x}_{\bar{S}})g(\mathbf{x}_{\bar{S}}|\mathbf{x}_S)d\mathbf{x}_{\bar{S}} \\ &= \int f\left(\varphi_S^{(-1)} \circ \varphi_S(\mathbf{x}_S), \varphi_{\bar{S}}^{(-1)} \circ \varphi_{\bar{S}}(\mathbf{x}_{\bar{S}})\right) g(\mathbf{x}_{\bar{S}}|\mathbf{x}_S)d\mathbf{x}_{\bar{S}} \\ &= \int \tilde{f}(\mathbf{u}_S, \mathbf{u}_{\bar{S}})g\left(\varphi_{\bar{S}}^{(-1)}(\mathbf{u}_{\bar{S}})|\varphi_S^{(-1)}(\mathbf{u}_S)\right) \prod_{i \in \bar{S}} |J\varphi^{(-1)}(u_i)| d\mathbf{u}_{\bar{S}} \\ &= \mathbb{E}\left[\tilde{f}(\mathbf{U}_S, \mathbf{U}_{\bar{S}})|\mathbf{U}_S = \mathbf{u}_S\right]. \end{aligned}$$

The equality of Shapley Values directly results from the equality of the reduced predictors. \square

2 Proof of SV invariance for encoded categorical variables

Proposition 2.1. *Given a predictor $f : \mathbb{R} \times \{1, \dots, K\} \rightarrow \mathbb{R}$ and its reparametrization \tilde{f} using Dummy Encoding defined as $\tilde{f} : \mathbb{R} \times \{0, 1\}^{K-1} \rightarrow \mathbb{R}$ such that $f(X, Z) \triangleq \tilde{f}(X, Z_1, \dots, Z_{K-1})$, we have*

$$\begin{cases} \phi_{z_{1:K-1}}(\tilde{f}) &= \phi_z(f) \\ \phi_x(\tilde{f}; z_{1:K-1}) &= \phi_x(f). \end{cases} \quad (3)$$

We recall the expression of the SV of the two variables $X \in \mathbb{R}$ and $Z \in \{1, \dots, K\}$ in Equation (4). The roles of the variables X, Z are symmetric, and the categorical or quantitative nature of the variable does not have any impact on the computation of the SV, as demonstrated below. Let's consider an observation $\mathbf{x} = (x, z)$, then

$$\begin{cases} \phi_x(f) = \frac{1}{2} (\mathbb{E}[f(X, Z) | X = x] - \mathbb{E}[f(X, Z)]) + \frac{1}{2} (f(x, z) - \mathbb{E}[f(X, Z) | Z = z]) \\ \phi_z(f) = \frac{1}{2} (\mathbb{E}[f(X, Z) | Z = z] - \mathbb{E}[f(X, Z)]) + \frac{1}{2} (f(x, z) - \mathbb{E}[f(X, Z) | X = x]) \end{cases} \quad (4)$$

Proof. Let us consider the Dummy Encoding (DE) $\varphi : z \mapsto (z_1, \dots, z_{K-1})$ without loss of generality, then the observation (x, z) is reparametrized as $(x, z_{1:K-1})$, and by construction of φ , $\exists! z \in \{1, \dots, K\}$ such that $\varphi(z) = z_{1:K-1}$. By taking the coalition of $z_{1:K-1}$ or considering them as a single variable, we have

$$\begin{aligned} \phi_{z_{1:K-1}}(\tilde{f}) &= \frac{1}{2} \left(\mathbb{E}_{\tilde{P}} [\tilde{f}(X, Z_{1:K-1}) | Z_{1:K-1} = z_{1:K-1}] - \mathbb{E}_{\tilde{P}} [\tilde{f}(X, Z_{1:K-1})] \right) \\ &\quad + \frac{1}{2} \left(\mathbb{E}_{\tilde{P}} [\tilde{f}(X, Z_{1:K-1}) | X = x, Z_{1:K-1} = z_{1:K-1}] - \mathbb{E}_{\tilde{P}} [\tilde{f}(X, Z_{1:K-1}) | X = x] \right). \end{aligned} \quad (5)$$

Recall that for any $z \in \{1, \dots, K\}$, and $\varphi(z) = z_{1:K-1}$, we have $\mathbb{P}(Z = z) = \mathbb{P}(Z_{1:K-1} = z_{1:K-1})$. We denote P the measure law of (X, Z) and \tilde{P} the pushforward measure of P under the transformation φ , then we have

$$\begin{aligned} \mathbb{E}_{\tilde{P}} [\tilde{f}(X, Z_{1:K-1}) | Z_{1:K-1} = z_{1:K-1}] &= \int \tilde{f}(x, z_{1:K-1}) \frac{\tilde{P}(dx, z_{1:K-1})}{\mathbb{P}(Z_{1:K-1} = z_{1:K-1})} \\ &= \int \tilde{f}(x, \varphi(z)) \frac{\tilde{P}(dx, \varphi(z))}{\mathbb{P}(Z_{1:K-1} = z_{1:K-1})} \\ &= \int f(x, z) \frac{P(dx, z)}{\mathbb{P}(Z = z)} \\ &= \mathbb{E}[f(X, Z) | Z = z]. \end{aligned}$$

As a result, we have

$$\begin{aligned} &\mathbb{E}_{\tilde{P}} [\tilde{f}(X, Z_{1:K-1}) | Z_{1:K-1} = z_{1:K-1}] - \mathbb{E}_{\tilde{P}} [\tilde{f}(X, Z_{1:K-1})] \\ &= E_P [\tilde{f}(X, \varphi(Z)) | Z = z] - E_P [\tilde{f}(X, \varphi(Z))] \\ &= E_P [f(X, Z) | Z = z] - E_P [f(X, Z)]. \end{aligned}$$

We also have

$$\begin{aligned}
& \mathbb{E}_{\tilde{P}} \left[\tilde{f}(X, Z_{1:K-1}) \mid X = x, Z_{1:K-1} = z_{1:K-1} \right] - \mathbb{E}_{\tilde{P}} \left[\tilde{f}(X, Z_{1:K-1}) \mid X = x \right] \\
&= \tilde{f}(x, z_{1:K-1}) - \mathbb{E}_P \left[\tilde{f}(X, \varphi(Z)) \mid X = x \right] \\
&= \tilde{f}(x, \varphi(z)) - \mathbb{E}_P \left[\tilde{f}(X, \varphi(Z)) \mid X = x \right] \\
&= f(x, z) - \mathbb{E}_P [f(X, Z) \mid X = x].
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
\phi_{z_{1:K-1}}(\tilde{f}) &= \frac{1}{2} (\mathbb{E}_P [f(X, Z) \mid Z = z] - \mathbb{E}_P [f(X, Z)]) + \frac{1}{2} (f(x, z) - \mathbb{E}_P [f(X, z) \mid X = x]) \\
&= \phi_z(f).
\end{aligned}$$

Similarly, we can derive that $\phi_x(\tilde{f}; z_{1:K-1}) = \phi_x(f)$. □

Proposition 2.2. *If $X \sim \mathcal{N}(\mu, \Sigma)$, then $X_{\bar{S}} \mid X_S = x_S$ is also multivariate gaussian with mean $\mu_{\bar{S} \mid S}$ and covariance matrix $\Sigma_{\bar{S} \mid S}$ equal:*

$$\mu_{\bar{S} \mid S} = \mu_{\bar{S}} + \Sigma_{\bar{S}, S} \Sigma_{S, S}^{-1} (x_S - \mu_S) \quad \text{and} \quad \Sigma_{\bar{S} \mid S} = \Sigma_{\bar{S} \bar{S}} - \Sigma_{\bar{S} S} \Sigma_{S S}^{-1} \Sigma_{S, \bar{S}}.$$

3 Proof of the limitation of SV as local explanation

Proposition 3.1. *Let us assume that we have $X \in \mathbb{R}^p$, $\mathbf{X} \in \mathcal{N}(0, I_p)$ and a piece-wise linear predictor f defined as:*

$$f(X) = (a_1 X_1 + a_2 X_2) \mathbb{1}_{X_5 \leq 0} + (a_3 X_3 + a_4 X_4) \mathbb{1}_{X_5 > 0}. \quad (6)$$

Even if we choose an observation \mathbf{x} such that $x_5 \leq 0$ and the predictor only uses x_1, x_2 , the SV of ϕ_{x_3}, ϕ_{x_4} is not necessarily zero.

Proof.

$$\begin{aligned}
\phi_{x_3} &= \frac{1}{p} \sum_{S \subseteq [p] \setminus \{3\}} \binom{p-1}{|S|}^{-1} \left(f_{S \cup 3}(\mathbf{x}_{S \cup 3}) - f_S(\mathbf{x}_S) \right) \\
&= \frac{1}{p} \sum_{S \subseteq [p] \setminus \{3, 5\}} \binom{p-1}{|S|}^{-1} \left(f_{S \cup 3}(\mathbf{x}_{S \cup 3}) - f_S(\mathbf{x}_S) \right) \\
&\quad + \frac{1}{p} \sum_{S \subseteq [p] \setminus \{3, 5\}} \binom{p-1}{|S|+1}^{-1} \left(f_{S \cup \{3, 5\}}(\mathbf{x}_{S \cup \{3, 5\}}) - f_{S \cup 5}(\mathbf{x}_{S \cup 5}) \right).
\end{aligned} \quad (7)$$

The second term is zero. Indeed, $\forall S \subseteq [p] \setminus \{3, 5\}$

$$f_{S \cup \{3, 5\}}(\mathbf{x}_{S \cup \{3, 5\}}) - f_{S \cup 5}(\mathbf{x}_{S \cup 5}) = 0$$

Because, if we condition on the event $\{X_5 = x_5\}$ with $x_5 \leq 0$

$$\begin{aligned}
f_{S \cup \{3,5\}}(\mathbf{x}_{S \cup \{3,5\}}) &= \mathbb{E} \left[(a_1 X_1 + a_2 X_2) \mathbf{1}_{X_5 \leq 0} + (a_3 X_3 + a_4 X_4) \mathbf{1}_{X_5 > 0} \mid X_{S \cup \{3,5\}} = \mathbf{x}_{S \cup \{3,5\}} \right] \\
&= \mathbb{E} \left[(a_1 X_1 + a_2 X_2) \mathbf{1}_{X_5 \leq 0} \mid X_{S \cup \{3,5\}} = \mathbf{x}_{S \cup \{3,5\}} \right] \quad (\text{because } x_5 \leq 0) \\
&= \mathbb{E} [(a_1 X_1 + a_2 X_2) \mid X_{S \cup 5} = \mathbf{x}_{S \cup 5}] \quad (\text{because the quantities are } \perp\!\!\!\perp \text{ of } X_3) \\
&= f_{S \cup 5}(\mathbf{x}_{S \cup 5})
\end{aligned}$$

□

The first term of (7) is the classic marginal contribution of SV in the linear model. For all $S \subseteq [p] \setminus \{3, 5\}$

$$\begin{aligned}
f_{S \cup 3}(\mathbf{x}_{S \cup 3}) &= \mathbb{E} [a_1 X_1 + a_2 X_2 \mid X_{S \cup 3} = \mathbf{x}_{S \cup 3}] \mathbb{P}(X_5 \leq 0 \mid X_{S \cup 3} = \mathbf{x}_{S \cup 3}) \\
&\quad + \mathbb{E} [a_3 X_3 + a_4 X_4 \mid X_{S \cup 3} = \mathbf{x}_{S \cup 3}] \mathbb{P}(X_5 > 0 \mid X_{S \cup 3} = \mathbf{x}_{S \cup 3}) \\
&= \mathbb{E} [a_1 X_1 + a_2 X_2 \mid X_S = \mathbf{x}_S] \mathbb{P}(X_5 \leq 0) + (\mathbb{E} [a_4 X_4 \mid X_S = \mathbf{x}_S] + a_3 \mathbf{x}_3) \mathbb{P}(X_5 > 0) \\
&= f_S(\mathbf{x}_S) + \mathbb{P}(X_5 > 0) (a_3 (\mathbf{x}_3 - \mathbb{E}[X_3]))
\end{aligned}$$

Therefore,

$$\begin{aligned}
\phi_{x_3} &= \frac{1}{p} \sum_{S \subseteq [p] \setminus \{3,5\}} \binom{p-1}{|S|}^{-1} \mathbb{P}(X_5 > 0) (a_3 (\mathbf{x}_3 - \mathbb{E}[X_3])) \\
&= K (a_3 (\mathbf{x}_3 - \mathbb{E}[X_3])),
\end{aligned}$$

where K is a constant. The computation of ϕ_{x_4} is obtained similarly by symmetry.

4 Relation between the Algorithm 1 (TreeSHAP with path-dependent) and \hat{f}^{SHAP}

In section 3.1, we claim that the recursive Algorithm 1 introduced in [Lundberg, 2018; Lundberg, 2020b] and shown in Figure 1 assumes that the probabilities can be factored using the path of the decision tree as follows:

$$\mathbb{P}_{P_{\mathbf{X}}^{SHAP}} \left(\prod_{k=1}^{d_m} I_k^m \mid \mathbf{X}_S = \mathbf{x}_S \right) = \delta_S(N_1^m) \times \prod_{i=2: N_i^m \notin S}^{d_m} \mathbb{P} \left(X_{N_i^m} \in I_i^m \mid \prod_{k=1}^i X_{N_{k-1}^m} \in I_{k-1}^m \right) \quad (8)$$

with $\delta_S(N_1^m) = \mathbb{P}(X_{N_1^m} \in I_1^m)$ if $N_1^m \notin S$, and 1 otherwise.

Algorithm 1 Estimating $E[f(\mathbf{x}) \mid \mathbf{x}_S]$

```

procedure EXPVALUE( $x, S, tree = \{v, a, b, t, r, d\}$ )
  procedure G( $j, w$ )
    if  $v_j \neq internal$  then
      return  $w \cdot v_j$ 
    else
      if  $d_j \in S$  then
        return G( $a_j, w$ ) if  $x_{d_j} \leq t_j$  else G( $b_j, w$ )
      else
        return G( $a_j, wr_{a_j}/r_j$ ) + G( $b_j, wr_{b_j}/r_j$ )
      end if
    end if
  end procedure
  return G(0, 1)
end procedure

```

Figure 1: Algorithm 1 (path-dependent Tree SHAP) in [Lundberg, 2018; Lundberg, 2020b], where v is a vector of node values, which takes the value *internal* for internal nodes. The vectors a and b represent the left and right node indexes for each internal node. The root node has index 0. The vector t contains the thresholds for each internal node, and d is a vector of indexes of the features used for splitting in internal nodes. The vector r represents the cover of each node (i.e., how many data samples fall in that sub-tree). The weight w measures what proportion of the training samples matching the conditioning set S fall into each leaf.

To show the link between between \hat{f}^{SHAP} and Algorithm 1 (path-dependent TreeSHAP), we choose an observation $\mathbf{x} = (x_0, x_1, x_2, x_3) = (2, 3, 0.5, -1)$ and aim to compute $\mathbb{E}[f(\mathbf{X}) \mid x_0 = 2, x_2 = 0.5]$ where f is the tree presented in Figure 2 using $\hat{f}^{(SHAP)}$. \mathbf{x} is compatible with Leaf 6, 7, 11, 13, 14, we denote $f_6, f_7, f_{11}, f_{13}, f_{14}$ the value of each leaf respectively.

The algorithm's steps and output are outlined below for this observation. Using Equation (8), we observe $\hat{f}^{(SHAP)}$ gives the same output. Let's denote \hat{P} the empirical distribution of \mathcal{D}_n , and $S = [0, 2]$ with $\mathbf{x}_S = (2, 0.5)$, then we have

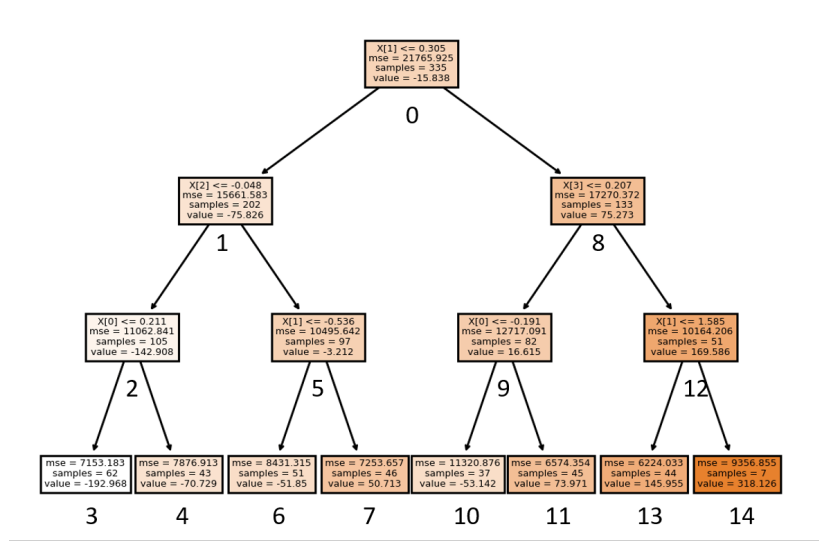


Figure 2: An example of decision tree used to illustrate the link between $\hat{f}^{(SHAP)}$ and Algorithm 1 (path-dependent TreeSHAP)

$$\begin{aligned}
& \hat{f}_S^{(SHAP)}(\mathbf{x}_S) \\
&= \mathbb{P}_{\hat{P}}(X_1 \leq 0.305) \times \mathbb{P}_{\hat{P}}(X_1 \leq -0.536 \mid X_2 > -0.048, X_1 \leq 0.305) \times f_6 \\
&\quad + \mathbb{P}_{\hat{P}}(X_1 \leq 0.305) \times \mathbb{P}_{\hat{P}}(X_1 > -0.536 \mid X_2 > -0.048, X_1 \leq 0.305) \times f_7 \\
&\quad + \mathbb{P}_{\hat{P}}(X_1 > 0.305) \times \mathbb{P}_{\hat{P}}(X_3 \leq 0.207 \mid X_1 > 0.305) \times f_{11} \\
&\quad + \mathbb{P}_{\hat{P}}(X_1 > 0.305) \times \mathbb{P}_{\hat{P}}(X_3 > 0.207 \mid X_1 > 0.305) \times \mathbb{P}_{\hat{P}}(X_1 \leq 1.585 \mid X_3 > 0.207, X_1 > 0.305) \times f_{13} \\
&\quad + \mathbb{P}_{\hat{P}}(X_1 > 0.305) \times \mathbb{P}_{\hat{P}}(X_3 > 0.207 \mid X_1 > 0.305) \times \mathbb{P}_{\hat{P}}(X_1 > 1.585 \mid X_3 > 0.207, X_1 > 0.305) \times f_{14} \\
&= (202/335) \times (51/97) \times (-51.85) + (202/335) \times (46/97) \times (50.713) + (133/335) \times (82/133) \times (73.971) \\
&\quad + (133/335) \times (51/133) \times (44/51) \times (145.955) + (133/335) \times (51/133) \times (7/51) \times (318.125) \\
&= 41.98
\end{aligned}$$

Step	Calculus
0	$G(0, 1)$
1	$G(1, 202/335) + G(8, 133/335)$
2	$G(5, 202/335) + G(9, 88/335) + G(12, 51/335)$
3	$G(6, (202/335) \times (51/97)) + G(7, (202/335) \times (46/97)) + G(11, 82/335) + G(13, 44/335) + G(14, 7/335)$
4	$-(202/335) \times (51/97) \times 51,85 + (202/335) \times (46/97) \times 50,713 + (82/335) \times 73,971 + (44/335) \times 145,955 + (7/335) \times 318,126$
5	$= 41.98$

5 Additional experiments

5.1 Impact of quantile discretization

The table below shows the impact of discretization on the performance of a Random Forest on UCI datasets.

