Tackling the XAI Disagreement Problem with Regional Explanations

Gabriel Laberge[1] **Yann Pequignot**[2] **Mario Marchand**[2] **Foutse Khomh**[1]
[1]Polytechnique Montréal
[2]Université Laval à Québec

Abstract

The XAI Disagreement Problem concerns the fact that various explainability methods yield different local/global insights on model behavior. Thus, given the lack of ground truth in explainability, practitioners are left wondering "Which explanation should I believe?". In this work, we approach the Disagreement Problem from the point of view of Functional Decomposition (FD). First, we demonstrate that many XAI techniques disagree because they handle feature interactions differently. Secondly, we reduce interactions locally by fitting a so-called FD-Tree, which partitions the input space into regions where the model is approximately additive. Thus instead of providing global explanations aggregated over the whole dataset, we advocate reporting the FD-Tree structure as well as the regional explanations extracted from its leaves. The beneficial effects of FD-Trees on the Disagreement Problem are demonstrated on toy and real datasets.

1 INTRODUCTION

The Machine Learning paradigm is growing in popularity in data-centric domains. However, there are rising concerns regarding the black-box nature of the performant models *e.g.* Random Forests (Breiman, 2001) and Gradient Boosted Trees (Friedman, 2001). To address these concerns, the field of eXplainable AI (XAI) was introduced, and various post-hoc explanation methods were proposed to provide insight into complex model behaviors (Arrieta et al., 2020). Notable examples of post-hoc explanations include Partial Dependence Plots (PDP) (Friedman, 2001), SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), and Permutation Feature Importance (PFI) (Breiman, 2001).

Despite the explosion in XAI research, there remain fundamental challenges to address. For instance, the so-called *Disagreement Problem* (DP) (Krishna et al., 2022) refers to the observation that different post-hoc explanation methods disagree on the local/global behavior of models. This is not necessarily an issue and it is actually expected since different XAI techniques characterize models differently. Yet, if practitioners are expected to make decisions based on those explanations or said explanations are used to justify decisions impacting human beings, then the DP becomes critical. Which explanations should a practitioner consider? Which explanation should be shown to a client being impacted by the model? Given the lack of ground truth in XAI, these questions are currently left unanswered.

Our partial mitigation of the disagreement problem is inspired by Hybrid Interpretable Models (Wang, 2019; Pan et al., 2020; Ferry et al., 2023). In these works, it is shown that regions exist where complex models can be replaced by simple rules without degradation in overall performance. These analyses suggest that models may be globally complex yet simple in certain regions. Note that it may be possible to explain the model on those regions without any disagreement from post-hoc explainers. As a toy example, consider the features $x_i \sim U(-1, 1)$ for $i = 0, 1, \dots 4$ and the model

$$h(\boldsymbol{x}) = \begin{cases} x_0 & \text{if } x_1 \geq 0 \\ x_2 & \text{otherwise,} \end{cases} \tag{1}$$

which is globally complex but locally simple (linear). To compute global feature importance with PDP/SHAP/PFI, one must provide reference data samples. The typical approach is to use the whole dataset which leads to Figure 1 (a). We observe strong disagreements regarding the importance of x_1. What can we do to reduce this disagreement? We propose to learn two disjoint regions $\Omega_- = \{\boldsymbol{x} \in \mathbb{R}^5 : x_i \leq \gamma\}$ and

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

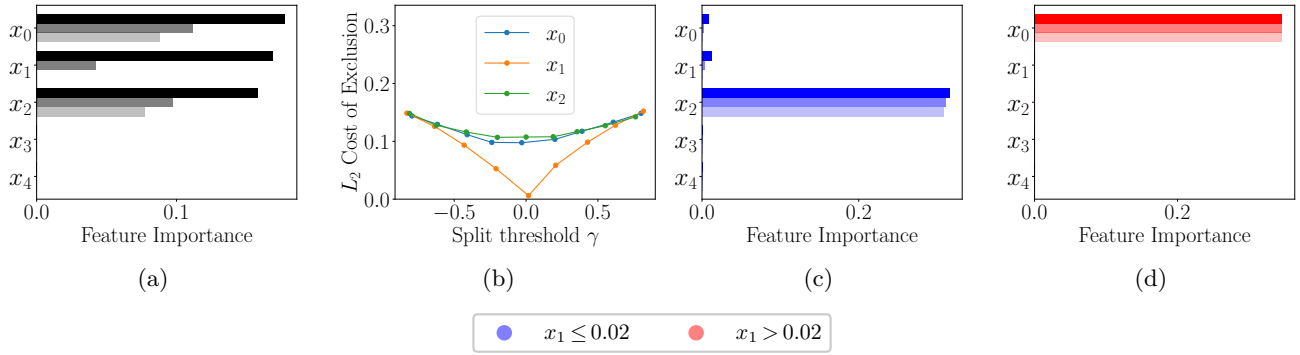


Figure 1: Toy Example. (a) Global Feature Importance when using the whole dataset as reference. The PDP (transparent), SHAP (semi-transparent), and PFI (opaque) importance are differentiated via their opacity. (b) A special loss function is minimized by splitting the input space along a feature. The chosen split is $x_1 \leq 0.02$. (c) & (d) Global feature importance when the reference data is restricted to each region. The two regions are indicated by red/blue colors.

$\Omega_+ = \{\boldsymbol{x} \in \mathbb{R}^5 : x_i > \gamma\}$ and explain the model separately on each region. To define these regions, we must choose the feature x_i to split upon and the threshold value γ. Figure 1 (b) shows the choice of feature i and threshold γ that minimizes an objective function that we will define in **Section 3.1**. This optimization procedure suggests defining two regions based on the conditions: $x_1 \leq 0.02$ and $x_1 > 0.02$. Then, instead of computing global feature importance using the whole dataset as a reference, we only provide reference samples that land in a given region. Doing so leads to a strong agreement between the different techniques, see Figure 1 (c)&(d). This simplified example demonstrates that explaining a model on a well-chosen input region can lead to more faithful results. The main contributions of this work are

1. By unifying the PDP/SHAP/PFI explainers via Functional Decomposition, we show that feature interactions are a necessary condition for explanation disagreement.

2. Since high-dimensional feature interactions are inherently hard to interpret, we argue that PDP, SHAP, and PFI are only **trustworthy** when they **agree** *i.e.* when there are no feature interactions.

3. To increase explanations agreement, we propose to learn partitions of the input space where the model has fewer interactions on each region. The disjoint regions are defined as the leaves of a decision tree.

4. We show on six real datasets that explaining Random Forests and Gradient Boosted Trees on well-chosen regions can increase the agreement between post-hoc explainers.

2 BACKGROUND

2.1 Functional Decompositions

Let $[d] := \{1, \ldots, d\}$ be a set of d features, $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space, and $h : \mathcal{X} \to \mathbb{R}$ be a model. For $u \subseteq [d]$ we denote by $\boldsymbol{x}_u = (x_i)_{i \in u}$ the restriction of the vector $\boldsymbol{x} \in \mathcal{X}$ to the indices in u. For convenience, we denote by ^-S the complement $[d] \setminus S$. Functional Decomposition (FD) aims to represent h as a sum of 2^d sub-functions

$$h(\boldsymbol{x}) = \sum_{u \subseteq [d]} h_u(\boldsymbol{x}), \tag{2}$$

where h_u only depends on $\boldsymbol{x}_u$. The term $h_\emptyset$ is a constant, the terms h_u for $|u| = 1$ are called "additive" while the terms $|u| \geq 2$ are referred to as "$|u|$-way interactions". The model h is said to be additive if there exists a decomposition where $h_u = 0$ whenever $|u| \geq 2$ *i.e.* there are no interactions. Additive models are advertised as being inherently interpretable because the impact of varying a feature on the output is independent of other features Lou et al. (2012). On the other hand, models with interactions h_u ($|u| \geq 2$) are more difficult to interpret because the impact of varying a feature on the output may depend on other features.

Functional Decompositions are far from unique yet they are often computed in the same recursive fashion. For instance, the *Anchored Decomposition* (Kuo et al., 2010) lets $\boldsymbol{z} \in \mathcal{X}$ be a reference input and follows

$$\begin{aligned} h_{\emptyset,\boldsymbol{z}} &:= h(\boldsymbol{z}) \quad \text{(Constant)} \\ h_{i,\boldsymbol{z}}(\boldsymbol{x}) &:= h(\boldsymbol{x}_i, \boldsymbol{z}_{-i}) - h_{\emptyset,\boldsymbol{z}} \\ &\cdots \\ h_{u,\boldsymbol{z}}(\boldsymbol{x}) &:= h(\boldsymbol{x}_u, \boldsymbol{z}_{-u}) - \sum_{v \subset u} h_{v,\boldsymbol{z}}(\boldsymbol{x}). \end{aligned} \tag{3}$$

The intuition is that we progressively construct interaction terms around $\boldsymbol{z}$ by replacing $\boldsymbol{z}_u$ components with $\boldsymbol{x}_u$ while substracting the effects of lower order interactions $v \subset u$. We can also use a distribution $\mathcal{B}$ as a reference

$$\begin{aligned} h_{\emptyset,\mathcal{B}} &:= \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{z})] \quad \text{(Constant)} \\ h_{i,\mathcal{B}}(\boldsymbol{x}) &:= \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{x}_i, \boldsymbol{z}_{-i})] - h_{\emptyset,\mathcal{B}} \\ &\dots \\ h_{u,\mathcal{B}}(\boldsymbol{x}) &:= \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{x}_u, \boldsymbol{z}_{-u})] - \sum_{v\subset u} h_{v,\mathcal{B}}(\boldsymbol{x}). \end{aligned} \tag{4}$$

We shall refer to this decomposition as the *Interventional Decomposition* since replacing $\boldsymbol{z}_u$ components with $\boldsymbol{x}_u$ breaks feature correlations within $\mathcal{B}$. Note that Anchored and Interventional Decompositions are related via $h_{u,\mathcal{B}}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{B}}[h_{u,\boldsymbol{z}}(\boldsymbol{x})]$. If $\mathcal{B} = \mathcal{B}_{\text{ind}} := \prod_{i=1}^{d} \mathcal{B}_i$ (*i.e.* input features are independent), the Interventional Decomposition falls back to the so-called *ANOVA Decomposition* (Hooker, 2004). In such cases, the total variance is the sum of the variance for each individual h_u

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}_{\text{ind}}}[\,(h(\boldsymbol{x}) - h_{\emptyset,\mathcal{B}_{\text{ind}}})^2] = \sum_{\substack{u\subseteq[d] \\ |u|\geq 1}} \sigma_u^2 \tag{5}$$

with $\sigma_u^2 := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{B}_{\text{ind}}}[h_{u,\mathcal{B}_{\text{ind}}}(\boldsymbol{x})^2]$. However, for general $\mathcal{B}$, the variance will not decompose in such a way.

The Anchored/Interventional Decompositions are of interest because they are *minimal* (Kuo et al., 2010). Simply speaking, these FDs do not introduce more interaction terms than necessary. If h does not depend on feature i then $h_{u,\mathcal{B}} = 0$ whenever $i \in u$. If h is additive then the decomposition is also additive. We refer to **Appendix A** for the formal definition of *minimality*.

2.2 Post-hoc Explanations

The field of XAI has risen with the promise of *explaining* black boxes. While there is no universal notion of *explanation*, multiple definitions have been proposed. Local feature attributions for instance are functionals $\boldsymbol{\phi} : \mathcal{H} \times \mathcal{X} \to \mathbb{R}^d$ whose ith component $\phi_i(h, \boldsymbol{x})$ is meant to convey how much each feature i contributes toward the prediction $h(\boldsymbol{x})$. On the other hand, global feature importance are functionals $\boldsymbol{\Phi} : \mathcal{H} \to \mathbb{R}_+^d$ whose ith component $\Phi_i(h)$ illustrates how much feature i is used "globally" by the model (not for a specific prediction). We now go through various functionals proposed in the literature. Partial Dependence Plots (PDP)(Friedman, 2001) take the following form

$$\phi_i^{\text{PDP}}(h, \boldsymbol{x}) := \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{x}_i, \boldsymbol{z}_{-i})]. \tag{6}$$

PDPs are typically visualized graphically as a function of x_i. They can also be extended to compute global feature importance by taking their variance (Greenwell et al., 2018)

$$\Phi_i^{\text{PDP}}(h) := \mathbb{V}_{\boldsymbol{x}\sim\mathcal{B}}[\phi_i^{\text{PDP}}(h, \boldsymbol{x})]. \tag{7}$$

The next post-hoc explainer comes from Cooperative Game Theory. Let, $\nu_{h,\boldsymbol{x}} : 2^{[d]} \to \mathbb{R}$ be a coalitional game which takes $S \subseteq [d]$ and returns $\nu_{h,\boldsymbol{x}}(S) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{x}_S, \boldsymbol{z}_{-S})]$, The Shapley values, as defined in the library SHAP (Lundberg and Lee, 2017), are

$$\phi_i^{\text{SHAP}}(h, \boldsymbol{x}) := \sum_{S\subseteq[d]\setminus\{i\}} W(|S|, d)\big[\nu_{h,\boldsymbol{x}}(S\cup\{i\}) - \nu_{h,\boldsymbol{x}}(S)\big] \tag{8}$$

where $W(|S|, d) := |S|!(d - |S| - 1)!/d!$. The Shapley values are the weighted average contribution of adding feature i to any coallition S that excludes it. They respect an important property called Efficiency:

$$\sum_{i=1}^{d} \phi_i^{\text{SHAP}}(h, \boldsymbol{x}) = h(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{z})]. \tag{9}$$

One way to provide global feature importance consists of averaging square Shapley values [1]

$$\Phi_i^{\text{SHAP}}(h) := \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}[\,\phi_i^{\text{SHAP}}(h, \boldsymbol{x})^2\,]. \tag{10}$$

Permutation Feature Importance (PFI) (Breiman, 2001) was introduced as a global feature importance technique for Random Forest although its definition is model-agnostic. The general idea is to *replace a feature with noise and report the impact on model performance*. Here, we will use the following definition

$$\Phi_i^{\text{PFI}}(h) := \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\Big[\big(\,h(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[\,h(\boldsymbol{x}_{-i}, \boldsymbol{z}_i)\,]\,\big)^2\Big]. \tag{11}$$

which is slightly different from other ones that have appeared in the literature. In **Appendix B** we discuss these differences. Equation 11 is of interest for this work because it can easily be expressed in terms of the Interventional Decomposition. This will allow us to compare PDP, SHAP, and PFI under a common theoretical framework.

We finally introduce disagreement metrics between local and global explanations

$$D(\boldsymbol{\phi}, \boldsymbol{\phi}') := \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\|\boldsymbol{\phi}(h, \boldsymbol{x}) - \boldsymbol{\phi}'(h, \boldsymbol{x})\|_2^2\right] \tag{12}$$

$$D(\boldsymbol{\Phi}, \boldsymbol{\Phi}') := \|\boldsymbol{\Phi}(h) - \boldsymbol{\Phi}'(h)\|_2^2. \tag{13}$$

We focus on norm-based metrics and not on metrics that compare the top-k features or feature rankings.