Dataset	Breiman’s RF	q=2	q=5	q=10	q=20
Authentication	0.0002	0.08	0.002	0.0005	0.0004
Diabetes	0.17	0.23	0.18	0.18	0.18
Haberman	0.32	0.35	0.30	0.32	0.30
Heart Statlog	0.10	0.10	0.10	0.10	0.10
Hepatitis	0.13	0.15	0.14	0.14	0.13
Ionosphere	0.02	0.07	0.03	0.02	0.02
Liver Disorders	0.23	0.32	0.27	0.25	0.24
Sonar	0.07	0.09	0.07	0.07	0.07
Spambase	0.01	0.14	0.03	0.02	0.01
Titanic	0.13	0.15	0.14	0.14	0.13
Wilt	0.007	0.15	0.03	0.02	0.02

Table 1: Accuracy, measured by 1-AUC on UCI datasets [Dua, 2017a], for two algorithms: Breiman’s random forests and random forests with splits limited to q-quantiles, for $q \in \{2, 5, 10, 20\}$. Table 5 in [Bénard, 2021c].

5.2 The differences between Coalition and sum on Census Data

We use the UCI Adult Census Dataset [Dua, 2017b]. We keep only four highly predictive categorical variables: Marital Status, Workclass, Race, Education and use a Random Forest which has a test accuracy of 86%. We compare the Global SV by taking the coalition or sum of the OHE modalities. Global SV for a feature X_j are defined as $I_j = \sum_{i=0}^N |\phi_{x_{i,j}}|/N$, where N is the number of observations.

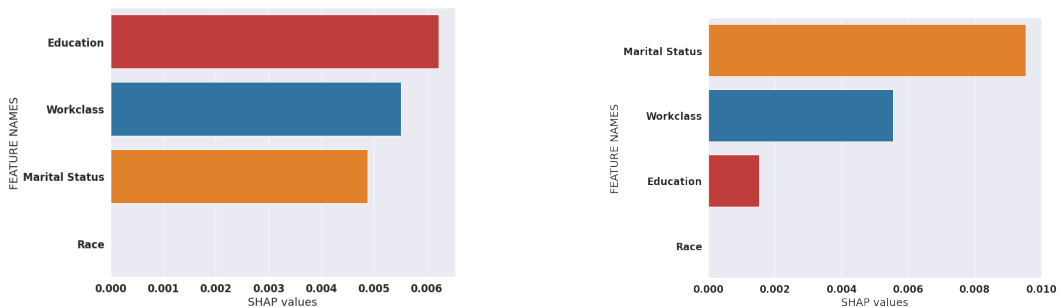


Figure 3: Difference between the global absolute value of SV: sum (left) vs coalition (right) of dummies of individual with modalities: Married, local gov, others, 1st-4th.

In Figure 3, we see differences between the global SV with coalition and sum with N=5000. The ranking of the variables changes, e.g. Education goes from important with sum to not important with the coalition. We also compute the proportion of order inversion over the 5000 observations that are chosen randomly. The ranking of variables is changed in 10% of the cases. Note that this difference may increase or decrease depending on the data.

6 Experiments setting

6.1 Parameters of the model from Section 2.3

Recall that the model is a linear predictor f , with categorical Z and 3 continuous variables $\mathbf{X} = (X_1, X_2, X_3)$, defined as $f(\mathbf{X}, Z) = B_Z \mathbf{X}$ with $B_Z \in \mathbb{R}^3$, $\mathbf{X}|Z = z \sim \mathcal{N}(\mu_z, \Sigma_z)$ and $\mathbb{P}(Z = z) = \pi_z$, $Z \in \{a, b, c\}$.

For the experiments in Figure 2.1 and 2.2, we set $\pi_z = \frac{1}{3}$, $\mu_z = 0$ for all $z \in \{a, b, c\}$. We generated random matrices from Wishart distribution for the covariance matrices, and the values used are:

$$\Sigma_a = \begin{bmatrix} 0.41871254 & -0.790061361 & 0.46956991 \\ -0.79006136 & 1.90865098 & -0.82571655 \\ 0.46956991 & -0.82571655 & 0.95835472 \end{bmatrix}, \Sigma_b = \begin{bmatrix} 0.55326081 & 0.11811951 & -0.70677924 \\ 0.11811951 & 2.73312979 & -2.94400196 \\ -0.70677924 & -2.94400196 & 4.22105088 \end{bmatrix},$$

$$\Sigma_c = \begin{bmatrix} 9.2859966 & 1.12872646 & 2.4224434 \\ 1.12872646 & 0.92891237 & -0.14373393 \\ 2.4224434 & -0.14373393 & 1.81601676 \end{bmatrix} \text{ for } y \in \{a, b, c\} \text{ respectively.}$$

The coefficients are $B_a = [1, 3, 5]$, $B_b = [-5, -10, -8]$, $B_c = [6, 1, 0]$, and the selected observation in Figure 2.1 is $x = [0.35, -1.61, -0.11, 1., 0., 0.]$.

6.2 Parameters of the model from Section 4

The data $\mathcal{D} = (\mathbf{X}_i, Y_i)_{1 \leq i \leq n}$ are generated from a linear regression $Y = B^t \mathbf{X}$ with $n = 10^4$, $\mathbf{X} \in \mathbb{R}^p$, $\mathbf{X} \in \mathcal{N}(0, \Sigma)$, $\Sigma = \rho J_p + (\rho - 1)I_p$ with $p = 5$, $\rho = 0.7$, I_p is the identity matrix, J_p is all-ones matrix and a linear predictor $Y = B^t \mathbf{X}$. The coefficient $B = [6.49, -2.44, -2.11, -4.29, 3.46]$ for the continuous case and $d=3$, $B = [6.49, -2.44, 0]$ for the discrete case.

We used the decision tree of scikit-learn trained on \mathcal{D} with the defaults parameters. The Mean Squared Error (MSE) are $\text{MSE} = 4.39$ for the continuous case and $\text{MSE} = 2.88$ for the discrete case.

Appendix for Chapter 3

7 Proof of Theorem 2.2

Theorem 7.1. *Let f be a piecewise linear function with m components defined by the collection $\{f_{|_{A_1}}, \dots, f_{|_{A_m}}\}$, where $\cup_{k=1}^m A_k = \mathcal{X}$. The regions A_k are disjoint hyperrectangles, specifically $A_k = \otimes_{i=1}^p A_{i,k}$, where $A_{i,k} = [l_{i,k}, r_{i,k}]$ with $l_{i,k}, r_{i,k} \in \overline{\mathbb{R}}$. Each component $f_{|_{A_k}}$ is represented as $f_k(\mathbf{X}) = \sum_{i=1}^p a_{i,k} X_i + b_k$, where the coefficients $a_{i,k}$ and b_k are real numbers. Consequently, f is defined as:*

$$f(\mathbf{X}) = \sum_{k=1}^m \left(\sum_{i=1}^p a_{i,k} X_i + b_k \right) \mathbf{1}_{A_k}(\mathbf{X}).$$

Consider an observation $\mathbf{x} = (x_1, \dots, x_p) \in A_{k^*}$, where $k^* \in \{1, \dots, m\}$, sampled from a distribution $P_{\mathbf{X}}$ with independent covariates such that the model only used $f_{k^*}(\mathbf{x})$ as $f(\mathbf{x}) = f_{k^*}(\mathbf{x})$ on A_{k^*} . The Local SV of a given feature-value $X_l = x_l$ is equal to

$$\phi_{x_l} = \sum_{k=1}^m \phi_{x_l}^k,$$

and $\phi_{x_l}^k$ is defined as

$$\begin{aligned} \phi_{x_l}^k = & \left(\frac{\mathbf{1}_{A_{l,k}}(x_l)}{\mathbb{P}(X_l \in A_{l,k})} - 1 \right) \sum_{S \subseteq D \setminus \{l\}} w(S) v_k(S) \\ & + a_{l,k} \left(x_l - \frac{\mathbb{E}[X_l \mathbf{1}_{A_{l,k}}(X_l)]}{\mathbb{P}(X_l \in A_{l,k})} \right) \sum_{S \subseteq D \setminus \{l\}} w(S) \times \prod_{i \in S \cup l} \mathbf{1}_{A_{i,k}}(x_i) \prod_{j \in \bar{S}} \mathbb{P}(X_j \in A_{j,k}), \end{aligned} \quad (9)$$

where $w(S) = \frac{1}{p} \binom{|D|-1}{|S|}^{-1}$ and $v_k(S) = \mathbb{E}[f_k(\mathbf{X}) \mathbf{1}_{A_k}(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$. Equation (9) demonstrates that even if the model only uses $f_{k^*}(\mathbf{x})$ for a given observation \mathbf{x} , the Local SV \mathbf{x} may depend on the coefficients of the unused linear models f_k for $k \in \{1, \dots, m\} \setminus \{k^*\}$.

Proof. Let's assume that $\prod_{i=1}^p \mathbb{P}(X_i \in A_{i,k}) > 0$ and intercepts $b_k = 0$ for all $k = 1, \dots, m$ without loss of generality. Given an observation $\mathbf{x} = (x_1, \dots, x_p)$, we consider the Shapley Value

of a feature-value $X_l = x_l$ defined as

$$\phi_{\mathbf{x}_l} = \sum_{S \subseteq D \setminus \{l\}} w(S) [\Delta_l(S)],$$

where $w(S) = \frac{1}{p} \binom{|D|-1}{|S|}^{-1}$, $\Delta_l(S) = v(S \cup l) - v(S)$ represent the marginal contribution, and $v(S) = \mathbb{E}[f(\mathbf{X}) | \mathbf{X}_S = \mathbf{x}_S]$. Note that we can decompose $v(S)$ into m separate terms as following:

$$\begin{aligned} v(S) &= \mathbb{E} \left[\sum_{k=1}^m \left(\sum_{i=1}^p a_{i,k} X_i \right) \mathbb{1}_{A_k}(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S \right] \\ &= \sum_{k=1}^m \mathbb{E} \left[\left(\sum_{i=1}^p a_{i,k} X_i \right) \mathbb{1}_{A_k}(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S \right] \\ &= \sum_{k=1}^m v_k(S) \end{aligned}$$

Hence, we can decompose the Shapley Value of $\phi_{\mathbf{x}_l}$ due to the linearity property of SV as

$$\phi_{\mathbf{x}_l} = \sum_{k=1}^m \phi_{\mathbf{x}_l}^k,$$

where $\phi_{\mathbf{x}_l}^k$ corresponds to the SV compute using the value function $v_k(S)$. Therefore, we only need to prove that $\phi_{\mathbf{x}_l}^k$ for $k \neq k^*$ is not necessarily null to prove the Theorem. We have

$$\begin{aligned} v_k(S) &= \mathbb{E} \left[\left(\sum_{i=1}^p a_{i,k} X_i \right) \mathbb{1}_{A_k}(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S \right] \\ &= \mathbb{E} \left[\left(\sum_{i \in S} a_{i,k} X_i + \sum_{i \in \bar{S}} a_{i,k} X_i \right) \prod_{j=1}^p \mathbb{1}_{A_{j,k}}(X_j) \mid \mathbf{X}_S = \mathbf{x}_S \right] \\ &= \left(\sum_{i \in S} a_{i,k} \mathbf{x}_i \right) \prod_{j \in S} \mathbb{1}_{A_{j,k}}(\mathbf{x}_j) \prod_{j \in \bar{S}} \mathbb{P}(X_j \in A_{j,k}) \\ &\quad + \sum_{i \in \bar{S}} a_{i,k} \mathbb{E} \left[X_i \mathbb{1}_{A_{i,k}}(X_i) \right] \prod_{j \in \bar{S}: j \neq i} \mathbb{P}(X_j \in A_{j,k}) \prod_{j \in S} \mathbb{1}_{A_{j,k}}(\mathbf{x}_j). \end{aligned}$$

Similarly, we can write $v(S \cup l)$ as:

$$\begin{aligned}
v_k(S \cup l) &= \left(\sum_{i \in S \cup l} a_{i,k} x_i \right) \prod_{j \in S \cup l} \mathbb{1}_{A_{j,k}}(x_j) \prod_{j \in \overline{S \cup l}} \mathbb{P}(X_j \in A_{j,k}) \\
&+ \sum_{i \in \overline{S \cup l}} a_{i,k} \mathbb{E} \left[X_i \mathbb{1}_{A_{i,k}}(X_i) \right] \prod_{j \in \overline{S \cup l}; j \neq i} \mathbb{P}(X_j \in A_{j,k}) \prod_{j \in S \cup l} \mathbb{1}_{A_{j,k}}(x_j) \\
&= \left(\sum_{i \in S} a_{i,k} x_i \right) \prod_{j \in S} \mathbb{1}_{A_{j,k}}(x_j) \prod_{j \in \overline{S}} \mathbb{P}(X_j \in A_{j,k}) \times \frac{\mathbb{1}_{A_{l,k}}(x_l)}{\mathbb{P}(X_l \in A_{l,k})} \\
&+ a_{l,k} x_l \prod_{j \in S} \mathbb{1}_{A_{j,k}}(x_j) \prod_{j \in \overline{S}} \mathbb{P}(X_j \in A_{j,k}) \times \frac{\mathbb{1}_{A_{l,k}}(x_l)}{\mathbb{P}(X_l \in A_{l,k})} \\
&+ \sum_{i \in \overline{S}} a_{i,k} \mathbb{E} \left[X_i \mathbb{1}_{A_{i,k}}(X_i) \right] \prod_{j \in \overline{S}; j \neq i} \mathbb{P}(X_j \in A_{j,k}) \prod_{j \in S} \mathbb{1}_{A_{j,k}}(x_j) \times \frac{\mathbb{1}_{A_{l,k}}(x_l)}{\mathbb{P}(X_l \in A_{l,k})} \\
&- a_{l,k} \mathbb{E} \left[X_l \mathbb{1}_{A_{l,k}}(X_l) \right] \prod_{j \in \overline{S}; j \neq l} \mathbb{P}(X_j \in A_{j,k}) \prod_{j \in S} \mathbb{1}_{A_{j,k}}(x_j) \times \frac{\mathbb{1}_{A_{l,k}}(x_l)}{\mathbb{P}(X_l \in A_{l,k})}
\end{aligned}$$

The terms highlighted in red and teal respectively represent the negative and positive contributions of the variable $X_l = x_l$ in $v_k(S \cup l)$, the other terms will be put together to form $v_k(S)$ as follows

$$\begin{aligned}
v_k(S \cup l) &= \frac{\mathbb{1}_{A_{l,k}}(x_l)}{\mathbb{P}(X_l \in A_{l,k})} v_k(S) \\
&+ \frac{1}{\mathbb{P}(X_l \in A_{l,k})} \prod_{j \in S \cup l} \mathbb{1}_{A_{j,k}}(x_j) \prod_{j \in \overline{S}} \mathbb{P}(X_j \in A_{j,k}) \times a_{l,k} \left(x_l - \frac{\mathbb{E} \left[X_l \mathbb{1}_{A_{l,k}}(X_l) \right]}{\mathbb{P}(X_l \in A_{l,k})} \right).
\end{aligned}$$

Hence, the marginal contribution of coalition S of the SV $\phi_{x_l}^k$ is equal to:

$$\begin{aligned}
\Delta_l^k(S) &= v_k(S \cup l) - v_k(S) \\
&= v_k(S) \left(\frac{\mathbb{1}_{A_{l,k}}(x_l)}{\mathbb{P}(X_l \in A_{l,k})} - 1 \right) \\
&+ \prod_{j \in S \cup l} \mathbb{1}_{A_{j,k}}(x_j) \prod_{j \in \overline{S}} \mathbb{P}(X_j \in A_{j,k}) \times a_{l,k} \left(x_l - \frac{\mathbb{E} \left[X_l \mathbb{1}_{A_{l,k}}(X_l) \right]}{\mathbb{P}(X_l \in A_{l,k})} \right)
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\phi_{x_l}^k &= \left(\frac{\mathbb{1}_{A_{l,k}}(x_l)}{\mathbb{P}(X_l \in A_{l,k})} - 1 \right) \sum_{S \subseteq D \setminus \{l\}} w(S) v_k(S) \\
&+ a_{l,k} \left(x_l - \frac{\mathbb{E} \left[X_l \mathbb{1}_{A_{l,k}}(X_l) \right]}{\mathbb{P}(X_l \in A_{l,k})} \right) \sum_{S \subseteq D \setminus \{l\}} w(S) \times \prod_{j \in S \cup l} \mathbb{1}_{A_{j,k}}(x_j) \prod_{j \in \overline{S}} \mathbb{P}(X_j \in A_{j,k})
\end{aligned}$$

□

Appendix for Chapter 4

8 Proof of the Projected CDF Forest consistency

Here, we prove the main result of this chapter, Theorem 4.4, which is the uniform a.s. consistency of the PRF CDF $\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S)$ to the Projected CDF $F_S(y|\mathbf{X}_S = \mathbf{x}_S)$.

8.1 Main assumptions

We recall the main assumptions (4.1, 4.2, 4.3) for the sake of clarity.

Assumption 8.1. For all $x \in \mathbb{R}^d$, the conditional cumulative distribution function $F(y|X = x)$ is continuous.

Assumption 8.1 is necessary to get the uniform convergence of the estimator.

Assumption 8.2. For any tree $l \in [k]$, we assume that the variation of the conditional cumulative distribution function within any cell goes to 0.

$$\forall x \in \mathbb{R}^d, \forall y \in \mathbb{R}, \sup_{z \in A_n(\mathbf{x}; \Theta_l)} |F(y|z) - F(y|\mathbf{x})| \xrightarrow[n \rightarrow +\infty]{a.s.} 0$$

Assumption 8.2 allows to control the approximation error of the estimator. If for all y , $F(y|\cdot)$ is continuous, Assumption 8.2 is satisfied provided that the diameter of the cell goes to zero. Note that the vanishing of the diameter of the cell is a common condition used to prove the consistency of general partitioning estimator (see chapter 4 in [Györfi, 2002]). [Scornet, 2015] show that this is true when the data come from additive regression models [Stone, 1985b], and [Elie-Dit-Cosaque, 2022] show that it holds for a more general class, such as product functions or sums of product functions. This result is also valid for all regression functions, with a slightly modified version of RF, where each child node contains at least a small fraction of the observations in the parent node, and the probability that each variable $j = 1, \dots, p$ is chosen for the split is positive for every node. Under these small modifications, Lemma 2 from [Meinshausen, 2006] shows that the diameter of each leaf node vanishes.

Assumption 8.3. Let k the number of trees and $N_n(\mathbf{x}; \Theta_l)$ number of bootstrap observations in the leaf node where \mathbf{x} falls, and assume that $k = \mathcal{O}(n^\alpha)$ with $\alpha > 0$, and $\forall \mathbf{x} \in \mathbb{R}^d$, $N_n(\mathbf{x}; \Theta_l) = \Omega^1(\sqrt{n}(\ln(n))^\beta)$, with $\beta > 1$ a.s.

¹ $f(n) = \Omega(g(n)) \iff \exists c > 0, \exists n_0 > 0 \mid \forall n \geq n_0, |f(n)| \geq c|g(n)|$

Assumption 8.3 allows us to control the estimation error and means that the cells should contain a sufficiently large number of points so that averaging among the observations is effective.

To prove the consistency of the PRF CDF $\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S)$, we only need to verify the assumptions 8.1, 8.2, 8.3 on the parameters of the Projected Forest and the Projected CDF $F_S(y|\mathbf{X}_S = \mathbf{x}_S) = \mathbb{P}(Y \leq y|\mathbf{X}_S = \mathbf{x}_S)$.

Assumptions 8.1 and 8.2 are satisfied for the Projected CDF and the PRF Forest's leaves. Since by definition $A_n^{(S)}(\mathbf{x}_S; \Theta_l)$ is included in $A_n(\mathbf{x}; \Theta_l)$, if the diameter goes to zero within the cells of the RF, it also vanishes in the Projected Forest. In addition, if the CDF $F(y|\mathbf{X} = \mathbf{x}) = F(y|\mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}})$ is continuous, by analysis of parameter-dependent integral we have that the Projected CDF $F_S(y|\mathbf{X}_S = \mathbf{x}_S) = \int F(y|\mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}})\mathbb{P}(\mathbf{x}_{\bar{S}}|\mathbf{x}_S)d\mathbf{x}_{\bar{S}}$ is also continuous. As we control the minimal number of observations in each leaf of the Projected Forest by construction, Assumption 8.3 is also verified. Then, the PRF CDF satisfies also Assumption 8.1-8.3 which ensures its consistency thanks to Theorem 4.4.