[1] SHAP actually takes the absolute value but taking the square facilitates our analysis.

Name	Local	Global
PDP	$\phi_i^{\text{PDP}}(h, \boldsymbol{x}) = h_{i,\mathcal{B}}(\boldsymbol{x}) + \text{Constant}$	$\Phi_i^{\text{PDP}}(h) = \mathbb{V}_{\boldsymbol{x}\sim\mathcal{B}}\big[h_{i,\mathcal{B}}(\boldsymbol{x})\big]$.
SHAP	$\phi_i^{\text{SHAP}}(h, \boldsymbol{x}) = \sum_{u\subseteq[d]:i\in u} \frac{h_{u,\mathcal{B}}(\boldsymbol{x})}{\lvert u\rvert}$	$\Phi_i^{\text{SHAP}}(h) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{B}}\left[\left(\sum_{u\subseteq[d]:i\in u} \frac{h_{u,\mathcal{B}}(\boldsymbol{x})}{\lvert u\rvert}\right)^2\right]$
PFI	$\phi_i^{\text{PFI}}(h, \boldsymbol{x}) = \sum_{u\subseteq[d]:i\in u} h_{u,\mathcal{B}}(\boldsymbol{x})$	$\Phi_i^{\text{PFI}}(h) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{B}}\left[\left(\sum_{u\subseteq[d]:i\in u} h_{u,\mathcal{B}}(\boldsymbol{x})\right)^2\right]$

Table 1: Expressing the various post-hoc explainers in terms of the Interventional Decomposition.

The reason is that, in **Section 3.1**, we will derive theoretical guarantees covering Equations 12 & 13. Note that it is extremely difficult to obtain theoretical guarantees with combinatorial quantities such as ranks.

3 METHODOLOGY

3.1 Lack of Additivity

We unify the various post-hoc explainers from the point of view of the Interventional Decomposition, see Table 1. Firstly, the PDPs are the first components of the FD up to a *Constant*. Since the profile of the PDP curve is more useful (not its value), the *Constant* can be fixed to zero without impacting interpretation. Secondly, SHAP computes local feature attributions by sharing the M-way interaction evenly between the M features involved (Herren and Hahn, 2022). Lastly, PFI yields importance to i by accounting for each component involving i, see **Appendix B** for details.

If there were no interactions ($h_{u,\mathcal{B}} = 0$ for $|u| \geq 2$), all explainers would agree and yield $\phi_i(h, \boldsymbol{x}) = h_{i,\mathcal{B}}(\boldsymbol{x})$ locally and $\Phi_i(h) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{B}}[h_{i,\mathcal{B}}(\boldsymbol{x})^2]$ globally. Therefore, tackling the Disagreement Problem requires quantifying interaction strength as a function of the model h and the background $\mathcal{B}$. Later, this will allow us to *design* backgrounds with reduced interactions and increased explanation agreement.

Definition 3.1. *A function $L_h(\mathcal{B}) \in \mathbb{R}^+$ measures Lack of Additivity (LoA) of h w.r.t $\mathcal{B}$ if it respects the following properties.*

1. *If h is additive on a rectangular domain $R \supseteq supp(\mathcal{B})$, then $L_h(\mathcal{B}) = 0$.*
2. *There is a function $w : 2^{[d]} \to \mathbb{R}^+$ such that*

$$L_h(\mathcal{B}_{ind}) = \sum_{u\subseteq[d]:|u|\geq 2} w(u)\sigma_u^2. \tag{14}$$

3. *If h is additive in feature j (i.e. $h(\boldsymbol{x}) = g_j(x_j) + g_{-j}(\boldsymbol{x}_{-j})$), then*

$$L_h(\mathcal{B}_j \times \mathcal{B}_{-j}) = L_h(\mathcal{B}'_j \times \mathcal{B}_{-j}) \tag{15}$$

for any distributions $\mathcal{B}_j$ and $\mathcal{B}'_j$ on feature j and $\mathcal{B}_{-j}$ on features $[d] \setminus \{j\}$. Simply put, the LoA is not affected by additive features unless they correlate with interacting features.

The three properties of this definition are desirable if L_h is to be used as a loss function to minimize w.r.t $\mathcal{B}$. Property 1 guarantees that if the model is additive over the support of the background, then minimization has converged. Property 2 is a sanity check that in the ideal scenario of independent features, the LoA penalizes the variances σ_u^2 which are well-established measures of interaction strength (Hooker, 2004; Owen, 2013). Property 3 allows a reduction of the search space for $\mathcal{B}$. Indeed, if only a subset $I \subset [d]$ of features interact, then we only need to minimize L_h w.r.t $\mathcal{B}_I$ and ignore other features.

Theorem 3.1. *Any function*

$$L_h(\mathcal{B}) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{\substack{u,v\subseteq[d]\\|u|\geq 2,|v|\geq 2}} a(u,v)h_{u,\mathcal{B}}(\boldsymbol{x})h_{v,\mathcal{B}}(\boldsymbol{x})\right].$$

for some $a : 2^{[d]} \times 2^{[d]} \to \mathbb{R}$ is a LoA.

Proof Sketch. We must show that function L_h in the theorem respects the three properties of **Definition 3.1**. Properties 1 and 3 are proven using the *minimality* of the Anchored/Interventional Decompositions (Kuo et al., 2010). Property 2 is proven with the characteristics of the ANOVA decomposition (Owen, 2013, Appendix A).

The implication of this Theorem is that many functions can quantify the LoA. A first possibility would be the L_2 Cost of Exclusion (CoE) (Hooker, 2004), which computes the error between the model and its additive components

$$L_h^{\text{CoE}}(\mathcal{B}) := \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\left(h(\boldsymbol{x}) - \sum_{\substack{u\subset[d]\\|u|\leq 1}} h_{u,\mathcal{B}}(\boldsymbol{x})\right)^2\right]. \tag{16}$$

Other possibilities would be any disagreement between local explainers.

Corollary 3.1. *The distances $D(\phi, \phi')$ between local PDP/SHAP/PFI explainers, as well as the L_2 CoE are all LoA functions.*

This Corollary clarifies the relationship between explanation disagreement and feature interactions. By the first property of LoA, when the model is additive, the disagreement will be null. Conversely, if disagreements occur between local PDP/SHAP/PFI then the model **must** have interactions.

3.2 FD-Trees

We now propose to minimize the Lack of Additivity with a methodology called *Functional Decomposition Trees* (FD-Trees). The main intuition behind it is that, although a complex model is non-additive over its whole domain, there may exist regions where it is approximately additive. More formally, given a background distribution $\mathcal{B}$, and a measurable subset $\Omega \subset \mathcal{X}$ such that $\mathcal{B}(\Omega) > 0$, we define the *local* background $\mathcal{B}_\Omega$ as the measure:

$$\mathcal{B}_\Omega(A) := \frac{\mathcal{B}(A \cap \Omega)}{\mathcal{B}(\Omega)}, \tag{17}$$

for any measurable subset $A \subset \mathcal{X}$. The assumption behind FD-Trees is that, although there are strong interactions $h_{u,\mathcal{B}}$, the local interactions $h_{u,\mathcal{B}_\Omega}$ ($|u| \geq 2$) may be of smaller magnitude.

3.2.1 Empirical Estimates

In XAI, we often set $\mathcal{B}$ to the distribution that generated the dataset $S := \{\boldsymbol{x}^{(i)}\}_{i=1}^N \sim \mathcal{B}^N$. Yet, since $\mathcal{B}$ is unknown, we must approximate it with the empirical distribution $\widehat{\mathcal{B}} := \frac{1}{N}\sum_{\boldsymbol{x}^{(i)} \in S} \delta(\boldsymbol{x}^{(i)})$. Applying Equation 17 leads to the following local backgrounds

$$\widehat{\mathcal{B}}_\Omega = \frac{1}{|S \cap \Omega|} \sum_{\boldsymbol{x}^{(i)} \in S \cap \Omega} \delta(\boldsymbol{x}^{(i)}). \tag{18}$$

The corresponding Interventional Decomposition $h_{u,\widehat{\mathcal{B}}_\Omega}$ describes h relative to the average prediction on datapoints that land in Ω. To simplify notation, the corresponding LoA will be renamed $\widehat{L}_h(S_\Omega) \equiv L_h(\widehat{\mathcal{B}}_\Omega)$ with $S_\Omega := S \cap \Omega$. We let $\boldsymbol{H}$ be the $N \times N \times d$ tensor whose components

$$H_{ijk} := h_{k,\boldsymbol{x}^{(j)}}(\boldsymbol{x}^{(i)}) := h(\, \boldsymbol{x}_k^{(i)}, \boldsymbol{x}_{-k}^{(j)}\,) - h(\boldsymbol{x}^{(j)}) \tag{19}$$

are the h_k of the $\boldsymbol{x}^{(j)}$-anchored decomposition evaluated at $\boldsymbol{x}^{(i)}$. Given $\boldsymbol{H}$, the L_2CoE is

$$\widehat{L}_h^{\text{CoE}}(S_\Omega) = \frac{1}{|S_\Omega|} \sum_{i \in S_\Omega} \left(H_{ii}^{\text{add}} - \frac{1}{|S_\Omega|} \sum_{j \in S_\Omega} H_{ij}^{\text{add}} \right)^2, \tag{20}$$

where $H_{ij}^{\text{add}} := \sum_{k=1}^d H_{ijk} + h(\boldsymbol{x}^{(j)})$. The disagreement between local PDP and PFI $D(\boldsymbol{\phi}^{\text{PDP}}, \boldsymbol{\phi}^{\text{PFI}})$ is also simple to compute given $\boldsymbol{H}$ and yields

$$\widehat{L}_h^{\text{PDP-PFI}}(S_\Omega) = \frac{1}{|S_\Omega|} \sum_{k \in [d]} \sum_{i \in S_\Omega} \left(\frac{1}{|S_\Omega|} \sum_{j \in S_\Omega} H_{ijk} + H_{jik} \right)^2. \tag{21}$$

We do not investigate other LoA functions involving SHAP since they would imply precomputing another $N \times N \times d$ tensor containing the Shapley values. With Equations 20&21, only the tensor $\boldsymbol{H}$ is required.

3.2.2 Learning a Partition

To reduce interactions and increase the agreement between various post-hoc explainers, we learn a partition of $\mathcal{X}$ into M regions $(\Omega_1, \Omega_2, \ldots, \Omega_M)$ that minimizes the total LoA

$$\underset{(\Omega_1, \Omega_2, \ldots, \Omega_M)}{\text{argmin}} \sum_{k=1}^M \widehat{L}_h(S_{\Omega_k}). \tag{22}$$

Equation 22 is intractable so solving it requires approximation. We shall restrict the set of all partitions of $\mathcal{X}$ to the set of leaves of depth $\log_2(M)$ decision trees. These decision trees will be grown in a greedy fashion where at each internal node n, we will search for the feature $i_n \in [d]$ and threshold $\gamma_n \in \mathbb{R}$ such that splitting examples according to $\mathbb{1}(x_{i_n} \leq \gamma_n)$ will minimize the objective.

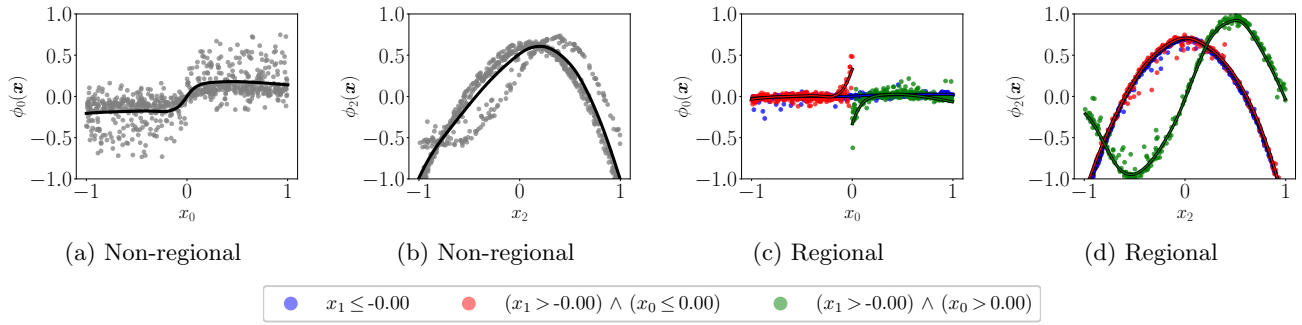
To illustrate the principle of FD-Trees, we study the following toy example

$$\begin{aligned} y(\boldsymbol{x}) =& \mathbb{1}(x_0 \geq 0 \wedge x_1 \geq 0)\sin(\pi x_2) \\ &+ \mathbb{1}(x_0 < 0 \vee x_1 < 0)(-2x_2^2 + x_3). \end{aligned} \tag{23}$$

To simulate a more realistic scenario, we fit this label with a Multi-Layered Perceptron (MLP) h and then explain it. This model is approximately additive in certain regions but not globally additive. Looking at Figure 2 (a) and (b), we note that there are many disagreements between PDP and SHAP which is indicative of strong feature interactions. The effect of x_0 on the model is especially difficult to understand given these plots. By learning an FD-Tree of depth 2 with CoE objective, we identified three regions indicated in blue/red/green in Figure 2 (c) and (d).

3.2.3 Computational Considerations

At each internal node, we must iterate over all features and identify the one along which it is optimal to split. The tree growth could be unnecessarily long if the number of input features is large. However, as the third property of **Definition 3.1** suggests, a split along an additive feature j can only decrease the loss



(a) Non-regional (b) Non-regional (c) Regional (d) Regional

Figure 2: Toy example showing how to increase agreement between PDP (lines) and SHAP (dots).

if it is correlated with another feature that interacts. Since in those cases, j only acts as a proxy, we argue that splitting should only be done on features that interact. As split candidates, we select the k features that interact most, see **Appendix D.1**.

Additionally, the LoA requires a $N \times N \times d$ tensor $\boldsymbol{H}$ which is too large for any realistic dataset. Hence, it is primordial to subsample $N' \ll N$ data points and use them to compute $\boldsymbol{H}$. In our experiments, we subsampled 600 points and addressed the stochasticity by reporting results on 5 random seeds. In **Appendix D.2** we study the impact of subsample size on the stability of the partitions.

3.3 Related Work

3.3.1 Disagreement Problem

Prior work proposes to increase agreement between post-hoc explainers by averaging their importance ranks (Pirie et al., 2023) or by only reporting the common top-k feature with consistent attribution sign (Roy et al., 2022). Such methods do not investigate the root cause of disagreements and thus offer no insight into the Disagreement Problem.

It has also been suggested to use Pearson/Spearman correlations of different explainers as a regularization for training Neural Networks (Schwarzschild et al., 2023). The authors observed that such regularization tends to make the model more linear w.r.t the input features. This coincides with our theoretical analysis showing that h being additive (which is more general than h being linear) is a sufficient condition for explanation agreement.