8.2 Proof of Theorem 4.4

Theorem 8.4. *Consider a random forest which satisfies Assumptions 8.1 to 8.3. Then,*

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad \sup_{y \in \mathbb{R}} |\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S) - F_S(y|\mathbf{X}_S = \mathbf{x}_S)| \xrightarrow[n \rightarrow +\infty]{a.s.} 0 \quad (10)$$

To prove this Theorem, we essentially follow [Elie-Dit-Cosaque, 2022]. The idea is first to prove the result for a honest forest [Wager, 2017], then demonstrate that the original forest and the Honest Forest are close a.s.

Let us assume we have a honest forest [Wager, 2017], which is a random forest that grows using \mathcal{D}_n but uses another independent sample $\mathcal{D}_n^\diamond = \{(\mathbf{X}_i^\diamond, Y_i^\diamond)\}_{i=1}^n$ (independent of \mathcal{D}_n and Θ) to estimate the weights and predictions. From this, we extract a projected CDF honest forest defined as follows:

$$F_S^\diamond(y|\mathbf{X}_S = \mathbf{x}_S) = \sum_{i=1}^n w_{n,i}^\diamond(\mathbf{x}_S) \mathbb{1}_{Y_i^\diamond \leq y} \quad \text{where} \quad w_{n,i}^\diamond(\mathbf{x}_S) = \frac{1}{k} \sum_{l=1}^k \frac{\mathbb{1}_{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)}}{N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)},$$

and $N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)$ is the number of observation of $\{\mathbf{X}_i^\diamond\}_{i=1}^n$ that fall into $A_n^{(S)}(\mathbf{x}_S; \Theta_l)$. Note that $F_S^\diamond(y|\mathbf{X}_S = \mathbf{x}_S)$ depends on $\Theta_1, \dots, \Theta_k, \mathcal{D}_n, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_n^\diamond$, but we will not specify them to ease notations. Consequently, we have for all $\mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$,

$$\begin{aligned} |\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S) - F_S(y|\mathbf{X}_S = \mathbf{x}_S)| &\leq |\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S) - F_S^\diamond(y|\mathbf{X}_S = \mathbf{x}_S)| \\ &\quad + |F_S^\diamond(y|\mathbf{X}_S = \mathbf{x}_S) - F_S(y|\mathbf{X}_S = \mathbf{x}_S)|. \end{aligned}$$

The convergence of the two right-hand terms is handled separately into the following Proposition 8.5 and Lemma 8.6 respectively.

Proposition 8.5. *Consider a random forest which satisfies Assumptions 8.1 to 8.3. Then,*

$$\forall \mathbf{x} \in \mathbb{R}^d, \forall y \in \mathbb{R}, \quad F_S^\diamond(y|\mathbf{X}_S = \mathbf{x}_S) \xrightarrow[n \rightarrow +\infty]{a.s.} F_S(y|\mathbf{X}_S = \mathbf{x}_S) \quad (11)$$

Proposition 8.5 shows that the Projected CDF honest forest is consistent and Lemma 8.6 shows that the honest and the original forest are close.

Lemma 8.6. *Consider a random forest which satisfies Assumptions 8.1 to 8.3. Then,*

$$\forall \mathbf{x} \in \mathbb{R}^d, \forall y \in \mathbb{R}, \quad |F_S^\diamond(y|\mathbf{X}_S = \mathbf{x}_S) - \widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S)| \xrightarrow[n \rightarrow +\infty]{a.s.} 0 \quad (12)$$

Hence, according to Proposition 8.5 and Lemma 8.6, we get

$$\forall \mathbf{x} \in \mathbb{R}^d, \forall y \in \mathbb{R}, \quad \widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S) \xrightarrow[n \rightarrow +\infty]{a.s.} F_S(y|\mathbf{X}_S = \mathbf{x}_S) \quad (13)$$

We use Dini's second Theorem to have the almost sure uniform convergence relative to y of the Projected CDF honest forest. Indeed, $\{Y_i^\diamond \leq y\} = \{U_i \leq F_S(y|\mathbf{X}_S = \mathbf{X}_{i,S}^\diamond)\}$, where $U_i, i = 1, \dots, n$ are i.i.d uniform random variables. Let $s_i = F_S(y|\mathbf{X}_S = \mathbf{X}_{i,S}^\diamond)$ and $s = F_S(y|\mathbf{X}_S = \mathbf{x}_S)$, we have

$$\begin{aligned} \widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S) &= \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{U_i \leq s_i\}} \\ &= \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{\tilde{U}_i \leq s\}}, \end{aligned}$$

where $\tilde{U}_i \sim \mathcal{U}(s - s_i, s - s_i + 1), i = 1, \dots, n$ are independent uniform random variables. Then, Equation 13 is equivalent to:

$$\forall \mathbf{x} \in \mathbb{R}^d, \forall s \in [0, 1], \quad \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{\tilde{U}_i \leq s\}} \xrightarrow[n \rightarrow +\infty]{a.s.} s. \quad (14)$$

Equation (14) states that, $\forall s \in [0, 1], \exists N_s \subset \Omega, \mathbb{P}(N_s) = 0$ such that

$$\forall \omega \in N_s^c, \quad \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{\tilde{U}_i(\omega) \leq s\}} \xrightarrow[n \rightarrow +\infty]{} s. \quad (15)$$

$w_{n,i}(\mathbf{x}_S)$ is also random but we do not write $w_{n,i}(\mathbf{x}_S)(\omega)$ to lighten the notations. Hence, we need to find a set N that does not depend on s , which satisfies Equation (15) to get the uniform convergence with Dini's second Theorem. To that aim, we will use the density of \mathbb{Q} in \mathbb{R} as in the proof of the Glivenko-Cantelli Theorem.

Since the countable union of null sets is a null set, $\exists N \subset \Omega, \mathbb{P}(N) = 0$ such that

$$\forall s \in [0, 1] \cap \mathbb{Q}, \forall \omega \in N^c, \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{\tilde{U}_i(\omega) \leq s\}} \xrightarrow{n \rightarrow +\infty} s. \quad (16)$$

In fact, Equation (16) is also true for all $s \in [0, 1]$. let $s \in [0, 1], \epsilon > 0, w \in N^c, \exists p, q \in \mathbb{Q}$ such that $s - \epsilon \leq p \leq s \leq q \leq s + \epsilon$, since $s \rightarrow \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{\tilde{U}_i(\omega) \leq s\}}$ is increasing, we have:

$$\sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{\tilde{U}_i(\omega) \leq p\}} \leq \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{\tilde{U}_i(\omega) \leq s\}} \leq \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{\tilde{U}_i(\omega) \leq q\}}. \quad (17)$$

Thus,

$$s - \epsilon \leq \liminf \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{\tilde{U}_i(\omega) \leq s\}} \leq \limsup \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{\tilde{U}_i(\omega) \leq s\}} \leq s + \epsilon. \quad (18)$$

So we have shown that $\exists N \subset \Omega, \mathbb{P}(N) = 0, \forall \omega \in N^c$

- $s \rightarrow \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{\tilde{U}_i(\omega) \leq s\}}$ is increasing for all $n \in \mathbb{N}^*$
- $\forall s \in [0, 1], \sum_{i=1}^n w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{\tilde{U}_i(\omega) \leq s\}} \xrightarrow{n \rightarrow +\infty} s$ and $s \rightarrow s$ is continuous

Then the Dini's second Theorem states that we have the almost sure uniform convergence proving Theorem 4.4. Now, we turn to the proof of Proposition 8.5 and Lemma 8.6. To that aim, we need the following lemma based on Vapnik-Chervonenkis classes.

Lemma 8.7 ([Elie-Dit-Cosaque, 2022]). *Consider $\mathcal{D}_n, \mathcal{D}_n^\diamond$, two independent datasets of n i.i.d samples of $(\mathbf{X}, Y) \sim P = P_{\mathbf{X}} P_{Y|\mathbf{X}}$ and a tree build using \mathcal{D}_n with bootstrap and bagging procedure driven by Θ . As before, $N_n(\mathbf{x}_S; \Theta_l)$ is the number of bootstrap observations of \mathcal{D}_n that fall into $A_n^{(S)}(\mathbf{x}_S; \Theta_l)$ and $N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)$ is the number of observations of \mathcal{D}_n^\diamond that fall into $A_n^{(S)}(\mathbf{x}_S; \Theta_l)$. Then:*

$$\forall \epsilon > 0, \quad \mathbb{P} \left\{ \left| N_n(\mathbf{x}_S; \Theta_l) - N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l) \right| > \epsilon \right\} \leq 24(n+1)^{2|S|} e^{-\epsilon^2/288n}. \quad (19)$$

Proof.

$$\begin{aligned}
& \mathbb{P} \left\{ \left| N_n(\mathbf{x}_S; \Theta_l) - N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l) \right| > \epsilon \right\} \\
& \leq \mathbb{P} \left\{ \left| \frac{N_n(\mathbf{x}_S; \Theta_l)}{n} - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_{i,S} \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)} \right| > \frac{\epsilon}{3n} \right\} \\
& \quad + \mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_{i,S} \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)} - \mathbb{P}_{P_{\mathbf{X}}} \{ \mathbf{X}_S \in A_n^{(S)}(\mathbf{x}_S; \Theta_l) \} \right| > \frac{\epsilon}{3n} \right\} \\
& \quad + \mathbb{P} \left\{ \left| \frac{N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)}{n} - \mathbb{P}_{P_{\mathbf{X}}} \{ \mathbf{X}_S \in A_n^{(S)}(\mathbf{x}_S; \Theta_l) \} \right| > \frac{\epsilon}{3n} \right\} \\
& \leq \mathbb{P} \left\{ \sup_{A \in \mathcal{B}_S} \left| \frac{1}{n} \sum_{i=1}^n B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{\mathbf{X}_{i,S} \in A} - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_{i,S} \in A} \right| > \frac{\epsilon}{3n} \right\} \\
& \quad + \mathbb{P} \left\{ \sup_{A \in \mathcal{B}_S} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_{i,S} \in A} - \mathbb{P}_{P_{\mathbf{X}}} \{ \mathbf{X}_S \in A \} \right| > \frac{\epsilon}{3n} \right\} \\
& \quad + \mathbb{P} \left\{ \sup_{A \in \mathcal{B}_S} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_{i,S}^{\diamond} \in A} - \mathbb{P}_{P_{\mathbf{X}}} \{ \mathbf{X}_S \in A \} \right| > \frac{\epsilon}{3n} \right\},
\end{aligned}$$

where $\mathcal{B}_S = \left\{ \prod_{i \in S} [a_i, b_i] : a_i, b_i \in \overline{\mathbb{R}} \right\}$ is the set of hyperrectangles. The last two terms are handled thanks to Theorem 2 in [Vapnik, 1971] which bounds the difference between the frequencies of events to their probabilities over the entire class \mathcal{B}_S whose VC dimension is $2|S|$. Hence,

$$\begin{aligned}
& \mathbb{P} \left\{ \sup_{A \in \mathcal{B}_S} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_{i,S} \in A} - \mathbb{P}_{P_{\mathbf{X}}} \{ \mathbf{X}_S \in A \} \right| > \frac{\epsilon}{3n} \right\} + \mathbb{P} \left\{ \sup_{A \in \mathcal{B}_S} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_{i,S}^{\diamond} \in A} - \mathbb{P}_{P_{\mathbf{X}}} \{ \mathbf{X}_S \in A \} \right| > \frac{\epsilon}{3n} \right\} \\
& \leq 2\mathbb{P} \left\{ \sup_{A \in \mathcal{B}_S} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_{i,S}^{\diamond} \in A} - \mathbb{P}_{P_{\mathbf{X}}} \{ \mathbf{X}_S \in A \} \right| > \frac{\epsilon}{3n} \right\} \\
& \leq 16(n+1)^{2|S|} e^{-\epsilon^2/288n}.
\end{aligned}$$

The first term is also handled using Theorem 2 in [Vapnik, 1971], but conditionally on \mathcal{D}_n . Recall that $\{B_n(\mathbf{X}_i; \Theta_l)\}_{i=1}^n$ is a multinomial random variable $\mathcal{M}(n; \frac{1}{n}, \dots, \frac{1}{n})$ given \mathcal{D}_n , thus

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{\mathbf{X}_{i,S} \in A} \mid \mathcal{D}_n \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_{i,S} \in A}.$$

Therefore, we can apply Vapnik's Theorem 2 to control the first term conditionally on \mathcal{D}_n as following

$$\begin{aligned}
& \mathbb{P} \left\{ \sup_{A \in \mathcal{B}_S} \left| \frac{1}{n} \sum_{i=1}^n B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{\mathbf{X}_{i,S} \in A} - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_{i,S} \in A} \right| > \frac{\epsilon}{3n} \right\} \\
& \leq \mathbb{E} \left[\mathbb{P} \left\{ \sup_{A \in \mathcal{B}_S} \left| \frac{1}{n} \sum_{i=1}^n B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{\mathbf{X}_{i,S} \in A} - \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{\mathbf{X}_{i,S} \in A} \mid \mathcal{D}_n \right] \right| > \frac{\epsilon}{3n} \mid \mathcal{D}_n \right\} \right] \\
& \leq 8(n+1)^{2|S|} e^{-\epsilon^2/288n}.
\end{aligned}$$

□

Proof of proposition 8.5.

We want to show that:

$$\forall \mathbf{x} \in \mathbb{R}^d, \forall y \in \mathbb{R}, \quad F_S^\diamond(y|\mathbf{X}_S = \mathbf{x}_S) \xrightarrow[n \rightarrow +\infty]{a.s.} F_S(y|\mathbf{X}_S = \mathbf{x}_S)$$

For all $x \in \mathbb{R}^d, y \in \mathbb{R}$, we have:

$$\begin{aligned} |F_S^\diamond(y|\mathbf{X}_S = \mathbf{x}_S) - F_S(y|\mathbf{X}_S = \mathbf{x}_S)| &\leq \left| \sum_{i=1}^n w_{n,i}^\diamond(\mathbf{x}_S) \left(\mathbb{1}_{\{Y_i^\diamond \leq y\}} - F_S(y|\mathbf{X}_S = \mathbf{X}_{i,S}^\diamond) \right) \right| \\ &\quad + \left| \sum_{i=1}^n w_{n,i}^\diamond(\mathbf{x}_S) \left(F_S(y|\mathbf{X}_S = \mathbf{X}_{i,S}^\diamond) - F_S(y|\mathbf{X}_S = \mathbf{x}_S) \right) \right| \end{aligned}$$

We define $W_n = \sum_{i=1}^n w_{n,i}^\diamond(\mathbf{x}_S) \left(\mathbb{1}_{\{Y_i^\diamond \leq y\}} - F_S(y|\mathbf{X}_S = \mathbf{X}_{i,S}^\diamond) \right)$ and $V_n = \sum_{i=1}^n w_{n,i}^\diamond(\mathbf{x}_S) \left(F_S(y|\mathbf{X}_S = \mathbf{X}_{i,S}^\diamond) - F_S(y|\mathbf{X}_S = \mathbf{x}_S) \right)$ and treat each term separately.

Let us prove that $|W_n| \xrightarrow[n \rightarrow +\infty]{a.s.} 0$. We can rewrite W_n as $W_n = \sum_{i=1}^n w_{n,i}^\diamond(\mathbf{x}_S) H_i^\diamond$ where H_i^\diamond is bounded by 1 and $\mathbb{E}[H_i^\diamond | \mathbf{X}_{i,S}^\diamond] = 0$. Then,

$$\begin{aligned} \mathbb{P}(W_n > \epsilon) &\leq e^{-t\epsilon} \mathbb{E}[e^{tW_n}] \\ &\leq e^{-t\epsilon} \mathbb{E} \left[\prod_{i=1}^n \mathbb{E} \left[e^{tw_{n,i}^\diamond(\mathbf{x}_S) H_i^\diamond} \mid \Theta_1, \dots, \Theta_k, \mathcal{D}_n, \mathbf{X}_{1,S}^\diamond, \dots, \mathbf{X}_{n,S}^\diamond \right] \right] \\ &\leq e^{-t\epsilon} \mathbb{E} \left[\prod_{i=1}^n e^{\frac{t^2}{2} w_{n,i}^\diamond(\mathbf{x}_S)^2} \right] \end{aligned}$$

The last inequality comes from the fact that $w_{n,i}^\diamond(\mathbf{x}_S)$ is constant given $\Theta_1, \dots, \Theta_k, \mathcal{D}_n, \mathbf{X}_{1,S}^\diamond, \dots, \mathbf{X}_{n,S}^\diamond$, and as H_i^\diamond is bounded by 1 with $\mathbb{E}[H_i^\diamond | \mathbf{X}_{i,S}^\diamond] = 0$, we used the following inequality: If $|X| \leq 1$ a.s and $\mathbb{E}[X] = 0$, then $\mathbb{E}[e^{tX}] \leq e^{\frac{t^2}{2}}$. Indeed, by using the convexity of exponential, we have $\mathbb{E}[e^{tX}] \leq \mathbb{E} \left[\frac{1-X}{2} \right] e^{-t} + \mathbb{E} \left[\frac{1+X}{2} \right] e^t \leq \cosh(t) \leq e^{\frac{t^2}{2}}$.

Using Assumption 8.2, let $K > 0$ be such that for all $l \in [k]$, $N_n(\mathbf{x}_S; \Theta_l) \geq K\sqrt{n} \ln(n)^\beta$ a.s., then we have $\Gamma(l) = \left\{ N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l) < \frac{K\sqrt{n} \ln(n)^\beta}{2} \right\} \subset \left\{ |N_n(\mathbf{x}_S; \Theta_l) - N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)| > \frac{K\sqrt{n} \ln(n)^\beta}{2} \right\}$. Thus, using Lemma 8.7, we have that $\mathbb{P}(\Gamma(l)) \leq 24(n+1)^{2|S|} \exp\left(-\frac{-K^2(\ln(n))^{2\beta}}{1152}\right)$.

We have

$$\begin{aligned} \sum_{i=1}^n w_{n,i}^\diamond(\mathbf{x}_S)^2 &= \sum_{i=1}^n w_{n,i}^\diamond(\mathbf{x}_S) \times \frac{1}{k} \left(\sum_{l=1}^k \frac{\mathbb{1}_{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)}}{N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)} (\mathbb{1}_{\{\Gamma(l)\}} + \mathbb{1}_{\{\Gamma(l)^c\}}) \right) \\ &\leq \sum_{i=1}^n w_{n,i}^\diamond(\mathbf{x}_S) \left(\frac{2}{K\sqrt{n} \ln(n)^\beta} + \frac{1}{k} \sum_{l=1}^k \mathbb{1}_{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)} \mathbb{1}_{\{\Gamma(l)\}} \right) \end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}(W_n > \epsilon) &\leq \exp(-t\epsilon + \frac{t^2}{K\sqrt{n}\ln(n)^\beta}) \mathbb{E} \left(\exp \left(\frac{t^2}{2} \mathbb{1}_{\cup_{l=1}^k \Gamma(l)} \right) \right) \\
&\leq \exp(-t\epsilon + \frac{t^2}{K\sqrt{n}\ln(n)^\beta}) \times \left(1 + e^{\frac{t^2}{2}} \sum_{l=1}^k \mathbb{P}(\Gamma(l)) \right) \\
&\leq \exp(-t\epsilon + \frac{t^2}{K\sqrt{n}\ln(n)^\beta}) \times \left(1 + 24k(n+1)^{2|S|} \exp \left(\frac{t^2}{2} - \frac{K^2 \ln(n)^{2\beta}}{1152} \right) \right)
\end{aligned}$$

Taking $t^2 = \frac{K^2 \ln(n)^{2\beta}}{576}$ leads to

$$\mathbb{P}(W_n > \epsilon) \leq (1 + 24k(n+1)^{2|S|}) \exp \left(\frac{K \ln(n)^\beta}{576\sqrt{n}} - \frac{\epsilon K \ln(n)^\beta}{24} \right)$$

We obtain the same bound for $\mathbb{P}(W_n \leq -\epsilon) = \mathbb{P}(-W_n > \epsilon)$, then by using the assumption 8.2, that is, $k = \mathcal{O}(n^\alpha)$ so that the right term is finite, we conclude by Borel cantelli that $|W_n|$ goes to 0 a.s.