3.3.2 Regional Explanations

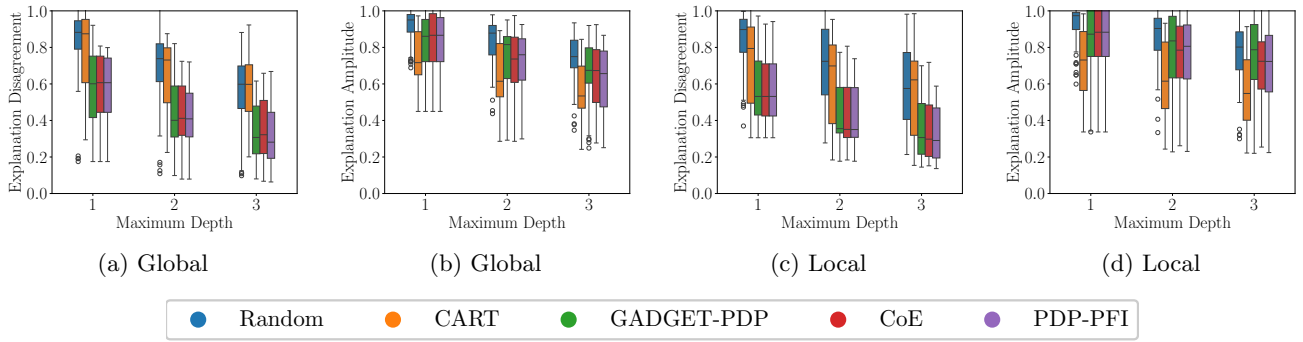
Partitioning the input space via decision trees to obtain regional explanations is not new. Regional Effect Plots with implicit Interaction Detection (REPID) have been previously proposed to grow such trees (Herbinger et al., 2022). The loss function minimized during feature splitting is the L_2 loss between the mean-centered PDP of feature i and the mean-centered ICE curves (Goldstein et al., 2015). Although REPID is similar to FD-Trees, it is only applicable to PDPs and not SHAP/PFI explanations. Moreover, a separate tree must be fitted for each individual feature. In constrast, growing a single FD-Tree can yield regional PDP for each feature.

Generalized Additive Decomposition of Global EffecTs (GADGET)(Herbinger et al., 2023) is a generalization of REPID to more explainers. It uses a decision tree to compute regional PDP, ALE(Apley and Zhu, 2020), and SHAP with reduced interactions. Still, there are two key differences in our methodology. First, GADGET uses different losses for PDP and SHAP. Thus, two tree-growth runs are required to compute regional PDP/SHAP explanations. On the contrary, FD-Trees unify PDP and SHAP via Interventional Decompositions. So a single FD-Tree is required for regional SHAP and PDP explanations. Second, the way SHAP is handled within GADGET is suboptimal. Indeed, GADGET recomputes SHAP values after each split to account for the fact that the background distribution has changed. Also, the loss computation requires regressing splines on the Shapley values. In contrast, growing an FD-Tree is extremely fast since the tensor $\boldsymbol{H}$ is precomputed before growing the tree.

Finally, as the following proposition highlights, GADGET with the PDP loss (henceforth called GADGET-PDP) is in fact a FD-Tree.

Proposition 3.1. *The GADGET-PDP loss is a LoA.*

Consequently, GADGET-PDP can reduce disagreements between PDP/SHAP/PFI explanations. This is an interesting result since GADGET-PDP was designed with only PDP and ICE in mind. For the details behind GADGET and the proof of **Proposition 3.1**, we refer to **Appendix C**.



(a) Global (b) Global (c) Local (d) Local

Figure 3: Explanation disagreements and amplitudes for two baselines (Random, CART) compared to various FD-Trees (GADGET-PDP, CoE, PDP-PFI). Results are presented for trees of depth 1, 2 and 3. The disagreements/amplitudes were **normalized** w.r.t disagreements/amplitudes obtained when the whole data is considered as the background.

Table 2: P-values of the Repeated-Measure-ANOVA tests comparing the explanation disagreements between the GADGET-PDP, CoE, and PDP-PFI objectives. For each p-value lower than 0.05, we also show the objective leading to the least disagreements : (1) GADGET-PDP (2) CoE (3) PDP-PFI.

Locality	Default-Credit	Adult	Marketing	Kin8nm	BikeSharing	California
Local	**0.003 (3)**	0.31	**0.015 (3)**	0.42	0.06	**0 (3)**
Global	**0.01 (3)**	**0.03 (3)**	0.25	**0 (3)**	**0.001 (3)**	0.7

4 EXPERIMENTS

We experimentally assessed the viability FD-Trees on various datasets and models[2]. The UCI classification datasets Adults, Default-Credit, and Marketing were employed while, for regression tasks, the UCI dataset BikeSharing, Kin8nm, and the StatLib dataset California were investigated. The two types of models that were considered are Random Forests (RF) and Gradient-Boosted Trees (GBT). The Scikit-Learn (Pedregosa et al., 2011) implementations of these models were used. For each dataset and model type, we trained a separate model for five different random seeds. Then, for each of the resulting 60 models, 9 FD-trees were fitted with maximum depth $1, 2, 3$ and losses GADGET-PDP, CoE, and PDP-PFI. A total of $60 \times 9 = 540$ FD-Trees were obtained. Since the sizes of the datasets were too high to store the full $N \times N \times d$ tensor $\boldsymbol{H}$ required for training FD-Trees, we sub-sampled 600 data points to generate $\boldsymbol{H}$ using the same seed as during model training. Thus, the seed controls all stochasticity in the FD-Tree growth since it impacts both the model h and the sample of data.

[2]The code to reproduce our experiments is available at `https://github.com/gablabc/UXAI_ANOVA`

4.1 Quantitative Results

Are FD-Trees able to significantly reduce the disagreements between post-hoc explainers? Before addressing this question, note that the disagreement metrics $D(\boldsymbol{\phi}, \boldsymbol{\phi}')$ are scale-sensitive: for any $\epsilon \in]0, 1[$ we have $D(\epsilon\boldsymbol{\phi}, \epsilon\boldsymbol{\phi}') < D(\boldsymbol{\phi}, \boldsymbol{\phi}')$. This introduces a bias since reducing the explanation norm can reduce explanation disagreements. Remember that PDP/SHAP/PFI all describe the model predictions $h(\boldsymbol{x})$ relative to the mean $\mathbb{E}_{\boldsymbol{z}\sim\mathcal{B}_\Omega}[h(\boldsymbol{z})]$. So, if the background $\mathcal{B}_\Omega$ is very local, the model may deviate less from its average and explanations will naturally be smaller. Consequently, even a **random** tree can identify regions with reduced disagreements, a fact that we consistently observed empirically. A basic sanity check for FD-Trees is to compare the reduction in explanation disagreements to those induced by random trees. A stronger sanity check is comparing FD-Trees to regions yielded by a Classification And Regression Tree (CART) fitted on the model output. CART minimizes the deviation $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{B}_\Omega}[\,(h(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{B}_\Omega}[h(\boldsymbol{z})]\,)^2\,]$ at each leaf Ω, and so directly minimizes the norms of the explanations.