Lastly, we show that $|V_n| \xrightarrow[n \rightarrow +\infty]{a.s.} 0$.

$$\begin{aligned}
|V_n| &= \left| \sum_{i=1}^n w_{n,i}^\diamond(\mathbf{x}_S) \left(F_S(y|\mathbf{X}_S = \mathbf{X}_{i,S}^\diamond) - F_S(y|\mathbf{X}_S = \mathbf{x}_S) \right) \right| \\
&\leq \sum_{i=1}^n \frac{1}{k} \sum_{l=1}^k \frac{\mathbb{1}_{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)}}{N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)} \left| \left(F_S(y|\mathbf{X}_S = \mathbf{X}_{i,S}^\diamond) - F_S(y|\mathbf{X}_S = \mathbf{x}_S) \right) \right| \\
&\leq \frac{1}{k} \sum_{l=1}^k \left(\sum_{i=1}^n \frac{\mathbb{1}_{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)}}{N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)} \left| F_S(y|\mathbf{X}_S = \mathbf{X}_{i,S}^\diamond) - F_S(y|\mathbf{X}_S = \mathbf{x}_S) \right| \right) \\
&\leq \frac{1}{k} \sum_{l=1}^k \left(\sum_{i=1}^n \frac{\mathbb{1}_{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)}}{N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)} \sup_{\mathbf{z} \in A_n^{(S)}(\mathbf{x}; \Theta_l)} |F_S(y|\mathbf{X}_S = \mathbf{z}_S) - F_S(y|\mathbf{X}_S = \mathbf{x}_S)| \right) \\
&\leq \frac{1}{k} \sum_{l=1}^k \sup_{\mathbf{z} \in A_n^{(S)}(\mathbf{x}; \Theta_l)} |F_S(y|\mathbf{X}_S = \mathbf{z}_S) - F_S(y|\mathbf{X}_S = \mathbf{x}_S)|
\end{aligned}$$

Using assumption 8.1, that is, variation of the Projected CDF within the cell of the Projected tree vanishes, we conclude that $|V_n| \xrightarrow[n \rightarrow +\infty]{a.s.} 0$ ending the proof of Proposition 8.5.

Proof of Lemma 8.6. Here, we show that:

$$\forall \mathbf{x} \in \mathbb{R}^d, \forall y \in \mathbb{R}, \quad |F_S^\diamond(y|\mathbf{X}_S = \mathbf{x}_S) - \widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S)| \xrightarrow[n \rightarrow +\infty]{a.s.} 0.$$

We have

$$\begin{aligned}
& \left| F_S^\diamond(y|\mathbf{X}_S = \mathbf{x}_S) - \widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S) \right| = \left| \sum_{i=1}^n w_{n,i}^\diamond(\mathbf{x}_S) \mathbb{1}_{\{Y_i^\diamond \leq y\}} - w_{n,i}(\mathbf{x}_S) \mathbb{1}_{\{Y_i \leq y\}} \right| \\
& = \left| \sum_{i=1}^n \frac{1}{k} \sum_{l=1}^k \frac{\mathbb{1}_{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)} \mathbb{1}_{\{Y_i^\diamond \leq y\}}}{N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)} - \frac{1}{k} \sum_{l=1}^k \frac{B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{\mathbf{X}_i \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)} \mathbb{1}_{\{Y_i \leq y\}}}{N_n^{(S)}(\mathbf{x}_S; \Theta_l)} \right| \\
& = \left| \frac{1}{k} \sum_{l=1}^k \left(\frac{\sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)} \mathbb{1}_{\{Y_i^\diamond \leq y\}}}{N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)} - \frac{\sum_{i=1}^n B_n(\mathbf{X}_i; \Theta_l) \mathbb{1}_{\mathbf{X}_i \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)} \mathbb{1}_{\{Y_i \leq y\}}}{N_n^{(S)}(\mathbf{x}_S; \Theta_l)} \right) \right|.
\end{aligned}$$

As in [Arenal-Gutiérrez, 1996], we replace the bootstrap component with $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ where $\mathbf{Z}_i = (Z_{i,1}, Z_{i,2})$ is distributed as $\mathbf{Z} = (Z_1, Z_2)$ which follows uniform distribution over the set $\mathcal{D}_n = \{(\mathbf{X}^1, Y^1), \dots, (\mathbf{X}^n, Y^n)\}$ and $\mathbf{Z}_{i,1}, \mathbf{Z}_{i,2}$ corresponds to the input and output, respectively.

$$\begin{aligned}
& |F_S^\diamond(y|\mathbf{X}_S = \mathbf{x}_S) - \widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S)| \\
& = \left| \frac{1}{k} \sum_{l=1}^k \left(\frac{\sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Y_i^\diamond \leq y\}}}{N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)} - \frac{\sum_{i=1}^n \mathbb{1}_{\{\mathbf{Z}_{i,1} \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Z_{i,2} \leq y\}}}{N_n^{(S)}(\mathbf{x}_S; \Theta_l)} \right) \right| \\
& = \left| \frac{1}{k} \sum_{l=1}^k \left(\frac{\sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Y_i^\diamond \leq y\}}}{N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)} - \frac{\sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Y_i^\diamond \leq y\}}}{N_n^{(S)}(\mathbf{x}_S; \Theta_l)} \right) \right. \\
& \quad \left. + \left(\frac{\sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Y_i^\diamond \leq y\}}}{N_n^{(S)}(\mathbf{x}_S; \Theta_l)} - \frac{\sum_{i=1}^n \mathbb{1}_{\{\mathbf{Z}_{i,1} \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Z_{i,2} \leq y\}}}{N_n^{(S)}(\mathbf{x}_S; \Theta_l)} \right) \right| \\
& = \left| \frac{1}{k} \sum_{l=1}^k \frac{\sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Y_i^\diamond \leq y\}} (N_n^{(S)}(\mathbf{x}_S; \Theta_l) - N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l))}{N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l) \times N_n^{(S)}(\mathbf{x}_S; \Theta_l)} \right. \\
& \quad \left. + \frac{\sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Y_i^\diamond \leq y\}} - \sum_{i=1}^n \mathbb{1}_{\{\mathbf{Z}_{i,1} \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Z_{i,2} \leq y\}}}{N_n^{(S)}(\mathbf{x}_S; \Theta_l)} \right|
\end{aligned}$$

Using the fact that $\sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Y_i^\diamond \leq y\}} \leq N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l) = \sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l)\}}$, we have

$$\begin{aligned}
& |F_S^\diamond(y|\mathbf{X}_S = \mathbf{x}_S) - \widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S)| \\
& \leq \frac{1}{k} \sum_{l=1}^k \frac{|N_n^{(S)}(\mathbf{x}_S; \Theta_l) - N_n^{\diamond(S)}(\mathbf{x}_S; \Theta_l)|}{N_n^{(S)}(\mathbf{x}_S; \Theta_l)} \\
& \quad + \frac{\left| \sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i^\diamond \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Y_i^\diamond \leq y\}} - \sum_{i=1}^n \mathbb{1}_{\{\mathbf{Z}_{i,1} \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Z_{i,2} \leq y\}} \right|}{N_n^{(S)}(\mathbf{x}_S; \Theta_l)} \\
& \stackrel{\text{def}}{=} \frac{1}{k} \sum_{l=1}^k (|G_l^1| + |G_l^2|)
\end{aligned}$$

with

$$G_l^1 = \frac{N_n^{(S)}(\mathbf{x}_S; \Theta_l) - N_n^{\circ(S)}(\mathbf{x}_S; \Theta_l)}{N_n^{(S)}(\mathbf{x}_S; \Theta_l)} \quad \text{and}$$

$$G_l^2 = \frac{\sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i^{\circ} \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Y_i^{\circ} \leq y\}} - \sum_{i=1}^n \mathbb{1}_{\{\mathbf{Z}_{i,1} \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Z_{i,2} \leq y\}}}{N_n^{(S)}(\mathbf{x}_S; \Theta_l)}.$$

Therefore, Lemma 8.6 is equivalent to show that for all $l \in [k]$, $|G_l^1|, |G_l^2| \xrightarrow[n \rightarrow +\infty]{a.s.} 0$.

Using Assumption 8.2, $\exists K > 0, N_n^{(S)}(\mathbf{x}_S; \Theta_l) \geq K\sqrt{n} \ln(n)^\beta$ a.s., then

$$\begin{aligned} \mathbb{P}(|G_l^1| > \epsilon) &= \mathbb{P}\left(\left|N_n^{(S)}(\mathbf{x}_S; \Theta_l) - N_n^{\circ(S)}(\mathbf{x}_S; \Theta_l)\right| > \epsilon N_n^{(S)}(\mathbf{x}_S; \Theta_l)\right) \\ &\leq \mathbb{P}\left(\left|N_n^{(S)}(\mathbf{x}_S; \Theta_l) - N_n^{\circ(S)}(\mathbf{x}_S; \Theta_l)\right| > \epsilon K\sqrt{n} \ln(n)^\beta\right) \\ &\leq 24(n+1)^{2|S|} \exp\left(\frac{-K^2\epsilon^2 \ln n^{2\beta}}{288}\right) \quad (\text{By Lemma 8.7}). \end{aligned}$$

As the right hand is summable, we have $|G_l^1| \xrightarrow[n \rightarrow +\infty]{a.s.} 0$ by Borel-Cantelli. Now, we treat the term $|G_l^2|$. Let's define $\mathcal{B} = \left\{ \prod_{i=1}^{|S|} [a_i, b_i] \times [-\infty, y] : a_i, b_i \in \bar{\mathbb{R}} \right\}$, then for all $\epsilon > 0$,

$$\begin{aligned} \mathbb{P}(|G_l^2| > \epsilon) &= \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i^{\circ} \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Y_i^{\circ} \leq y\}} - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\mathbf{Z}_{i,1} \in A_n^{(S)}(\mathbf{x}_S; \Theta_l), Z_{i,2} \leq y\}}\right| > \frac{\epsilon N_n^{(S)}(\mathbf{x}_S; \Theta_l)}{n}\right) \\ &\leq \mathbb{P}\left(\sup_{A \in \mathcal{B}} \left|\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\mathbf{X}_i^{\circ}, Y_i^{\circ}\} \in A} - \mathbb{P}((\mathbf{X}, Y) \in A)\right| > \frac{\epsilon K\sqrt{n} \ln(n)^\beta}{3n}\right) \\ &+ \mathbb{P}\left(\sup_{A \in \mathcal{B}} \left|\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{(\mathbf{X}_i, Y_i) \in A\}} - \mathbb{P}((\mathbf{X}, Y) \in A)\right| > \frac{\epsilon K\sqrt{n} \ln(n)^\beta}{3n}\right) \\ &+ \mathbb{P}\left(\sup_{A \in \mathcal{B}} \left|\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{(\mathbf{Z}_{i,1}, Z_{i,2}) \in A\}} - \sum_{i=1}^n \mathbb{1}_{\{(\mathbf{X}_i, Y_i) \in A\}}\right| > \frac{\epsilon K\sqrt{n} \ln(n)^\beta}{3n}\right) \\ &\leq 2\mathbb{P}\left(\sup_{A \in \mathcal{B}} \left|\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{(\mathbf{X}_i, Y_i) \in A\}} - \mathbb{P}((\mathbf{X}, Y) \in A)\right| > \frac{\epsilon K\sqrt{n} \ln(n)^\beta}{3n}\right) \\ &+ \mathbb{P}\left(\sup_{A \in \mathcal{B}} \left|\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{(\mathbf{Z}_{i,1}, Z_{i,2}) \in A\}} - \sum_{i=1}^n \mathbb{1}_{\{(\mathbf{X}_i, Y_i) \in A\}}\right| > \frac{\epsilon K\sqrt{n} \ln(n)^\beta}{3n}\right) \end{aligned}$$

As above, the first term are handled thanks to a direct application of the Theorem 2 in [Vapnik, 1971] that bounds the difference between the frequencies of events to their probabilities over the entire class \mathcal{B} . As a result, we have

$$\mathbb{P}\left(\sup_{A \in \mathcal{B}} \left|\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{(\mathbf{X}_i, Y_i) \in A\}} - \mathbb{P}((\mathbf{X}, Y) \in A)\right| > \frac{\epsilon K\sqrt{n} \ln(n)^\beta}{3n}\right) \leq 8(n+1)^{2|S|+1} \exp\left(\frac{-K^2\epsilon^2 \ln(n)^{2\beta}}{288}\right)$$

To handle the last term, we apply the Theorem 2 in [Vapnik, 1971] under the conditional

distribution given \mathcal{D}_n ,

$$\begin{aligned}
& \mathbb{P} \left(\sup_{A \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{(Z_{i,1}, Z_{i,2}) \in A\}} - \sum_{i=1}^n \mathbb{1}_{\{(X_i, Y_i) \in A\}} \right| > \frac{\epsilon K \sqrt{n} \ln(n)^\beta}{3n} \right) \\
&= \mathbb{E} \left(\mathbb{P} \left(\sup_{A \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{(Z_{i,1}, Z_{i,2}) \in A\}} - \sum_{i=1}^n \mathbb{1}_{\{(X_i, Y_i) \in A\}} \right| > \frac{\epsilon K \sqrt{n} \ln(n)^\beta}{3n} \mid \mathcal{D}_n \right) \right) \\
&= \mathbb{E} \left(\mathbb{P} \left(\sup_{A \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{(Z_{i,1}, Z_{i,2}) \in A\}} - \mathbb{P}((\mathbf{Z}_1, Z_2) \in A \mid \mathcal{D}_n) \right| > \frac{\epsilon K \sqrt{n} \ln(n)^\beta}{3n} \mid \mathcal{D}_n \right) \right) \\
&\leq 8(n+1)^{2|S|+1} \exp\left(-\frac{K^2 \epsilon^2 \ln(n)^{2\beta}}{288}\right)
\end{aligned}$$

Finally, we get the overall upper bound,

$$\mathbb{P}(|G_l^2| > \epsilon) \leq 24(n+1)^{2|S|+1} \exp\left(\frac{-\epsilon^2 \ln(n)^{2\beta}}{288}\right)$$

By Borel-Cantelli, we conclude that $|G_l^2| \xrightarrow[n \rightarrow +\infty]{a.s.} 0$.

This concludes the proof of Lemma 8.6, thus the proof of Theorem 4.4.

9 Empirical evaluations of the estimator \widehat{F}_S

In order to compare the PRF CDF $\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S)$ and $F_S(y|\mathbf{X}_S = \mathbf{x}_S)$, we use a Monte Carlo approach to effectively compute $F_S(y|\mathbf{X}_S = \mathbf{x}_S)$. We use the synthetic dataset of Section 5: $\mathbf{X} \in \mathbb{R}^p$, $\mathbf{X} \in \mathcal{N}(0, \Sigma)$, $\Sigma = 0.8J_p + 5I_p$ with $p = 100$, I_p is the identity matrix, J_p is all-ones matrix and a piece-wise linear predictor defined as:

$$Y = (X_1 + X_2)\mathbb{1}_{X_5 \leq 0} + (X_3 + X_4)\mathbb{1}_{X_5 > 0}. \tag{20}$$

The variables X_i for $i = 6 \dots 100$ are noise variables. We fit a RF with a sample size $n = 10^4$, $k = 20$ trees and the minimal number of samples by leaf node is set to $t_n = \lfloor \sqrt{n} \times \ln(n)^{1.5} / 250 \rfloor$ for the original and the Projected Forest.

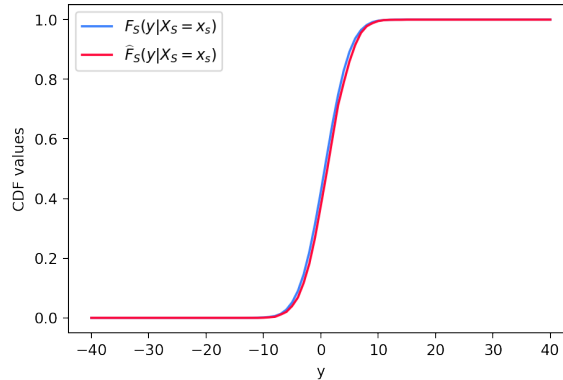


Figure 4: Comparison of $\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_S)$ and $F_S(y|\mathbf{X}_S = \mathbf{x}_S)$ with $S = [1, 2, 5]$ and $\mathbf{x}_S = [-0.13, 1.29, -1.31]$

We chose a randomly chosen point $\mathbf{x}_S = [-0.13, 1.29, -1.31]$ with $S = [1, 2, 5]$ from the test set. The experiment is replicated 100 times. Figure 4 shows that the estimator works well for almost all points $y \in \mathbb{R}$.

We also compute two global metrics. For a given S , we compute the average Kolmogorov-Smirnov $MKS = \frac{1}{n} \sum_{i=1}^n \sup_{y \in \mathbb{R}} |\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_{S,i}) - F_S(y|\mathbf{X}_S = \mathbf{x}_{S,i})|$ and the average mean absolute deviation $MAD = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} |\widehat{F}_S(y|\mathbf{X}_S = \mathbf{x}_{S,i}) - F_S(y|\mathbf{X}_S = \mathbf{x}_{S,i})| dy$.

We have $MAD = 0.008$ and the $MKS=0.26$ on all the observations with $S = [1, 2, 3, 5]$ showing the estimator’s efficiency. We also compute them with small $S = [0, 4]$, it works even better with $MAD=0.068$, $MKS=0.0098$.

10 Additional experiments

10.1 Local rules of Anchors and Sufficient Rules with ground truth explanations

In this section, we compare Anchors and Sufficient Rules using a synthetic dataset with strong dependencies between the important features. In this case, we can evaluate their capacity of providing the ground truth minimal rules since we know the distribution of the data. We use the moon dataset $(X_1, X_2, Y) \in \mathbb{R}^2 \times \{0, 1\}$, see Figure 5, and we add gaussian features $\mathbf{Z} \in \mathbb{R}^{100}$ with the μ, Σ of the previous section so that the final data is $(X_1, X_2, \mathbf{Z}, Y)$. Also, if $Z_1 > 0$, we flip the label Y of the observations.

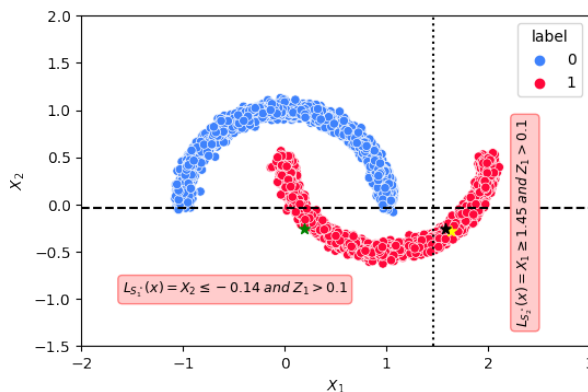


Figure 5: Explanations of $\mathbf{x}, \tilde{\mathbf{x}}$ by the two Sufficient Rules, the horizontal/vertical rectangle is associate with $S_1^* = [x_1, z_1], S_2^* = [x_2, z_1]$ respectively. The background samples are the observations with $z_1 > 0$.