Figure 3 presents the results comparing the baselines (Random, CART) to various FD-Trees (GADGET-PDP, CoE, PDP-PFI). From Figure 3(a)&(c), we note that FD-Trees lead to greater reductions in disagreements compared to both baselines. These differences are statistically significant according to paired

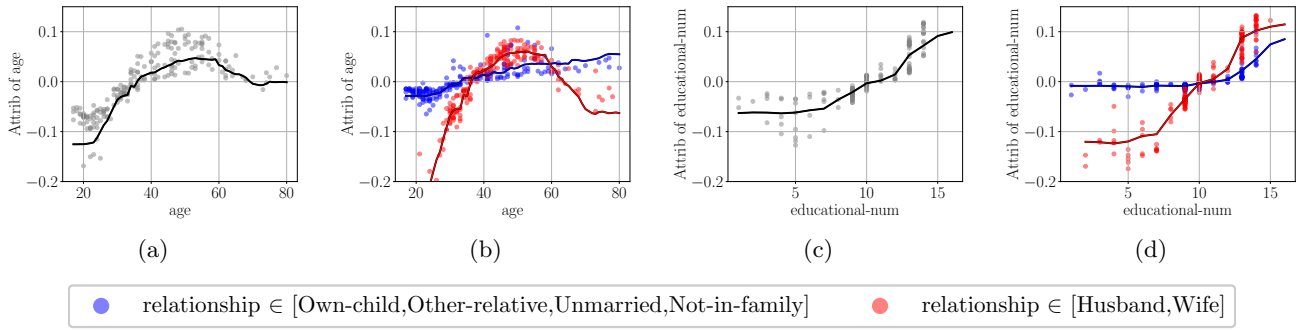


Figure 4: Adult Income. Lines are PDPs while points are SHAP values. (a)&(c) represent the SHAP and PDP explanations when the background is set to the whole dataset. (b)&(d) plot regional explanations with backgrounds restricted to the two regions indicated in red/blue colors.

Student-t tests. Still, is there more agreement simply because explanations are smaller? Looking at the explanation amplitudes in Figure 3(b)&(d), CART is by far the method that leads to the smallest explanation norms. Yet, CART did not manage to reduce explanation disagreements as much as FD-Trees could. This demonstrates that solving the Disagreement Problem is not as trivial as making explanations smaller, and that feature interactions are a key quantity to minimize.

Given that FD-Trees (GADGET-PDP, CoE, PDP-PFI) can identify useful regions, we compared them more thoroughly via Repeated-Measure-ANOVA tests, see Table 2. Repeated-Measure-ANOVA aims to identify if there are significant differences in "outcome" for various "treatments" applied to recurring "subjects". In our setting the "outcome" is the explanation disagreement, the "subjects" are the 30 combinations of model type, random seed, and depth of the FD-Tree, while the "treatment" is the objective employed when growing the tree. According to Table 2, there are often no significant differences between GADGET-PDP, CoE, and PDP-PFI. When the differences are significant, it is systematically the PDP-PFI objective that leads to the least disagreements.

4.2 Qualitative Results

We discuss Adults and BikeSharing here and the other four datasets in **Appendix D.3**. An industrial use-case is also presented in **Appendix E**.

4.2.1 Adult-Income

The Adult-Income task is to predict if someone makes more than 50k USD based on demographic attributes. The first step of the analysis was to compute post-hoc explanations using the whole dataset as the background distribution. According to Figure 4 (a)&(c), the resulting PDP local attributions are a poor estimate of the SHAP values. This warns us that strong feature interactions make the local attributions of `age` and `educational-num` unreliable. To reduce disagreements, we studied the regional explanations over the leaves of a FD-Tree. All FD-Trees that were trained on RFs and Adult-Income identified the same first split: separating married from unmarried people. Figure 4 (b)&(d) show the corresponding regional explanations. Our first observation is that errors between PDP and SHAP suddenly decrease. Secondly, the model has very distinctive behaviors between married and unmarried people. From Figure 4 (b), the attribution of `age` tends to be negative for younger people, but to a larger extent if you are married. Also, the attribution of `age` becomes negative for older married people while it remains positive for older unmarried ones. Similar observations can be made for `education-num`.

4.2.2 Bike-Sharing

The BikeSharing dataset aims to predict the hourly count of bike rentals between years 2011 and 2012 in Washington state, based on time and weather features. Figure 5 (a) presents the global feature importance of a RF when the background is the whole dataset. We note that the importance of the feature `workingday` is highly uncertain. Indeed, PFI ranks it 2, SHAP ranks it 4, and PDP assigns it no importance. This implies that `workingday` interacts strongly with other features. Interestingly, all the FD-Trees trained on BikeSharing used `workingday=True` and `workingday=False` as their first split. However, the trees differ in their subsequent splits. The rest of Figure 5 presents the results on a depth-2 FD-Tree trained on a RF model with CoE objective. The four identified regions differentiate working days from non-working days and early morning hours from daytime hours. According to Figure 5 (b) presenting the at-

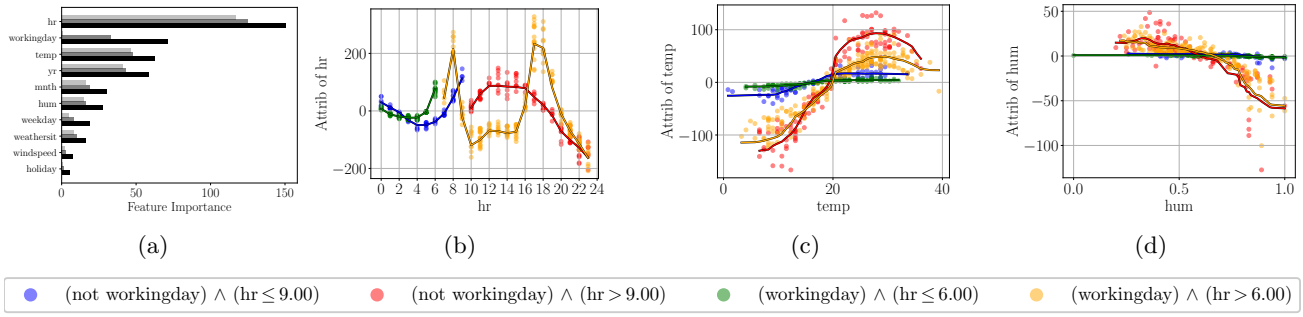


Figure 5: BikeSharing. Lines are PDPs while points are SHAP values. The bar chart (a) is the global feature importance when the whole dataset is used as a background distribution. For all other sub-figures, the regional explanations are displayed.

tribution of `hr`, the amplitude peaks at rush hours during working days while it peaks in the afternoon on non-working days. Inspecting the local attribution of `temp` from Figure 5 (c), the temperature has minimal effects on the model during early morning hours. Nevertheless, past 9am, the temperature considerably influences model predictions, and to a larger extent for non-working days compared to working days. Similar conclusions are drawn from the humidity feature `hum`.

4.3 Discussion & Limitations

Our method induces a higher cognitive load on users because they must understand the regions description, as well as the model behavior on each separate region. For example, instead of providing a single ranking of global feature importance, one such ranking must be reported for each FD-Tree leaf. We view this as a necessary price to pay in order to gain faithful insight into model behavior. Nevertheless, practitioners can still use the whole dataset as background as long as the explanations of various techniques are shown in tandem to reveal potential interactions. For example, we advocate simultaneously reporting the PDP/SHAP/PFI global importance as in Figure 5(a).

FD-Trees are currently grown in a greedy fashion : we split each node along a feature that is locally optimal, and we never consider future impacts of a split, nor do we backtrack on any previous choice. Greedy strategies are common in tree induction because of the considerable search space (Louppe, 2014). Even so, investing more time to find a better solution (*e.g.* via look-ahead strategies (Esmeir and Markovitch, 2007)) may prove beneficial for FD-Trees.

Discontinuities in the local explanation of a feature occur when FD-Trees split along them, see Figure 5(b) for example. These discontinuities must not be interpreted as model discontinuities. They are simply artifacts of considering disjoint regions.