We train a RF with the parameters of the previous section. It has $AUC=99\%$ on the test set of 10^4 observations. We use Anchors with threshold $\tau = 0.95$, tolerance $\delta = 0.05$, and the Minimal Sufficient Rules with $\pi = 0.95$ to explain 1000 observations of the test set. We observe that on average Anchors tend to give much longer rules. The mean size for Sufficient Rules is 2, and for Anchors it is 10. In addition, the Minimal Sufficient Explanations detect local relevant variables more accurately. It has $FDR=3\%$, $TDR=100\%$ and Anchors has $FDR=48\%$,

TDR=80%. Finally, we qualitatively observe the rules on a given example \mathbf{x} (black star in Figure 5). We also assess the stability of the explanations by comparing the rules of \mathbf{x} and $\tilde{\mathbf{x}}$ a nearby observation such that $\max_{i \in \{1,2\}} |x_i - \tilde{x}_i| \leq 0.05$ (yellow star in Figure 5). The rules given by Anchors for $\mathbf{x}, \tilde{\mathbf{x}}$ are $L_{\text{Anchors}}(\mathbf{x}) = \{X_1 > -0.03 \text{ AND } Z_1 > 0.01 \text{ AND } Z_9 > -1.66 \text{ AND } Z_{44} > 1.66 \text{ AND } Z_{32} \leq -1.57\}$ and $L_{\text{Anchors}}(\tilde{\mathbf{x}}) = \{X_1 > 1.04 \text{ AND } X_2 \leq -0.20 \text{ AND } Z_1 > 0.01 \text{ AND } Z_{28} > 0.01 \text{ AND } Z_{45} \leq -1.57\}$. The rules given by Anchors are very different, showing instability. Moreover, we also note that Anchors is very sensitive to random seed.

In contrast, the SDP approach gives the same explanations for $\mathbf{x}, \tilde{\mathbf{x}}$. The observations have two Minimal Sufficient Explanations $S_1^* = [X_1, Z_1], S_2^* = [X_2, Z_1]$. These explanations lead to two Sufficient Rules, which, when visualized along the X_1 and X_2 axes (as illustrated in Figure 5), effectively elucidated the model’s predictions. Nevertheless, the vertical rule could be slightly more to the left. We think this imprecision comes from the estimation of the RF, which is not perfect.

As these observations have multiple explanations, we provide additional insight about the important variables by computing their Local eXplanatory Importances (LXI). The LXI of $\mathbf{x}, \tilde{\mathbf{x}}$ are $[\mathbf{x}_1 = 0.5, \mathbf{x}_2 = 0.5, \mathbf{z}_1 = 1, \mathbf{z}_2 = 0, \dots, \mathbf{z}_{100} = 0]$. It shows that the variables $\{Z_i\}_{i \in [2,100]}$ are irrelevant for these observations. The relevant variables are X_1, X_2, Z_1 and especially Z_1 is the most important. It is a necessary feature as it appears in every Sufficient Explanations.

Comparison of Shapley Values and LXI on Moon Data: In Figure 6 - 7, we compare the Shapley Values and LXI of an observation with $Z_1 > 0$ (the green star). We observe that the LXI gives non-null values only on the active variable (X_2, Z_1), while the SV gives non-null values also on noise variables. Moreover, SV gives a non-negligible value to the feature X_1 that is not important for this prediction. Indeed, by analyzing Figure 5, we observe that whatever the value of X_1 , if we fix X_2 and the sign of Z_1 , the prediction will not change.

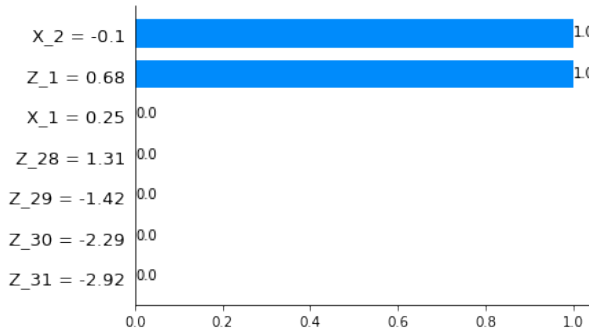


Figure 6: LXI of the green star

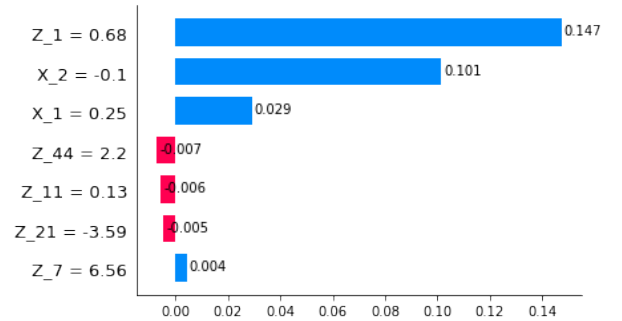


Figure 7: Shapley Values of the green star

We also compute the mean importance score across the population in Figure 8 - 9. For the SV, we take the mean absolute values as it may have negative contributions. The three important values that come out for both methods are X_1, X_2, Z_1 . However, as in the local case, SV assign values to the noise variables.

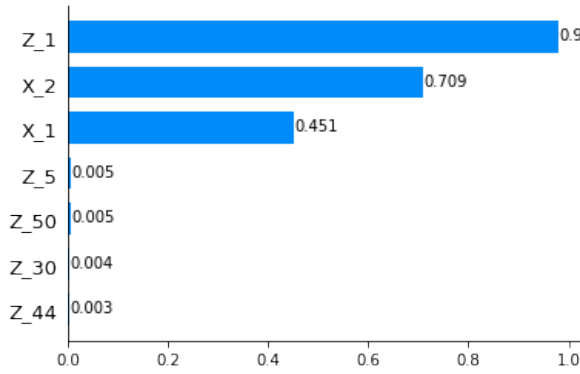


Figure 8: Mean LXI

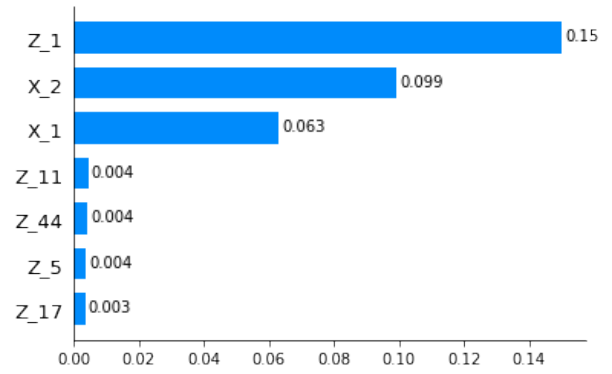
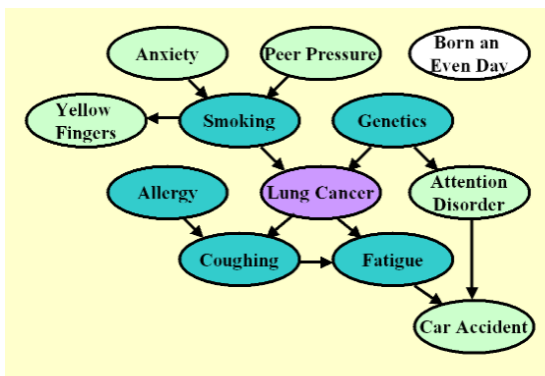


Figure 9: Mean absolute SHAP

10.2 Shapley Values and Local eXplanatory importance (LXI) on LUCAS dataset

In this section, we want to highlight a case where the LXI permit to drastically simplify the diversity of the possible explanations. We use a semi-synthetic dataset **LUCAS** (Lung Cancer Simple set), a dataset generated by causal Bayesian networks with 12 binary variables. The causal graph is drawn in Figure 10a and the probability table in Figure 10b .



(a)

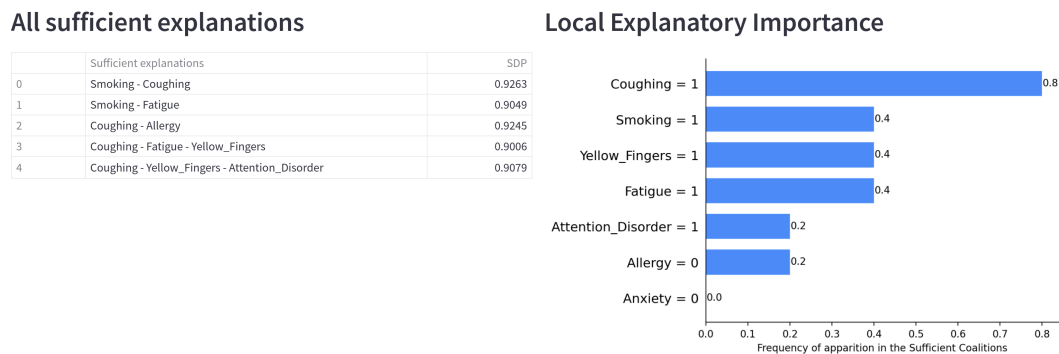
```

P(Anxiety=T)=0.64277
P(Peer Pressure=T)=0.32997
P(Smoking=T|Peer Pressure=F, Anxiety=F)=0.43118
P(Smoking=T|Peer Pressure=T, Anxiety=F)=0.74591
P(Smoking=T|Peer Pressure=F, Anxiety=T)=0.8686
P(Smoking=T|Peer Pressure=T, Anxiety=T)=0.91576
P(Yellow Fingers=T|Smoking=F)=0.23119
P(Yellow Fingers=T|Smoking=T)=0.95372
P(Genetics=T)=0.15953
P(Lung cancer=T|Genetics=F, Smoking=F)=0.23146
P(Lung cancer=T|Genetics=T, Smoking=F)=0.86996
P(Lung cancer=T|Genetics=F, Smoking=T)=0.83934
P(Lung cancer=T|Genetics=T, Smoking=T)=0.99351
P(Attention Disorder=T|Genetics=F)=0.28956
P(Attention Disorder=T|Genetics=T)=0.68706
P(Born an Even Day=T)=0.5
P(Allergy=T)=0.32841
P(Coughing=T|Allergy=F, Lung cancer=F)=0.1347
P(Coughing=T|Allergy=T, Lung cancer=F)=0.64592
P(Coughing=T|Allergy=F, Lung cancer=T)=0.7664
P(Coughing=T|Allergy=T, Lung cancer=T)=0.99947
P(Fatigue=T|Lung cancer=F, Coughing=F)=0.35212
P(Fatigue=T|Lung cancer=T, Coughing=F)=0.56514
P(Fatigue=T|Lung cancer=F, Coughing=T)=0.80016
P(Fatigue=T|Lung cancer=T, Coughing=T)=0.89589
P(Car Accident=T|Attention Disorder=F, Fatigue=F)=0.2274
P(Car Accident=T|Attention Disorder=T, Fatigue=F)=0.779
P(Car Accident=T|Attention Disorder=F, Fatigue=T)=0.78861
P(Car Accident=T|Attention Disorder=T, Fatigue=T)=0.97169
  
```

(b)

Figure 10: (a): Bayesian network that represents the causal relationships between variables and (b): Probabilities table used to generate the data.

In Figure 11, we observe the different explanations of an observation chosen randomly, its features values are {**Smoking = True**, **Yellow Fingers = True**, **Anxiety = False**, **Peer Pressure = False**, **Genetic = False**, **Attention Disorder = True**, **Born an Even Day = False**, **Car Accident = True**, **Fatigue = True**, **Allergy = False**, **Coughing = True**} and its label is **True**. We see in the left of Figure 11 that it has many Sufficient Explanations. Therefore, as seen in the right of Figure 11, the LXI permit to synthesize all the different explanations in a single feature contributions that exhibits the local importance of the variables. Each value corresponds to the frequency of apparition of the corresponding feature in the set of all the sufficient explanations.



Feature values highlight by SDP

This observation has 5 different explanations, below to observe their values

Change the explanations

0	Smoking	Yellow_Fingers	Anxiety	Peer_Pressure	Genetics	Attention_Disorder	Born_an_Even_Day	Car_Accident	Fatigue	Allergy	Coughing	Output	SDP
0	1	1	0	0	0	1	0	1	1	0	1	1	0.9263

Local rule explanation

$0.5 \leq \text{Smoking} \leq \text{inf}$ and $0.5 \leq \text{Coughing} \leq \text{inf}$

Figure 11: Screenshot of a web-app developed in <https://github.com/salimamoukou/acv00> showing the Sufficient Explanations (upper left) and LXI (upper right) of an observation chosen randomly

At the bottom of Figure 11, we observe the rule associated with the first Sufficient Explanation which is {**Smoking = True**, **Coughing = True**}. Note that this rule is very powerful as it has a coverage of 46% and an accuracy of 93%.

Comparison of Shapley Values and LXI on LUCAS: We can also compare the Shapley Values and LXI on this data set. In Figure 12, we observe that it is the value of Coughing that is really important for this observation. Indeed, it appears in several Sufficient Explanations (80%). On the other side, in Figure 13, SHAP associates values to many more variables, and it has difficulties to discriminate between the important values. It is difficult to deduce from the values of Smoking, Coughing, Fatigue, Allergy which is the most important variable with Shapley Values.

All sufficient explanations

Sufficient explanations	SDP
0 Smoking - Coughing	0.9387
1 Smoking - Fatigue	0.9078
2 Coughing - Yellow_Fingers	0.9225
3 Coughing - Anxiety	0.9036
4 Coughing - Fatigue - Peer_Pressure	0.9034

Local Explanatory Importance

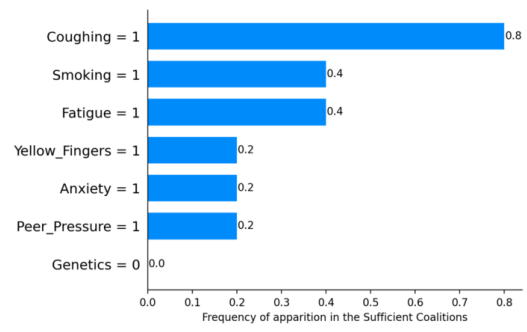


Figure 12: Sufficient Explanations and Local Explanatory Importance

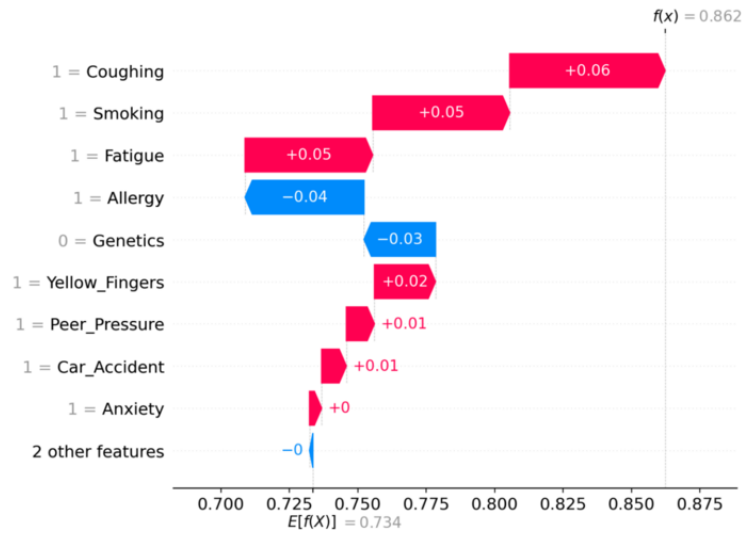


Figure 13: Shapley Values

10.3 Stability of Anchors and Sufficient Rules

Here, we run the last experiment of Section 5 on the stability of the local rules (Anchors, Sufficient Rules). The objective was to evaluate how these methods handle input perturbations. To do this, we compared the rules generated by each method against the rules derived from 50 noisy versions of a given observation \mathbf{x} . The noise was introduced by adding random Gaussian perturbations $\mathcal{N}(0, \epsilon \times I)$ to the feature values, with two different values of $\epsilon = 0.01, 0.001$.

Figure 15 illustrates the distribution of the number of different rules obtained from this experiment. We observed that Anchors tend to produce a larger variety of rules when the observation is slightly modified, whereas Sufficient Rules exhibit higher stability.

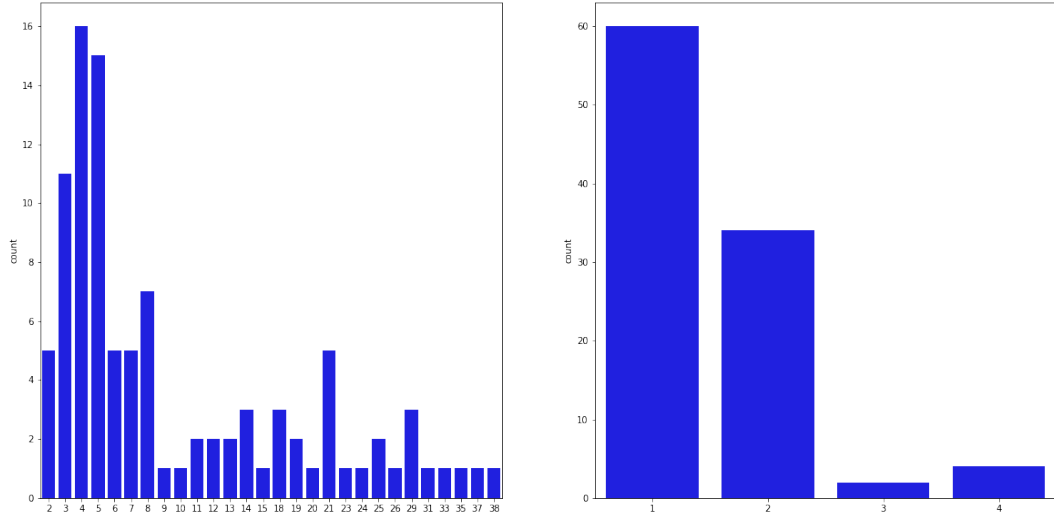


Figure 14: Number of different rules of Anchors (left) and Sufficient Rules (right) when $\epsilon = 0.001$

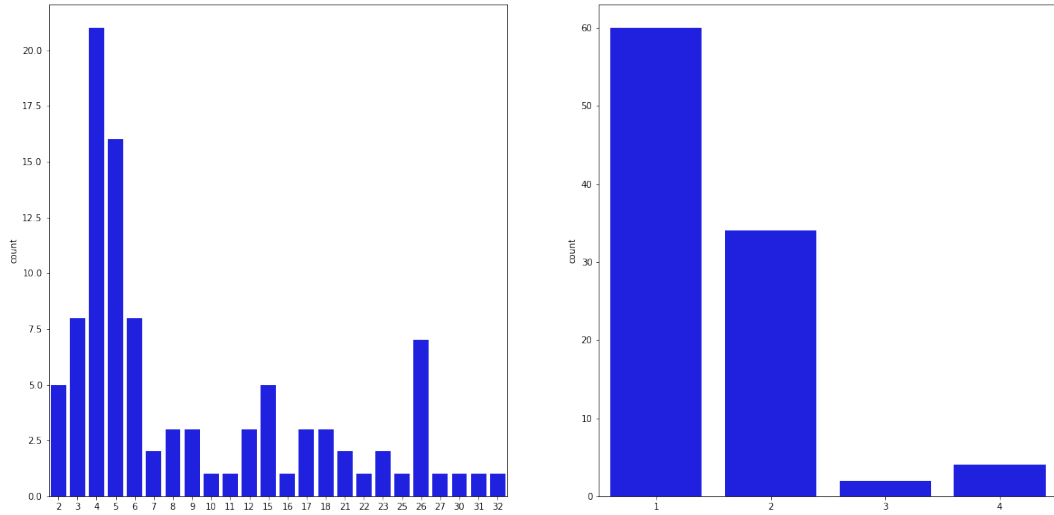


Figure 15: Number of different rules of Anchors (left) and Sufficient Rules (right) when $\epsilon = 0.01$

11 From Sufficient Rules to Global Interpretable model

In this section, we investigate the capacity to transform the Sufficient Rules explanations into a global competitive model. Indeed, we can build a global model by combining all the Sufficient Rules found for the observations in the training set. We set the output of each rule as the majority class (resp. average values) for classification (resp. regression) of the training observations that satisfy this rule. Note that some rules can overlap, and an observation can satisfy multiple rules. To resolve these conflicts, we use the output of the rule with the best precision such as accuracy or R^2 for classification or regression respectively. We have experimented on 2 real-world datasets: Diabetes [Kaggle, 2016] contains diagnostic measurements and aims to predict whether or not a patient has diabetes, Breast Cancer Wisconsin (BCW) [Dua, 2017a] consists of predicting if a tumor is benign or not using the characteristic of the cell nuclei. Thus, we perform comparisons between the global model induced by the Sufficient Rules (G-SR) and SOTA global rule-based

models as baseline. We use the package `imodel` [Singh, 2021] for RuleFit, Skoped Rule (SkR) and `scikit-learn` [Pedregosa, 2011] for Decision Tree (DT), and Random Forest (RF). In table 2, we observe that the G-SR performs as well as the best baseline models while being transparent in its decision process. These experiments increase the trustworthiness of our explanations because we derive an interpretable (by-design) global model without paying a high trade-off with performance. As a by-product, SR can be used as a new way of building glass-box models, but this line of research is beyond the scope of the current work.

Table 2: Accuracy of the different models on Diabetes and Breast Cancer Wisconsin dataset (BCW). For G-SR, we add the coverage of the model on the test-set in brackets.

DATA SET	G-SR	RULEFIT	SKR	DT	RF
DIABETES	0.98 (81%)	0.76	0.71	0.90	0.92
BCW	0.95 (92%)	0.95	0.93	0.95	0.96

12 Projected Forest CDF algorithm

Algorithm 5: Projected Forest CDF: \hat{F}_S

Input : A random forest fit with \mathcal{D}_n , a query point $\mathbf{x}_S, y, \text{min_nodes_size}$

Output: $\hat{F}(y|\mathbf{X}_S = \mathbf{x}_S)$

- 1: **for** all trees in the forest **do**
- 2: initialize *nodes_level* as a list of nodes containing only the root node;
- 3: initialize *nodes_child* as an empty list of child nodes;
- 4: initialize *samples* as the list of observation indices of the full training data of the tree;
- 5: **for** all levels in the tree **do**
- 6: **for** all nodes in *nodes_level*: **do**
- 7: **if** the node splits on a variable in S **then**
- 8: compute whether \mathbf{x}_S falls in the left or right child node;
- 9: append the child node to *nodes_child*;
- 10: set *samples_child* as the observations in *samples* which satisfy the split;
- 11: **else**
- 12: append both the left and right children nodes to *nodes_child*;
- 13: set *samples* = *samples_child*;
- 14: **if** the size of *samples_child* is lower than *min_node_size* **then**
- 15: break the loop through the tree levels;
- 16: **else**
- 17: set *samples* = *samples_child*;
- 18: set *nodes_level* = *nodes_child*;
- 19: compute the tree prediction as the average of $\mathbb{1}_{Y_i \leq y}$ for all i in *samples*;
- 20: **return** average the prediction of all trees;

Appendix for Chapter 5

13 Regional RF detailed

In this section, we give a simple application of the Regional RF algorithm to better understand how it works. Recall that the Regional RF is a generalization of the RF's algorithm to give prediction even when we condition given a region, e.g., to estimate $E(f(\mathbf{X}) | \mathbf{X}_S \in C_S(\mathbf{x}), \mathbf{X}_{\bar{S}} = \mathbf{x}_{\bar{S}})$ with $C_S(\mathbf{x}) = \prod_{i=1}^{|\bar{S}|} [a_i, b_i]$, $a_i, b_i \in \bar{\mathbb{R}}$ a hyperrectangle. The algorithm works as follows: we drop the observations in the initial trees, if a split used variable $i \in \bar{S}$, a fixed value-based condition, we used the classic rules, i.e., if $x_i \leq t$, the observations go to the left children, otherwise the right children. However, if a split used variable $i \in S$, regional-based condition, we used the hyperrectangle $C_S(\mathbf{x}) = \prod_{i=1}^{|\bar{S}|} [a_i, b_i]$. The observations are sent to the left children if $b_i \leq t$, right children if $a_i > t$ and if $t \in [a_i, b_i]$ the observations are sent both to the left and right children.

To illustrate how it works, we use a two dimensional variables $\mathbf{X} = (X_0, X_1) \in \mathbb{R}^2$, a simple decision tree f represented in Figure 16, and want to compute for $\mathbf{x} = [1.5, 1.9]$, $\mathbb{E}(f(\mathbf{X}) | X_1 \in [2, 3.5], X_0 = 1.5)$. We assume that $\mathbb{P}(X_1 \in [2, 3.5] | X_0 = 1.5) > 0$ and denoted T_1 as the set of the values of the splits based on variables X_1 of the decision tree. One way of estimating this conditional mean is by using Monte Carlo sampling. Therefore, there are two cases :

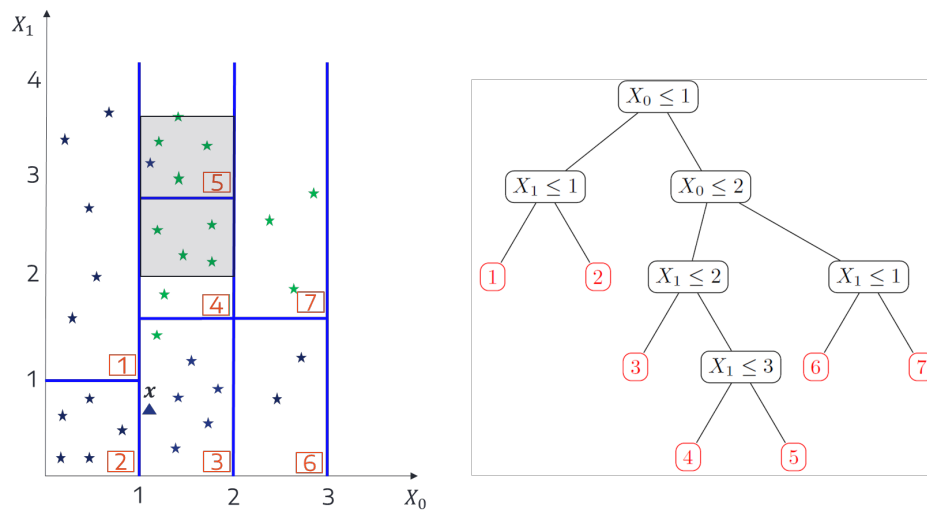


Figure 16: Representation of a simple decision tree (right Figure) and its associated partition (left Figure). The gray part in the partition corresponds to the region $[2, 3.5] \times [1, 2]$

-
- If $\forall t \in T_1, t \leq 2$ or $t > 3.5$, then all the observations sampled s.t. $\tilde{X}_i \sim P_{\mathbf{X}} |_{X_1 \in [2, 3.5], X_0 = 1.5}$ follow the same path and fall in the same leaf. The Monte Carlo estimator of the decision tree $\mathbb{E}(f(\mathbf{X}) | X_1 \in [2, 3.5], X_0 = 1.5)$ is equal to the output of the Regional RF algorithm.
 - For instance, a special case of the case above is: if $\forall t \in T_1, t \leq 2$, and we sample using $P_{\mathbf{X}} |_{X_1 \in [2, 3.5], X_0 = 1.5}$, then all the observations go to the right children when they encounter a node using X_1 and fall in the same leaf.
 - If $\exists t \in T_1$ and $t \in [2, 3.5]$, then the observations sampled s.t. $\tilde{X}_i \sim P_{\mathbf{X}} |_{X_1 \in [2, 3.5], X_0 = 1.5}$ can fall in multiple terminal leaf depending on if their coordinates x_1 is lower than t . Following our example, if we generate samples using $P_{\mathbf{X}} |_{X_1 \in [2, 3.5], X_0 = 1.5}$, the observations will fall in the gray region of Figure 16, and thus can fall in node 4 or 5. Therefore, the true estimate is:

$$\begin{aligned} \mathbb{E}(f(\mathbf{X}) | X_1 \in [2, 3.5], X_0 = 1.5) &= \mathbb{P}(X_1 \leq 2.9 | X_0 = 1.5) \times \mathbb{E}[f(\mathbf{X}) | \mathbf{X} \in L_4] \\ &+ \mathbb{P}(X_1 > 2.9 | X_0 = 1.5) \times \mathbb{E}[f(\mathbf{X}) | \mathbf{X} \in L_5] \quad (21) \end{aligned}$$

Concerning the last case ($t \in [2, 3.5]$), we need to estimate the different probabilities $\mathbb{P}(X_1 \leq 2.9 | X_0 = 1.5), \mathbb{P}(X_1 > 2.9 | X_0 = 1.5)$ to compute $\mathbb{E}(f(\mathbf{X}) | X_1 \in [2, 3.5], X_0 = 1.5)$, but these probabilities are difficult to estimate in practice. However, we argue that we can ignore these splits, and thus do not need to fragment the query region using the leaves of the tree. Indeed, as we are no longer interested in a point estimate but regional (population mean) we do not need to go to the level of the leaves. We propose to ignore the splits of the leaves that divide the query region. For instance, the leaves 4 and 5 split the region $[2, 3.5]$ in two cells, by ignoring these splits we estimate the mean of the gray region by taking the average output of the leaves 4 and 5 instead of computing the mean weighted by the probabilities as in Equation (21). Roughly, it consists to follow the classic rules of a decision tree (if the region is above or below a split) and ignore the splits that are in the query region, i.e., we average the output of all the leaves that are compatible with the condition $X_1 \in [2, 3.5], X_0 = 1.5$. We think it leads to a better estimation for two reasons. First, we observe that the case where t is in the region and thus divides the query region does not occur often. Moreover, the leaves of the trees are very small in practice, and taking the mean of observations that fall into the union of leaves that belong to the query region is more reasonable than computing the weighted mean and thus trying to estimate the different probabilities $\mathbb{P}(X_1 \leq 2.9 | X_0 = 1.5), \mathbb{P}(X_1 > 2.9 | X_0 = 1.5)$.

14 Additional experiments

In table 3, we compare the *Accuracy* (Acc), *Plausibility* (Psb), and *Sparsity* (Sprs) of the different methods on additional real-world datasets: FICO [FICO, 2018], NHANESI [CDC, 1999-2022].

We observe that the L-CR, and R-CR outperform the baseline methods by a large margin on *Accuracy* and *Plausibility*. The baseline methods still struggle to change at the same time the positive and negative class. AReS and CET give better sparsity, but their counterfactual samples

are less plausible than the ones generated by the CR.

Table 3: Results of the *Accuracy* (Acc), *Plausibility*, and *Sparsity* (Sprs) of the different methods. We compute each metric according to the positive (Pos) and negative (Neg) class.

	FICO						NHANESI					
	Acc		Psb		Sps		Acc		Psb		Sps	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
L-CR	0.98	0.94	0.98	0.99	5	5	0.99	0.98	0.98	0.97	5	6
R-CR	0.90	0.94	0.98	0.99	9	8.43	0.86	0.95	0.96	0.99	7	7
AReS	0.34	0.01	0.85	0.86	2	1	0.06	1	0.87	0.92	1	1
CET	0.76	0	0.76	0.60	2	2	0	0.40	0.82	0.56	0	5

15 Parameters detailed

In this section, we give the different parameters of each method. For all methods and datasets, we first used a greedy search given a set of parameters. For AReS, we use the following set of parameters:

- max rule = {4, 6, 8}, max rule length = {4, 8}, max change num = {2, 4, 6},
- minimal support = 0.05, discretization bins = {10, 20},
- $\lambda_{acc} = \lambda_{cov} = \lambda_{cst} = 1$.

For CET, we search in the following set of parameters:

- max iterations = {500, 1000},
- max leaf size = {4, 6, 8, -1},
- $\lambda = 0.01, \gamma = 1$.

Lastly, for the Counterfactual Rules, we used the following parameters:

- nb estimators = {20, 50}, max depth= {8, 10, 12},
- $\pi = 0.9, \pi_C = 0.9$.

We obtained the same optimal parameters for all datasets:

- AReS: max rule = 4, max rule length= 4, max change num = 4, minimal support = 0.05, discretization bins = 10, $\lambda_{acc} = \lambda_{cov} = \lambda_{cst} = 1$
- CET: max iterations = 1000, max leaf size = -1, $\lambda = 0.01, \gamma = 1$
- CR: nb estimators= 20, max depth= 10, $\pi = 0.9, \pi_C = 0.9$

Appendix for Chapter 6

16 Proof of Lemma 4.1

The Lemma 4.1, which is the cornerstone of the LCP framework, shows how to achieve marginal coverage by properly selecting the level $\tilde{\alpha}$ of the quantile of the localizer.

Lemma 16.1. *Let $\tilde{\alpha}$ be the smallest value in $\Gamma = \left\{ \sum_{j=1}^k w_n(\mathbf{X}_i, \mathbf{X}_j) : i, k \in [n+1] \right\}$ such that*

$$\sum_{i=1}^{n+1} \frac{1}{n+1} \mathbf{1}_{\widehat{V}_i \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_i)} \geq 1 - \alpha, \quad (22)$$

then $\mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}) \right\} \geq 1 - \alpha$, or equivalently $\mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}^\infty) \right\} \geq 1 - \alpha$.

It is important to keep in mind that both $\tilde{\alpha}$ and \mathcal{F}_{n+1} depends on $\widehat{\mathcal{D}}_n = \left\{ \widehat{Z}_1, \dots, \widehat{Z}_n \right\}$ and $(\mathbf{X}_{n+1}, \widehat{V}_{n+1})$ where $\widehat{Z}_i = (\mathbf{X}_i, \widehat{V}_i)$, but we will not specify them for ease of reading.

Proof. Let define the event $E_{n+1} = \left\{ \widehat{Z}_1 = \widehat{z}_1, \dots, \widehat{Z}_{n+1} = \widehat{z}_{n+1} \right\}$ where $\widehat{Z}_i = (\mathbf{X}_i, \widehat{V}_i)$ and $\widehat{z}_i = (\mathbf{x}_i, \widehat{v}_i) \in \mathcal{X} \times \mathbb{R}$. The exchangeability of the residuals implies that $\widehat{V}_{n+1} | E_{n+1}$ is uniform on the set $\{\widehat{v}_1, \dots, \widehat{v}_{n+1}\}$, and

$$\begin{aligned} \mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}) \mid E_{n+1} \right\} &= \sum_{i=1}^{n+1} \mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} = \widehat{v}_i \mid E_{n+1} \right\} \mathbf{1}_{\widehat{v}_i \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_i)} \\ &= \sum_{i=1}^{n+1} \frac{1}{n+1} \mathbf{1}_{\widehat{v}_i \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_i)} \geq 1 - \alpha. \quad (\text{by Eq. 22}) \end{aligned}$$

The formulation $\widehat{V}_{n+1} | E_{n+1}$ aims to provide another way to represent the uniformity of ranks when variables are exchangeable. It corresponds to a scenario where we had observed an unordered set of variables $E_{n+1} = \left\{ \widehat{Z}_1 = \widehat{z}_1, \dots, \widehat{Z}_{n+1} = \widehat{z}_{n+1} \right\}$ and have forgotten which value v_i each random variable V_j is associated with. By leveraging the uniformity of ranks of exchangeable random variables (see Chapter 1, Lemma 2.2), we establish that $P(V_j = v_i | E_{n+1}) = \frac{1}{n+1}$.

By marginalizing over the event E_{n+1} , we have $\mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}) \right\} \geq 1 - \alpha$. Additionally, we can remove the dependence on the unknown residuals \widehat{V}_{n+1} using the well-known fact that $\widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}) \iff \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}^\infty)$ (see Chapter 1, Lemma 2.3). Thus, we also have $\mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}^\infty) \right\} \geq 1 - \alpha$. \square

We refer to the original paper [Guan, 2022] for the proof of Theorem 4.2 and Lemma 4.3.

17 Proof of Theorem 4.4: Training-conditional of LCP-RF

In this section, we prove Theorem 4.4 that shows how to adapt the LCP-RF approach to have training-conditional or PAC coverage.

Theorem 17.1. *Suppose that all observations are i.i.d. drawn from the distribution P . For any given $\epsilon > 0$ and $\alpha - \epsilon > 0$, let $\hat{\alpha}$ be the smallest value in the uniform grid $T = \{\alpha_1 = \frac{1}{K}, \dots, \alpha_K = 1\}$ of size K such that*

$$\sum_{i=1}^{n_2} \frac{1}{n_2} \mathbb{1}_{\hat{V}_i^2 \leq \mathcal{Q}(1 \wedge (\hat{\alpha}(\mathbf{X}_i^2) + \hat{\alpha}); \mathcal{F}_i^{2, \infty})} \geq 1 - \alpha. \quad (23)$$

Then, we have

$$\mathbb{P}_{P^{n_1}} \left\{ \text{cov}(\mathcal{D}_{n_1}) \geq 1 - \alpha - \epsilon \right\} \geq 1 - \delta, \quad (24)$$

with $\delta = K \exp(-2n_2\epsilon^2)$ and $\text{cov}(\mathcal{D}_{n_1}) = \mathbb{P}_P \left\{ \hat{V}_{n+1} \leq \mathcal{Q}(1 \wedge (\tilde{\alpha}(\mathbf{X}_{n+1}) + \hat{\alpha}); \mathcal{F}_{n+1}^\infty) \mid \mathcal{D}_{n_1} \right\}$.

Remark. This result is valid under the i.i.d assumption and not under exchangeability as the other results of this chapter. We suggest choosing a grid with $K = 10, T \subset [0, \alpha]$ as we have observed in most practical scenarios that $\tilde{\alpha}(\mathbf{X}_{n+1}) \approx 1 - \alpha$. However, the central idea remains unaltered - to select a grid that enables transitioning from $\tilde{\alpha}(\mathbf{X}_{n+1})$ to 1. Additionally, as $\tilde{\alpha}(\mathbf{X}_{n+1}) + \hat{\alpha}$ may be above 1, we use $1 \wedge (\tilde{\alpha}(\mathbf{X}_{n+1}) + \hat{\alpha})$ to ensure that it does not exceed 1.

Proof. Recall that $\tilde{\alpha}, \hat{\alpha}$ and \mathcal{F}_{n+1}^∞ is a function of $\hat{\mathcal{D}}_{n_1} = \{\hat{Z}_1, \dots, \hat{Z}_{n_1}\}$ and \mathbf{X}_{n+1} as the RF has been trained on $\hat{\mathcal{D}}_{n_1}$, but we will not specify $\hat{\mathcal{D}}_{n_1}$ for ease of reading. We also assume that $\tilde{\alpha}(\mathbf{X}_{n+1}) + \alpha \leq 1$ for all $\alpha \in T$ without loss of generality to lighten the notations.

$$\begin{aligned} & \mathbb{P}_{P^{n_1}} \left\{ \mathbb{P}_P \left\{ \hat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_{n+1}) + \hat{\alpha}; \mathcal{F}_{n+1}^\infty) \mid \mathcal{D}_{n_1} \right\} \leq 1 - \alpha - \epsilon \right\} \\ & \leq \mathbb{P}_{P^{n_1}} \left\{ \mathbb{P}_P \left\{ \hat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_{n+1}) + \hat{\alpha}; \mathcal{F}_{n+1}^\infty) \mid \mathcal{D}_{n_1} \right\} \leq \sum_{i=1}^{n_2} \frac{1}{n_2} \mathbb{1}_{\hat{V}_i^2 \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_i^2) + \hat{\alpha}; \mathcal{F}_i^{2, \infty})} - \epsilon \right\} \\ & = \mathbb{E} \left[\mathbb{P}_{P^{n_1}} \left\{ \mathbb{P}_P \left\{ \hat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_{n+1}) + \hat{\alpha}; \mathcal{F}_{n+1}^\infty) \mid \mathcal{D}_{n_1} \right\} \leq \sum_{i=1}^{n_2} \frac{1}{n_2} \mathbb{1}_{\hat{V}_i^2 \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_i^2) + \hat{\alpha}; \mathcal{F}_i^{2, \infty})} - \epsilon \right\} \mid \mathcal{D}_{n_1} \right] \\ & \leq \sum_{\alpha \in T} \mathbb{E} \left[\mathbb{P}_{P^{n_1}} \left\{ \mathbb{P}_P \left\{ \hat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_{n+1}) + \alpha; \mathcal{F}_{n+1}^\infty) \mid \mathcal{D}_{n_1} \right\} \leq \sum_{i=1}^{n_2} \frac{1}{n_2} \mathbb{1}_{\hat{V}_i^2 \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_i^2) + \alpha; \mathcal{F}_i^{2, \infty})} - \epsilon \right\} \mid \mathcal{D}_{n_1} \right] \end{aligned}$$

Note that conditionally on \mathcal{D}_{n_1} , $\sum_{i=1}^{n_2} \frac{1}{n_2} \mathbb{1}_{\hat{V}_i^2 \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_i^2) + \alpha; \mathcal{F}_i^{2, \infty})}$ is the average of n_2 bernoulli-trial with mean $\mathbb{P}_P \left\{ \hat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_{n+1}) + \alpha; \mathcal{F}_{n+1}^\infty) \mid \mathcal{D}_{n_1} \right\}$, therefore we can bound the conditional probability using Hoeffding's inequality. Finally, we have

$$\begin{aligned}
& \mathbb{P}_{P^{n_1}} \left\{ \mathbb{P}_P \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_{n+1}) + \widehat{\alpha}; \mathcal{F}_{n+1}^\infty) \mid \mathcal{D}_{n_1} \right\} \leq 1 - \alpha - \epsilon \right\} \\
& \leq \sum_{\alpha \in T} \mathbb{E} \left[\mathbb{P}_{P^{n_1}} \left\{ \mathbb{P}_P \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_{n+1}) + \alpha; \mathcal{F}_{n+1}^\infty) \mid \mathcal{D}_{n_1} \right\} \leq \sum_{i=1}^{n_2} \frac{1}{n_2} \mathbb{1}_{\widehat{V}_i^2 \leq \mathcal{Q}(\tilde{\alpha}(\mathbf{X}_i^2) + \alpha; \mathcal{F}_i^{2,\infty})} - \epsilon \right\} \mid \mathcal{D}_{n_1} \right] \\
& \leq K \exp(-2\epsilon^2 n_2).
\end{aligned}$$

□

18 Proof of marginal coverage of groupwise LCP-RF

Here, we show that there is no loss in coverage guarantee when calibrating by group. We demonstrate the case of marginal coverage, the groupwise training-conditional is obtained similarly.