The Accumulated Local Effect (ALE)(Apley and Zhu, 2020) explainer was not investigated because, although its *theoretical* definition agrees with PDP/SHAP/PFI when the model is additive, its *empirical* estimate may not. Indeed, the empirical ALE estimate requires binning the ith feature and the choice of bins impacts the resulting explanation. Thus, it is hard to disentangle the disagreements induced by interactions and those induced by binning. In constrast, PDP/SHAP/PFI are **guaranteed** to agree when the model is additive, irrespective of the sample of data used. Adding ALE to our framework and characterizing the effects of binning is part of future work.

Our entire methodology assumes that more agreement between explanations is better. However, recent work has demonstrated that disagreements can be leveraged to detect irregularities in the model or data (Stando et al., 2023).

5 CONCLUSION

We unified the Partial Dependence Plots, SHapley Additive exPlanations, and Permutation Feature Importance through the lens of Functional Decomposition. We showed that disagreements between these explainers are caused by feature interactions. Thus, we proposed to reduce interactions by partitioning the input space. The background distributions used to compute explanations are then restricted to each region, leading to explanations with reduced disagreements. The benefit of regional explanations was demonstrated both quantitatively and qualitatively on a variety of toy/real datasets and models.

Future work should investigate other definitions of regions. For example, a region could be an open-ball around an instance similar to Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016), or it could be obtained via K-means clustering.

6 Acknowledgements

This work is supported by the DEEL Project CRDPJ 537462-18 funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Consortium for Research and Innovation in Aerospace in Québec (CRIAQ), together with its industrial partners Thales Canada inc, Bell Textron Canada Limited, CAE inc and Bombardier inc.[3]

References

European Union Aviation Safety Agency. Easa artificial intelligence concept paper - proposed issue 2, 2023. URL `https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-concept-paper-proposed-issue-2`.

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):1059–1086, 2020.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

Clément Bénard, Sébastien Da Veiga, and Erwan Scornet. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the sobol-mda. *Biometrika*, 109(4):881–900, 2022.

Leo Breiman. Random forests. *Machine learning*, 45 (1):5–32, 2001.

Saher Esmeir and Shaul Markovitch. Anytime learning of decision trees. *Journal of Machine Learning Research*, 8(5), 2007.

Julien Ferry, Gabriel Laberge, and Ulrich Aïvodji. Learning hybrid interpretable models: Theory, taxonomy, and methods. *arXiv preprint arXiv:2303.04437*, 2023.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. 2008.

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):44–65, 2015.

Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018.

Julia Herbinger, Bernd Bischl, and Giuseppe Casalicchio. Repid: Regional effect plots with implicit interaction detection. In *International Conference on Artificial Intelligence and Statistics*, pages 10209–10233. PMLR, 2022.

Julia Herbinger, Bernd Bischl, and Giuseppe Casalicchio. Decomposing global feature effects based on feature interactions. *arXiv preprint arXiv:2306.00541*, 2023.

Andrew Herren and P Richard Hahn. Statistical aspects of shap: Functional anova for model interpretation. *arXiv preprint arXiv:2208.09970*, 2022.

Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580, 2004.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

F Kuo, I Sloan, Grzegorz Wasilkowski, and Henryk Woźniakowski. On decompositions of multivariate functions. *Mathematics of computation*, 79(270): 953–966, 2010.

Gabriel Laberge and Yann Pequignot. Understanding interventional treeshap: How and why it works. *arXiv preprint arXiv:2209.15123*, 2022.

Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158, 2012.

Gilles Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

Art B. Owen. *Monte Carlo theory, methods and examples.* 2013. URL `https://statweb.stanford.edu/~owen/mc/`.

[3] `https://deel.quebec`

Danqing Pan, Tong Wang, and Satoshi Hara. Interpretable companions for black-box models. In *International conference on artificial intelligence and statistics*, pages 2444–2454. PMLR, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Craig Pirie, Nirmalie Wiratunga, Anjana Wijekoon, and Carlos Francisco Moreno-Garcia. Agree: a feature attribution aggregation framework to address explainer disagreements with alignment metrics. CEUR Workshop Proceedings, 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Saumendu Roy, Gabriel Laberge, Banani Roy, Foutse Khomh, Amin Nikanjam, and Saikat Mondal. Why dont xai techniques agree? characterizing the disagreements between post-hoc explanations of defect predictions. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 444–448. IEEE, 2022.

Avi Schwarzschild, Max Cembalest, Karthik Rao, Keegan Hines, and John Dickerson. Reckoning with the disagreement problem: Explanation consensus as a training objective. *arXiv preprint arXiv:2303.13299*, 2023.

Adrian Stando, Mustafa Cavus, and Przemysław Biecek. The effect of balancing methods on model behavior in imbalanced classification problems. *arXiv preprint arXiv:2307.00157*, 2023.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International conference on machine learning*, pages 9259–9268. PMLR, 2020.

Tong Wang. Gaining free or low-cost interpretability with interpretable partial substitute. In *International Conference on Machine Learning*, pages 6505–6514. PMLR, 2019.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, we state that FD-Trees require computing a tensor $\boldsymbol{H}$ with time and space complexity $\mathcal{O}(N^2d)$.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes, the code is provided in supplementary materials.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes.
   (b) Complete proofs of all theoretical results. Yes, all proofs are available in supplementary materials.
   (c) Clear explanations of any assumptions. Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, the code to reproduce the results is provided in supplementary materials.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not Applicable
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Not Applicable
   (d) A description of the computing infrastructure used. Not Applicable

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes, we cite the `SHAP` library.
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable

(d) Information about consent from data providers/curators. Not Applicable, the data is open source.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

Supplementary Materials

A PROOFS

We first present the minimality property of the Anchored Decomposition.

Theorem A.1 (**Theorem 3.1** from (Kuo et al., 2010))**.** *Let $R \subseteq \mathbb{R}^d$ be a rectangle and let $h : R \to \mathbb{R}$ be a function that can be expressed as $\sum_{u \subseteq [d]} g_u$ where g_u only depends on $\boldsymbol{x}_u$. Also, assume that a subset $v \subset [d]$ exists such that*

$$v \subseteq u \Rightarrow \forall \boldsymbol{x} \in R \;\; g_u(\boldsymbol{x}) = 0.$$

Then, $\boldsymbol{z}$-Anchored Decomposition respects

$$v \subseteq u \Rightarrow \forall \boldsymbol{x}, \boldsymbol{z} \in R \;\; h_{u,\boldsymbol{z}}(\boldsymbol{x}) = 0.$$

Critically, this Theorem requires that the function h has a rectangular domain. Otherwise, the expressions $h(\boldsymbol{x}_u, \boldsymbol{z}_{-u})$ from Equation 3 would not make sense. We can now easily derive the minimality of the Interventional Decomposition.