Theorem 18.1. *Given a partition of the calibration data \mathcal{D}_n in G_1, \dots, G_L and their associated regions $\mathbf{R} = \{R_1, \dots, R_L\}$ defined by the weighted adjacency matrix G with $G_{i,j} = w_n(\mathbf{X}_i, \mathbf{X}_j)$ of the RF. We denote $R(\mathbf{X}) \in \mathbf{R}$ the region where \mathbf{X} falls and $|\mathbf{R}(\mathbf{X})|$ the number of observations in $\mathbf{R}(\mathbf{X})$. At $\widehat{V}_{n+1} = v$, let define $\tilde{\alpha}(v, R(\mathbf{X}_{n+1}))$ to be the smallest value $\tilde{\alpha} \in \Gamma = \left\{ \sum_{j=1}^k w_n(\mathbf{X}_i, \mathbf{X}_j) : i = 1, \dots, n+1; k = 1, \dots, n+1 \right\}$ such that*

$$\sum_{i: \mathbf{X}_i \in R(\mathbf{X}_{n+1})} \frac{1}{|\mathbf{R}(\mathbf{X}_{n+1})| + 1} \mathbb{1}_{\widehat{V}_i \leq \mathcal{Q}(\tilde{\alpha}(v, R(\mathbf{X}_{n+1})); \mathcal{F}_i^v)} \geq 1 - \alpha. \quad (25)$$

If $C_V(\mathbf{X}_{n+1}) = \{v : v \leq \mathcal{Q}(\tilde{\alpha}(v, R(\mathbf{X}_{n+1})); \mathcal{F}_{n+1}^\infty)\}$, then $\mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \in C_V(\mathbf{X}_{n+1}) \right\} \geq 1 - \alpha$.

Proof.

$$\begin{aligned}
\mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \in C_V(\mathbf{X}_{n+1}) \right\} &= \mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}(\widehat{V}_{n+1}, R(\mathbf{X}_{n+1})); \mathcal{F}_{n+1}^\infty) \right\} \\
&= \sum_{l=1}^L \mathbb{P}(R_l) \mathbb{P}_{P^{n+1}} \left\{ \widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}(\widehat{V}_{n+1}, R(\mathbf{X}_{n+1})); \mathcal{F}_{n+1}^\infty) \mid \mathbf{X}_{n+1} \in R_l \right\} \\
&\geq \sum_{l=1}^L \mathbb{P}(R_l) (1 - \alpha) \quad (\text{by Eq. 25}) \\
&\geq 1 - \alpha.
\end{aligned}$$

□

19 Proof of Theorem 5.4: asymptotic conditional coverage

Here, we prove the asymptotic conditional coverage of the LCP-RF approach or Theorem 5.4. Our primary contribution is Lemma 19.3, which enables us to control the weights of the RF and, subsequently, to proceed with [Guan, 2022]'s proof.

Theorem 19.1. *Suppose that all observations are i.i.d. and let $\tilde{\alpha}(v)$ and $\widehat{C}_V(\mathbf{X}_{n+1})$ define as in Theorem 4.2, i.e., $\tilde{\alpha}(v)$ is the smallest value in $\Gamma = \left\{ \sum_{j=1}^k w_n(\mathbf{X}_i, \mathbf{X}_j) : i, k \in [n+1] \right\}$ such that $\sum_{i=1}^{n+1} \frac{1}{n+1} \mathbb{1}_{\widehat{V}_i \leq \mathcal{Q}(\tilde{\alpha}(v); \mathcal{F}_i^v)} \geq 1 - \alpha$ and $\widehat{C}_V(\mathbf{X}_{n+1}) = \{v : v \leq \mathcal{Q}(\tilde{\alpha}(v); \mathcal{F}_{n+1}^\infty)\}$. Under assumptions 5.1-5.3, we have for all $\epsilon > 0$ and any nonatomic points \mathbf{x}_{n+1} of $P_{\mathbf{X}}$,*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\widehat{V}_{n+1} \in C_V(\mathbf{X}_{n+1}) \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1} \right) = 1 - \alpha \quad \text{and}$$

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\max_v |\tilde{\alpha}(v) - (1 - \alpha)| < \epsilon \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1} \right) = 1.$$

The bootstrap step in Random Forest makes its theoretical analysis difficult, which is why it has been replaced by subsampling without replacement in most studies that investigate the asymptotic properties of Random Forests [Scornet, 2015; Wager, 2017; Goehry, 2020]. To circumvent this difficulty, we will use Honest Forest [Wager, 2017] as a theoretical surrogate. Honest Forest is a variation of random forest that is simpler to analyze, and [Elie-Dit-Cosaque, 2022] have shown that asymptotically, the original forest and the Honest Forest are close a.s. (see Chapter 4, Lemma 8.6), thus we can extend the results from the Honest Forest to the original forest.

The main idea is to use a second independent sample $\mathcal{D}_n^\diamond = \{(\mathbf{X}_i^\diamond, Y_i^\diamond)\}_{i=1}^n$. We assume that we have a Honest Forest, which is a random forest that is grown using \mathcal{D}_n , but uses another sample \mathcal{D}_n^\diamond (independent of \mathcal{D}_n and $\Theta_{1:k}$) to estimate the weights and the prediction. Consequently, akin to the approach detailed in Section 3, the Honest version of the RF Localizer is defined as follows:

$$\widehat{F}^\diamond(r \mid \mathbf{X} = \mathbf{x}, \Theta_1, \dots, \Theta_k, \mathcal{D}_n, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_{n+1}^\diamond) = \sum_{j=1}^{n+1} w_n^\diamond(\mathbf{x}, \mathbf{X}_j^\diamond) \mathbb{1}_{V_j^\diamond \leq r}$$

where $\mathbf{X}_{n+1}^\diamond = \mathbf{X}_{n+1} = \mathbf{x}_{n+1}$ is the test observation, $\widehat{V}_{n+1}^\diamond = +\infty$ unless specified, and

$$w_n^\diamond(\mathbf{x}, \mathbf{X}_j^\diamond) = \frac{1}{k} \sum_{l=1}^k \frac{\mathbb{1}_{\mathbf{X}_i^\diamond \in A_n(\mathbf{x}; \Theta_l)}}{N_{n+1}^\diamond(A_n(\mathbf{x}; \Theta_l))}.$$

$N_{n+1}^\diamond(A_n(\mathbf{x}; \Theta_l))$ is the number of observation of $\{\mathbf{X}_1^\diamond, \dots, \mathbf{X}_n^\diamond\} \cup \{\mathbf{X}_{n+1}\}$ that fall into $A_n(\mathbf{x}; \Theta_l)$. To ease the notations, we do not write $\Theta_1, \dots, \Theta_k, \mathcal{D}_n, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_{n+1}^\diamond$ if not necessary, thus we write $\widehat{F}^\diamond(r \mid \mathbf{x})$ instead of $\widehat{F}^\diamond(r \mid \mathbf{X} = \mathbf{x}, \Theta_1, \dots, \Theta_k, \mathcal{D}_n, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_{n+1}^\diamond)$. The following Lemma 19.2 allows for control of the weights of the Honest Forest.

Lemma 19.2. *Consider $\mathcal{D}_n, \mathcal{D}_n^\diamond$, two independent datasets of independent n samples of (\mathbf{X}, Y) and a tree build using \mathcal{D}_n with bootstrap and bagging procedure driven by Θ . If $N_{n+1}(A_n(\mathbf{x}; \Theta_l))$ is the number of bootstrap observations of $\{\mathbf{X}_i\}_{i=1}^n \cup \{\mathbf{X}_{n+1}\}$ that fall into $A_n(\mathbf{x}; \Theta_l)$ and $N_{n+1}^\diamond(A_n(\mathbf{x}; \Theta_l))$ is the number of observations of $\{\mathbf{X}_i^\diamond\}_{i=1}^n \cup \{\mathbf{X}_{n+1}\}$ that fall into $A_n(\mathbf{x}; \Theta_l)$,*

$$\forall \epsilon > 0, \quad \mathbb{P} \left(|N_{n+1}(A_n(\mathbf{x}; \Theta_l)) - N_{n+1}^\diamond(A_n(\mathbf{x}; \Theta_l))| > \epsilon \right) \leq 24(n+2)^{2p} e^{-\epsilon^2/288(n+1)}. \quad (26)$$

This lemma is a minor adjustment of Lemma 8.7 from Chapter 4, including an additional observation in the training set when computing the forest weights. See the proof here (8.7).

The following Lemma is the key element to prove Theorem 19.1 for Honest RF Localizer.

Lemma 19.3. *Let define for all $i = 1, \dots, n+1$,*

$$R_i = \sum_{j=1}^n w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j) \left(\mathbb{1}_{\widehat{V}_j^\diamond < \widehat{V}_i} - F(\widehat{V}_i | \mathbf{X}_j^\diamond) \right) \quad \text{and} \quad I_i = \sum_{j=1}^n w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j) F(\widehat{V}_i | \mathbf{X}_j^\diamond),$$

where $F(v|\mathbf{x}) = \mathbb{P}(\widehat{V} \leq v | \mathbf{X} = \mathbf{x})$. For any $\epsilon > 0$, under assumptions 5.1-5.3, we have

$$\mathbb{P}(|R_i| > \epsilon) \leq 2(1 + 24k(n+2)^{2p}) \exp\left(\frac{K \ln(n+1)^\beta}{576\sqrt{n+1}} - \frac{\epsilon K \ln(n+1)^\beta}{24}\right)$$

$$I_i \in \left[F(\widehat{V}_i | \mathbf{X}_i) - r(n) - \frac{1}{K\sqrt{n+1} \ln(n+1)^\beta}, \quad F(\widehat{V}_i | \mathbf{X}_i) + r(n) + \frac{1}{K\sqrt{n+1} \ln(n+1)^\beta} \right] \text{ a.s.}$$

where $r(n)$ is a sequence s.t. $r(n) \xrightarrow{n \rightarrow \infty} 0$.

Proof. First, let's rewrite R_i as

$$R_i = \sum_{j=1}^n w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j) \left(\mathbb{1}_{\widehat{V}_j^\diamond < \widehat{V}_i} - F(\widehat{V}_i | \mathbf{X}_j^\diamond) \right) = \sum_{j=1}^n w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j) H_j^\diamond$$

where H_j^\diamond is bounded by 1 and $E[H_j^\diamond | \mathbf{X}_j^\diamond, \mathcal{D}_n] = 0$. Then, for all $\epsilon > 0$

$$\begin{aligned} \mathbb{P}(R_i > \epsilon) &\leq e^{-t\epsilon} \mathbb{E}[e^{tR_i}] \\ &\leq e^{-t\epsilon} \mathbb{E} \left[\prod_{j=1}^n \mathbb{E} \left[e^{t w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j) H_j^\diamond} | \Theta_1, \dots, \Theta_k, \mathcal{D}_n, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_n^\diamond \right] \right] \\ &\leq e^{-t\epsilon} \mathbb{E} \left[\prod_{j=1}^n e^{\frac{t^2}{2} w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j)^2} \right] \end{aligned}$$

The last inequality comes from $w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j)$ being constant given $\Theta_1, \dots, \Theta_k, \mathcal{D}_n, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_n^\diamond$, and as H_j^\diamond is bounded by 1 with $E[H_j^\diamond | \mathbf{X}_j^\diamond, \mathcal{D}_n] = 0$, we leverage the subsequent inequality: If $|X| \leq 1$ a.s and $\mathbb{E}[X] = 0$, then $\mathbb{E}[e^{tX}] \leq e^{\frac{t^2}{2}}$. Indeed, by using the convexity of exponential, we have $\mathbb{E}[e^{tX}] \leq \mathbb{E}\left[\frac{1-X}{2}\right] e^{-t} + \mathbb{E}\left[\frac{1+X}{2}\right] e^t \leq \cosh(t) \leq e^{\frac{t^2}{2}}$.

Using assumption 5.3, there exists $K > 0$ such that for all $l \in [k]$, $N_{n+1}(A_n(\mathbf{X}_i; \Theta_l)) \geq N_n(A_n(\mathbf{X}_i; \Theta_l)) \geq K\sqrt{n+1} \ln(n+1)^\beta$ a.s., then we have the event $\Upsilon(l) = \{N_{n+1}^\diamond(A_n(\mathbf{X}_i; \Theta_l)) < \frac{K\sqrt{n+1} \ln(n+1)^\beta}{2}\} \subset \{|N_{n+1}(A_n(\mathbf{X}_i; \Theta_l)) - N_{n+1}^\diamond(A_n(\mathbf{X}_i; \Theta_l))| > \frac{K\sqrt{n+1} \ln(n+1)^\beta}{2}\}$. Thus, using Lemma 19.2, we have that $\mathbb{P}(\Upsilon(l)) \leq 24(n+2)^{2p} \exp(-\frac{K^2(\ln(n+1)^{2\beta})}{1152})$. We have

$$\begin{aligned} \sum_{j=1}^n w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j)^2 &= \sum_{j=1}^n w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j) \times \frac{1}{k} \left(\sum_{l=1}^k \frac{\mathbb{1}_{\mathbf{X}_j^\diamond \in A_n(\mathbf{X}_i; \Theta_l)}}{N_{n+1}^\diamond(A_n(\mathbf{X}_i; \Theta_l))} (\mathbb{1}_{\{\Upsilon(l)^c\}} + \mathbb{1}_{\{\Upsilon(l)\}}) \right) \\ &\leq \sum_{j=1}^n w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j) \left(\frac{2}{K\sqrt{n+1} \ln(n+1)^\beta} + \frac{1}{k} \sum_{l=1}^k \mathbb{1}_{\mathbf{X}_j^\diamond \in A_n(\mathbf{X}_i; \Theta_l, \mathcal{D}_m)} \mathbb{1}_{\{\Upsilon(l)\}} \right). \end{aligned}$$

So that,

$$\begin{aligned}
\mathbb{P}(R_i > \epsilon) &\leq \exp(-t\epsilon + \frac{t^2}{K\sqrt{n+1}\ln(n+1)^\beta}) \mathbb{E} \left[\exp \left(\frac{t^2}{2} \mathbf{1}_{\cup_{l=1}^k \Upsilon(l)} \right) \right] \\
&\leq \exp(-t\epsilon + \frac{t^2}{K\sqrt{n+1}\ln(n+1)^\beta}) \times \left(1 + e^{\frac{t^2}{2}} \sum_{l=1}^k \mathbb{P}(\Upsilon(l)) \right) \\
&\leq \exp(-t\epsilon + \frac{t^2}{K\sqrt{n+1}\ln(n+1)^\beta}) \times \left(1 + 24k(n+2)^{2p} \exp \left(\frac{t^2}{2} - \frac{K^2 \ln(n+1)^{2\beta}}{1152} \right) \right).
\end{aligned}$$

Taking $t^2 = \frac{K^2 \ln(n+1)^{2\beta}}{576}$ leads to

$$\mathbb{P}(R_i > \epsilon) \leq (1 + 24k(n+2)^{2p}) \exp \left(\frac{K \ln(n+1)^\beta}{576\sqrt{n+1}} - \frac{\epsilon K \ln(n+1)^\beta}{24} \right).$$

We obtain the same bound for $\mathbb{P}(R_i \leq -\epsilon) = \mathbb{P}(-R_i > \epsilon)$, then by using assumption 5.2, there exists $k = \mathcal{O}(n^\alpha)$ so that the right term is finite, we conclude by Borel-Cantelli that $|R_i|$ goes to 0 a.s. Finally, we have

$$\mathbb{P}(|R_i| > \epsilon) \leq 2(1 + 24k(n+2)^{2p}) \exp \left(\frac{K \ln(n+1)^\beta}{576\sqrt{n+1}} - \frac{\epsilon K \ln(n+1)^\beta}{24} \right).$$

Now, we consider I_i . By assumption 5.2, we have $\forall \mathbf{x} \in \mathbb{R}^d, \forall r \in \mathbb{R}, \sup_{\mathbf{z} \in A_n(\mathbf{x}; \Theta_l)} |F(r|\mathbf{z}) - F(r|\mathbf{x})| \xrightarrow[n \rightarrow +\infty]{a.s.} 0$, then we can assume that there exists a sequence $r(n) \rightarrow 0$ s.t.

$$\forall \mathbf{x} \in \mathbb{R}^d, \forall r \in \mathbb{R}, \sup_{\mathbf{z} \in A_n(\mathbf{x}; \Theta_l)} |F(r|\mathbf{z}) - F(r|\mathbf{x})| \leq r(n) \quad a.s. \quad (27)$$

Consequently,

$$\begin{aligned}
|I_i - F(\widehat{V}_i|\mathbf{X}_i)| &= \left| \sum_{j=1}^{n+1} w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j^\diamond) \left(F(\widehat{V}_i|\mathbf{X}_j^\diamond) - F(\widehat{V}_i|\mathbf{X}_i) \right) - w_n^\diamond(\mathbf{X}_i, \mathbf{X}_{n+1}) F(\widehat{V}_i|\mathbf{X}_{n+1}^\diamond) \right| \\
&\leq \sum_{j=1}^{n+1} w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j^\diamond) \left| F(\widehat{V}_i|\mathbf{X}_j^\diamond) - F(\widehat{V}_i|\mathbf{X}_i) \right| + w_n^\diamond(\mathbf{X}_i, \mathbf{X}_{n+1}) \\
&\leq \sum_{j=1}^{n+1} w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j^\diamond) \sup_{\mathbf{z} \in A_n(\mathbf{X}_i; \Theta_l)} \left| F(\widehat{V}_i|\mathbf{z}) - F(\widehat{V}_i|\mathbf{X}_i) \right| + \frac{1}{K\sqrt{n+1}\ln(n+1)^\beta} \\
&\leq r(n) + \frac{1}{K\sqrt{n+1}\ln(n+1)^\beta}
\end{aligned}$$

We use the fact that by assumption 5.3, we can lower bound the weights of the forest since $N_{n+1}(A_n(\mathbf{X}_i; \Theta_l)) \geq K\sqrt{n+1}\ln(n+1)^\beta$ for all $l \in [k]$, thus we have $w_n^\diamond(\mathbf{X}_i, \mathbf{X}_{n+1}) \leq \frac{1}{K\sqrt{n+1}\ln(n+1)^\beta}$. \square

19.1 Proof of Theorem 19.1

As in [Guan, 2022], we first prove that $\tilde{\alpha}(v) \rightarrow 1 - \alpha$ for any v and then show that the resulting PI of the Honest RF Localizer has a coverage rate with the desired level $1 - \alpha$.

Proof. Let consider $R_i = \sum_{j=1}^n w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j^\diamond) \left(\mathbb{1}_{\widehat{V}_j^\diamond < \widehat{V}_i} - F(\widehat{V}_i | \mathbf{X}_j^\diamond) \right)$, $I_i = \sum_{j=1}^n w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j^\diamond) F(\widehat{V}_i | \mathbf{X}_j^\diamond)$ and $\mathcal{F}_i^{\diamond v} = \widehat{F}^\diamond(\cdot | \mathbf{X}_i^\diamond) = \sum_{j=1}^n w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j^\diamond) \mathbb{1}_{\widehat{V}_j^\diamond \leq \cdot} + w_n^\diamond(\mathbf{X}_i, \mathbf{X}_{n+1}^\diamond) \mathbb{1}_{\widehat{V}_{n+1}^\diamond \leq \cdot}$. when $\widehat{V}_{n+1}^\diamond = v$, with $v \in [0, \infty]$, for all $i = 1, \dots, n+1$. For any $\tilde{\alpha}$ and v , we have

$$J_i(v, \tilde{\alpha}) := \left\{ \widehat{V}_i \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_i^{\diamond v}) \right\} = \left\{ \tilde{\alpha} > \sum_{j \leq n: \widehat{V}_j^\diamond < \widehat{V}_i} w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j^\diamond) + w_n^\diamond(\mathbf{X}_i, \mathbf{X}_{n+1}^\diamond) \mathbb{1}_{v < \widehat{V}_i} \right\}$$

which is the event in the sum defined in Equation (6.7) of Theorem 4.2 using the weights of the Honest Forest. Let consider the right-hand term in the definition of the set $J_i(v, \tilde{\alpha})$, we have

$$\begin{aligned} R_i + I_i - w_n^\diamond(\mathbf{X}_i, \mathbf{X}_{n+1}^\diamond) &\leq R_i + I_i \leq \sum_{j \leq n: \widehat{V}_j^\diamond < \widehat{V}_i} w_n^\diamond(\mathbf{X}_i, \mathbf{X}_j^\diamond) + w_n^\diamond(\mathbf{X}_i, \mathbf{X}_{n+1}^\diamond) \mathbb{1}_{v < \widehat{V}_i} \\ &\leq R_i + I_i + w_n^\diamond(\mathbf{X}_i, \mathbf{X}_{n+1}^\diamond). \end{aligned} \quad (28)$$

Let $\epsilon > 0$, and denote $G = \{i \in \{1, \dots, n\} : |R_i| \leq \epsilon\}$. By Lemma 19.3, we have $I_i \in \left[F(\widehat{V}_i | \mathbf{X}_i) - r(n) - \frac{1}{K\sqrt{n+1}\ln(n+1)^\beta}, F(\widehat{V}_i | \mathbf{X}_i) + r(n) + \frac{1}{K\sqrt{n+1}\ln(n+1)^\beta} \right]$ a.s. Using the upper bound of Equation (28), for any $i \in G$, we have

$$J_i^{\text{down}}(\tilde{\alpha}) := \left\{ \tilde{\alpha} > F(\widehat{V}_i | \mathbf{X}_i) + \epsilon + r(n) + \frac{2}{K\sqrt{n+1}\ln(n+1)^\beta} \right\} \subseteq J_i(v, \tilde{\alpha}) \quad (29)$$

and similiary with the lower bound of Equation (28), we have

$$J_i^{\text{up}}(\tilde{\alpha}) := \left\{ \tilde{\alpha} > F(\widehat{V}_i | \mathbf{X}_i) - \epsilon - r(n) - \frac{2}{K\sqrt{n+1}\ln(n+1)^\beta} \right\} \supseteq J_i(v, \tilde{\alpha}). \quad (30)$$

Hence, we can upper and lower bound the left side of Equation (6.7) in Theorem 4.2, i.e., $\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}[J_i(v, \tilde{\alpha})]$, using $J_i^{\text{up}}(\tilde{\alpha})$ and $J_i^{\text{down}}(\tilde{\alpha})$ as follows,

$$\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}[J_i(v, \tilde{\alpha})] \leq \frac{1}{n+1} + \frac{1}{n+1} \sum_{i \in G} \mathbb{1}[J_i^{\text{up}}(\tilde{\alpha})] + \frac{|\bar{G}|}{n+1}, \quad (31)$$

$$\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}[J_i(v, \tilde{\alpha})] \geq \frac{1}{n+1} \sum_{i \in G} \mathbb{1}[J_i^{\text{down}}(\tilde{\alpha})], \quad (32)$$

where $\mathbb{1}[J]$ is 1 if the event J is true, and 0 otherwise. Note that $W_i = F(\widehat{V}_i | \mathbf{X}_i)$ is an i.i.d. uniform distribution as $\widehat{V}_i | \mathbf{X}_i$ is a continuous random variable. Consequently, on the event $\{|\bar{G}| = 0\}$, if $\tilde{\alpha}$ satisfy the marginal coverage of Equation (6.7) of Theorem 4.2 using the weights of the

Honest RF Localizer, i.e., $\tilde{\alpha}$ is the smallest value in $\Gamma^\diamond = \left\{ \sum_{j=1}^k w_n^\diamond(\mathbf{X}_i^\diamond, \mathbf{X}_j^\diamond) : i, k \in [n+1] \right\}$ s.t. $\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}[J_i(v, \tilde{\alpha})] \geq 1 - \alpha$, then

- By Equation (31), we must have

$$\begin{aligned} \frac{1}{n+1} \left(1 + \sum_{i=1}^n \mathbb{1}[J_i^{up}(\tilde{\alpha})] \right) &\geq 1 - \alpha \\ \implies \tilde{\alpha} &\geq \mathcal{Q} \left(\frac{n+1}{n}(1-\alpha) - \frac{1}{n}; \frac{1}{n} \sum_{i=1}^n W_i \right) - \epsilon - r(n) - \frac{2}{K\sqrt{n+1} \ln(n+1)^\beta}. \end{aligned} \quad (33)$$

The implication comes from the fact that $\frac{1}{n+1} (1 + \sum_{i=1}^n \mathbb{1}[J_i^{up}(\tilde{\alpha})]) \geq 1 - \alpha$ implies that at least $\lceil (n+1)(1-\alpha) \rceil - 1$ of the events $J_i^{up}(\tilde{\alpha}) := \left\{ \tilde{\alpha} > F(\widehat{V}_i | \mathbf{X}_i) - \epsilon - r(n) - \frac{2}{K\sqrt{n+1} \ln(n+1)^\beta} \right\}$ are true. Assuming $J_i^{up}(\tilde{\alpha})$ is true, and replacing $W_i = F(\widehat{V}_i | \mathbf{X}_i)$ then the order statistics $W_{(\lceil (n+1)(1-\alpha) \rceil - 1)}$ should also satisfy the condition by definition.

Note that $\mathcal{Q} \left(\frac{n+1}{n}(1-\alpha) - \frac{1}{n}; \frac{1}{n} \sum_{i=1}^n W_i \right) = W_{(\lceil (n+1)(1-\alpha) \rceil - 1)}$.

- Similarly, with Equation (32), $\tilde{\alpha}$ satisfy the marginal coverage of Equation (6.7) of Theorem 4.2 using the Honest RF Localizer's weight as described above, as long as

$$\begin{aligned} \frac{1}{n+1} \sum_{i=1}^n \mathbb{1}[J_i^{down}(\tilde{\alpha})] &\geq 1 - \alpha \\ \implies \tilde{\alpha} &\geq \mathcal{Q} \left(\frac{n+1}{n}(1-\alpha); \frac{1}{n} \sum_{i=1}^n W_i \right) + \epsilon + r(n) + \frac{2}{K\sqrt{n+1} \ln(n+1)^\beta}. \end{aligned}$$

As $\tilde{\alpha}$ is the smallest value in $\Gamma^\diamond = \left\{ \sum_{j=1}^k w_n^\diamond(\mathbf{X}_i^\diamond, \mathbf{X}_j^\diamond) : i, k \in [n+1] \right\}$ that makes Eq. (6.7) of Theorem 4.2 holds using the weights of the Honest Forest and the maximal deviation between two adjacent weights of the forest is $\frac{1}{K\sqrt{n+1} \ln(n+1)^\beta}$, there exists C such that $\tilde{\alpha}$ is upper bounded by

$$\tilde{\alpha} \leq \mathcal{Q} \left(\frac{n+1}{n}(1-\alpha); \frac{1}{n} \sum_{i=1}^n W_i \right) + \epsilon + r(n) + \frac{2C}{K\sqrt{n+1} \ln(n+1)^\beta}. \quad (34)$$

In addition, even if $\tilde{\alpha}$ satisfies Equation (6.7) of Theorem 4.2 and not $\frac{1}{n+1} \sum_{i=1}^n \mathbb{1}[J_i^{down}(\tilde{\alpha})] \geq 1 - \alpha$, we systematically have $\tilde{\alpha} \leq \mathcal{Q} \left(\frac{n+1}{n}(1-\alpha); \frac{1}{n} \sum_{i=1}^n W_i \right) + \epsilon + r(n) + \frac{2}{K\sqrt{n+1} \ln(n+1)^\beta}$.

Therefore, on the event $\{|\bar{G}| = 0\}$, we have

$$\begin{aligned} \mathcal{Q} \left(\frac{n+1}{n}(1-\alpha) - \frac{1}{n}; \frac{1}{n} \sum_{i=1}^n W_i \right) - \epsilon - r(n) - \frac{2}{K\sqrt{n+1} \ln(n+1)^\beta} &\leq \tilde{\alpha} \\ &\leq \mathcal{Q} \left(\frac{n+1}{n}(1-\alpha); \frac{1}{n} \sum_{i=1}^n W_i \right) + \epsilon + r(n) + \frac{2C}{K\sqrt{n+1} \ln(n+1)^\beta}. \end{aligned}$$

In addition, let $\epsilon' > 0$ and consider the event $H = \left\{ \sup_t |\mathcal{Q}(t; \frac{1}{n} \sum_{i=1}^n W_i) - t| \leq \epsilon' \right\}$, on this even we have

$$\begin{aligned} (1 - \alpha) + \frac{1}{n}(1 - \alpha) - \frac{1}{n} - \epsilon' - \epsilon - r(n) - \frac{2}{K\sqrt{n+1}\ln(n+1)^\beta} &\leq \tilde{\alpha} \\ &\leq (1 - \alpha) + \frac{1}{n}(1 - \alpha) + \epsilon' + \epsilon + r(n) + \frac{2C}{K\sqrt{n+1}\ln(n+1)^\beta}. \end{aligned} \quad (35)$$

Then, there exists C and ϵ s.t.

$$(1 - \alpha) - \epsilon - r(n) - \frac{2C}{K\sqrt{n+1}\ln(n+1)^\beta} \leq \tilde{\alpha} \leq (1 - \alpha) + \epsilon + r(n) + \frac{2C}{K\sqrt{n+1}\ln(n+1)^\beta}$$

We can simplify the previous Equation as

$$|\tilde{\alpha} - (1 - \alpha)| \leq \epsilon + r(n) + \frac{2C}{K\sqrt{n+1}\ln(n+1)^\beta}. \quad (36)$$

Finally, we have

$$\mathbb{P} \left(|\tilde{\alpha} - (1 - \alpha)| > \epsilon + r(n) + \frac{2C}{K\sqrt{n+1}\ln(n+1)^\beta} \right) \leq \mathbb{P}(\bar{G}) + \mathbb{P}(\bar{H}).$$

Using DKW inequality [Massart, 1990] for \bar{H} , and union bound for \bar{G} , we have

$$\mathbb{P}(\bar{H}) = \mathbb{P}(\sup_t |\mathcal{Q}(t; \frac{1}{n} \sum_{i=1}^n W_i) - t| > \epsilon') \leq 2 \exp(-2n\epsilon'^2)$$

$$\mathbb{P}(\bar{G}) = \mathbb{P}(\exists i \in \{1, \dots, n\} : |R_i| > \epsilon) \leq n \times 2(1 + 24k(n+2)^{2p}) \exp \left(\frac{K \ln(n+1)^\beta}{576\sqrt{n+1}} - \frac{\epsilon K \ln(n+1)^\beta}{24} \right).$$

Consequently, we have for any $\hat{V}_{n+1} = v$, if $\tilde{\alpha}(v)$ is the smallest value in Γ^\diamond s.t. $\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}[J_i(v, \tilde{\alpha})] \geq 1 - \alpha$, then $\mathbb{P} \left(|\tilde{\alpha}(v) - (1 - \alpha)| > \epsilon + r(n) + \frac{2C}{K\sqrt{n+1}\ln(n+1)^\beta} \right) \xrightarrow{n \rightarrow \infty} 0$ with $\epsilon + r(n) + \frac{2C}{K\sqrt{n+1}\ln(n+1)^\beta} \xrightarrow{n \rightarrow \infty} 0$ which conclude the first part of the proof.

Now, let's prove that $\lim_{n \rightarrow \infty} \mathbb{P} \left(\hat{V}_{n+1} \in C_V(\mathbf{X}_{n+1}) \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1} \right) = 1 - \alpha$ at almost any nonatomic points of $P_{\mathbf{X}}$. By definition, we have

$$\hat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}^\infty) \iff \sum_{j=1}^n w_n^\diamond(\mathbf{X}_{n+1}, \mathbf{X}_j^\diamond) \mathbb{1}_{\hat{V}_j^\diamond < \hat{V}_{n+1}} = I_{n+1} + R_{n+1} < \tilde{\alpha}. \quad (37)$$

Let's denote $G = \{|R_{n+1}| \leq \epsilon\}$ with $\epsilon = \frac{1}{n}$. On the event G , we can lower and upper bound the left side of Equation (37) using Lemma 19.3 as above. As a result, we have:

$$I_{n+1} + R_{n+1} \leq F(\hat{V}_{n+1} | \mathbf{X}_{n+1}) + r(n) + \frac{2C}{K\sqrt{n+1}\ln(n+1)^\beta} + \epsilon \quad (38)$$

$$I_{n+1} + R_{n+1} \geq F(\hat{V}_{n+1} | \mathbf{X}_{n+1}) - r(n) - \frac{2C}{K\sqrt{n+1}\ln(n+1)^\beta} - \epsilon \quad (39)$$

Since $F(\widehat{V}_{n+1} | \mathbf{X}_{n+1} = \mathbf{x}_{n+1})$ is a uniform distribution, and $\mathbb{P}(\bar{G}) \rightarrow 0$, we have

$$\mathbb{P}(I_{n+1} + R_{n+1} < \tilde{\alpha} | \mathbf{X}_{n+1} = \mathbf{x}_{n+1}) \leq \tilde{\alpha} + r(n) + \frac{2C}{K\sqrt{n+1}\ln(n+1)^\beta} + \epsilon + \mathbb{P}(\bar{G}) \rightarrow \tilde{\alpha} \quad (40)$$

$$\mathbb{P}(I_{n+1} + R_{n+1} < \tilde{\alpha} | \mathbf{X}_{n+1} = \mathbf{x}_{n+1}) \geq \tilde{\alpha} - r(n) - \frac{2C}{K\sqrt{n+1}\ln(n+1)^\beta} - \epsilon \rightarrow \tilde{\alpha}. \quad (41)$$

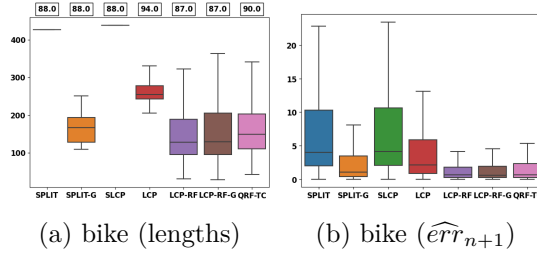
Therefore, we have

$$\mathbb{P}\left(\widehat{V}_{n+1} \leq \mathcal{Q}(\tilde{\alpha}; \mathcal{F}_{n+1}^\infty) | \mathbf{X}_{n+1} = \mathbf{x}_{n+1}\right) = \mathbb{P}(I_{n+1} + R_{n+1} < \tilde{\alpha} | \mathbf{X}_{n+1} = \mathbf{x}_{n+1}) \rightarrow \tilde{\alpha}$$

As we have shown that $\tilde{\alpha} = \tilde{\alpha}(v) \rightarrow 1 - \alpha$ for any v , the LCP-RF achieve the asymptotic conditional coverage at level $1 - \alpha$. \square

20 Additional experiments

In this section, we present additional experiments on real-world datasets. First, we show the lengths and residuals of the PI when $\hat{\mu}$ is a linear model with $\hat{V}(\mathbf{X}, Y) = |Y - \hat{\mu}(\mathbf{X})|$ on bike sharing demand from UCI [Dua, 2017a].



Now, we run the experiment above on star and bike dataset using quantile score $V(\mathbf{X}, Y, \{\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}\}) = \max(\hat{q}_{\alpha/2}(\mathbf{X}) - Y, Y - \hat{q}_{1-\alpha/2}(\mathbf{X}))$. We first estimate $\{\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}\}$ using quantile linear regression [Chernozhukov, 2010] (QLR), then we use Quantile Regression Forest (QRF). Note that in this case, split-CP corresponds to Conformalized Quantile Regression [Romano, 2019].

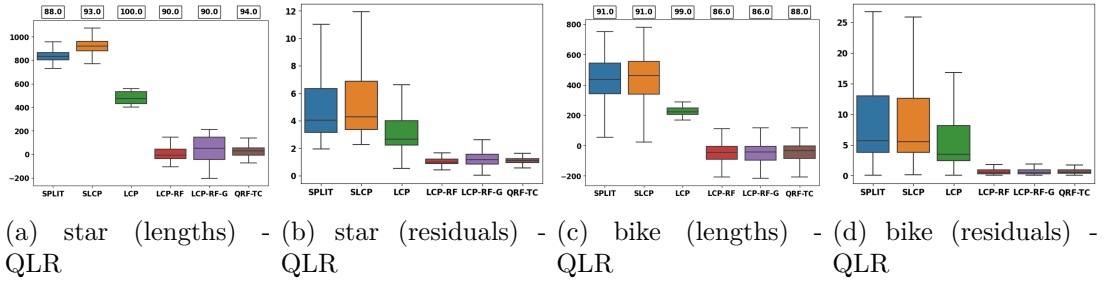


Figure 18: Lengths and errors distribution of quantile score using Quantile Linear Regression

In figure 18c-18d, we observe that QLR gives negative interval lengths. Indeed, this stems from the fact that unlike Quantile Regression Forests, linear quantile regression can face a problem known as "crossed quantile" or "non-monotonicity" [Saleh, 2021; He, 1997] where $\hat{q}_{\alpha_1}(\mathbf{X}) < \hat{q}_{\alpha_2}(\mathbf{X})$ even if $\alpha_1 > \alpha_2$. This leads to negative intervals in our calculations.

We also compute the quantile score using Quantile Regression Forest in the figure below.

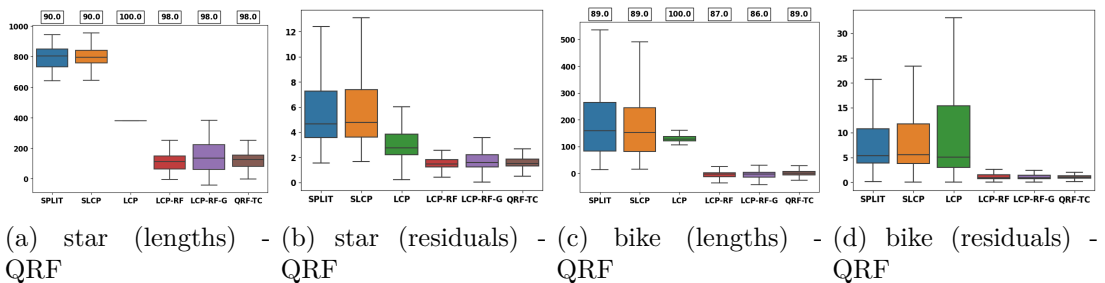


Figure 19: Lengths and errors distribution of quantile score using Quantile Random Forest

All these figures show that the RF Localizer performs much better than the other methods.