Corollary A.1. *Let $R \subseteq \mathbb{R}^d$ be a rectangle and let $h : R \to \mathbb{R}$ be a function that can be expressed as $\sum_{u \subseteq [d]} g_u$ where g_u only depends on $\boldsymbol{x}_u$. Also, assume that a subset $v \subset [d]$ exists such that*

$$v \subseteq u \Rightarrow \forall \boldsymbol{x} \in R \;\; g_u(\boldsymbol{x}) = 0.$$

Then, for any probability distribution $\mathcal{B}$ such that supp$(\mathcal{B}) \subseteq R$ the Interventional Decomposition respects

$$v \subseteq u \Rightarrow \forall \boldsymbol{x} \in R \;\; h_{u,\mathcal{B}}(\boldsymbol{x}) = 0.$$

Proof. Let u be a super-set of v ($v \subseteq u$). By **Theorem A.1**, for any $\boldsymbol{x}, \boldsymbol{z} \in R$ we have $h_{u,\boldsymbol{z}}(\boldsymbol{x}) = 0$. Since supp$(\mathcal{B}) \subseteq R$, any sample $\boldsymbol{z} \sim \mathcal{B}$ from the background will land inside the domain R. Hence, for any $\boldsymbol{x} \in R$

$$h_{u,\mathcal{B}}(\boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{B}}[h_{u,\boldsymbol{z}}(\boldsymbol{x})] = \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{B}}[0] = 0.$$

□

Lemma A.1. *If h is additive in feature j then*

$$j \notin u \Rightarrow h_{u,\mathcal{B}_j \times \mathcal{B}_{-j}} = h_{u,\mathcal{B}'_j \times \mathcal{B}_{-j}} \tag{24}$$

for any two distributions $\mathcal{B}_j$ and $\mathcal{B}'_j$ on x_j.

Proof. According to Kuo et al. (2010), the $h_{u,\mathcal{B}}$ component of the Interventional Decomposition can be written

$$h_{u,\mathcal{B}}(\boldsymbol{x}) = \sum_{v \subseteq u} (-1)^{|u \setminus v|} \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{B}}[h(\boldsymbol{x}_v, \boldsymbol{z}_{-v})]. \tag{25}$$

By our assumptions, $h(\boldsymbol{x}) = g_j(x_j) + g_{-j}(\boldsymbol{x}_{-j})$ for some g_j and g_{-j}, and the probability density under $\mathcal{B}$ can be expressed $\rho(\boldsymbol{x}) = \rho_j(x_j)\rho_{-j}(\boldsymbol{x}_{-j})$. Now, for any subset u excluding j we have

$$\begin{aligned}
h_{u,\mathcal{B}}(\boldsymbol{x}) &= \sum_{v \subseteq u} (-1)^{|u \setminus v|} \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{B}}[h(\boldsymbol{x}_v, \boldsymbol{z}_{-v})] \\
&= \sum_{v \subseteq u} (-1)^{|u \setminus v|} \left(\mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{B}}[g_j(\boldsymbol{x}_v, \boldsymbol{z}_{-v})] + \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{B}}[g_{-j}(\boldsymbol{x}_v, \boldsymbol{z}_{-v})] \right) \\
&= \sum_{v \subseteq u} (-1)^{|u \setminus v|} \left(\mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{B}}[g_j(z_j)] + \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{B}}[g_{-j}(\boldsymbol{x}_v, \boldsymbol{z}_{-v})] \right) && \text{(Since } j \notin v\text{)} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{B}}[g_j(z_j)] \sum_{v \subseteq u} (-1)^{|u \setminus v|} + \sum_{v \subseteq u} (-1)^{|u \setminus v|} \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{B}}[g_{-j}(\boldsymbol{x}_v, \boldsymbol{z}_{-v})] && \text{(Note that } \textstyle\sum_{v \subseteq u} (-1)^{|u \setminus v|} = 0\text{)} \\
&= \sum_{v \subseteq u} (-1)^{|u \setminus v|} \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{B}}[g_{-j}(\boldsymbol{x}_v, \boldsymbol{z}_{-v})] \\
&= \sum_{v \subseteq u} (-1)^{|u \setminus v|} \int_{\mathcal{X}} g_{-j}(\boldsymbol{x}_v, \boldsymbol{z}_{-v}) \rho(\boldsymbol{z}) d\boldsymbol{z} \\
&= \sum_{v \subseteq u} (-1)^{|u \setminus v|} \int_{\mathcal{X}} \underbrace{g_{-j}(\boldsymbol{x}_v, \boldsymbol{z}_{-v})}_{\text{does not depend on } z_j} \rho_{-j}(\boldsymbol{z}_{-j}) \rho_j(z_j) d\boldsymbol{z}_{-j} dz_j \\
&= \sum_{v \subseteq u} (-1)^{|u \setminus v|} \int_{\mathcal{X}_{-j}} g_{-j}(\boldsymbol{x}_v, \boldsymbol{z}_{-v}) \rho_{-j}(\boldsymbol{z}_{-j}) d\boldsymbol{z}_{-j} \int \rho_j(z_j) dz_j \\
&= \sum_{v \subseteq u} (-1)^{|u \setminus v|} \int_{\mathcal{X}_{-j}} g_{-j}(\boldsymbol{x}_v, \boldsymbol{z}_{-v}) \rho_{-j}(\boldsymbol{z}_{-j}) d\boldsymbol{z}_{-j}.
\end{aligned}$$

This expression is independent of the choice of $\rho_j(z_j)$. □

Theorem A.2 (**Theorem 3.1**). *Any function*

$$L_h(\mathcal{B}) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\Bigg[\sum_{\substack{u,v\subseteq[d]\\|u|\geq 2,|v|\geq 2}} a(u,v)h_{u,\mathcal{B}}(\boldsymbol{x})h_{v,\mathcal{B}}(\boldsymbol{x})\Bigg] \tag{26}$$

for some $a : 2^{[d]} \times 2^{[d]} \to \mathbb{R}$ *is a LoA.*

Proof. We demonstrate that the function L_h from Equation 26 respects the three properties of **Definition 3.1**.

Property 1 By the minimality of the Interventional Decomposition (cf. **Corollary A.1**), if h is additive over a rectangle $R \supseteq \text{supp}(\mathcal{B})$, then $|u| \geq 2 \Rightarrow \forall \boldsymbol{x} \in R \;\; h_{u,\mathcal{B}}(\boldsymbol{x}) = 0$. So we have

$$L_h(\mathcal{B}) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\Bigg[\sum_{\substack{u,v\subseteq[d]\\|u|\geq 2,|v|\geq 2}} a(u,v)h_{u,\mathcal{B}}(\boldsymbol{x})h_{v,\mathcal{B}}(\boldsymbol{x})\Bigg] = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}[0] = 0.$$

Property 2 By feature independence, we have (Owen, 2013, Appendix A)

$$u \neq v \Rightarrow \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}[h_{u,\mathcal{B}_{\text{ind}}}(\boldsymbol{x})\, h_{v,\mathcal{B}_{\text{ind}}}(\boldsymbol{x})] = 0. \tag{27}$$

Letting $\sigma_u^2 := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{B}_{\text{ind}}}[h_{u,\mathcal{B}_{\text{ind}}}(\boldsymbol{x})^2]$, we therefore obtain

$$L_h(\mathcal{B}_{\text{ind}}) = \sum_{u\subseteq[d]:|u|\geq 2} a(u,u)\,\sigma_u^2, \tag{28}$$

and by identification the interaction penalization weights are $w(u) = a(u,u)$.

Property 3 Let h be additive in feature j meaning that $h(\boldsymbol{x}) = g_j(x_j) + g_{-j}(\boldsymbol{x}_{-j})$. Since the Interventional Decomposition is minimal, there will not be any interaction h_u with $j \in u$. Moreover, let the background factorize as $\mathcal{B} := \mathcal{B}_j \times \mathcal{B}_{-j}$ implying that the corresponding probability density is $\rho(\boldsymbol{x}) = \rho_j(x_j)\rho_{-j}(\boldsymbol{x}_{-j})$.

$$\begin{aligned}
L_h(\mathcal{B}_j \times \mathcal{B}_{-j}) &:= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\Bigg[\sum_{\substack{u,v\subseteq[d]\\|u|\geq 2,|v|\geq 2}} a(u,v)h_{u,\mathcal{B}}(\boldsymbol{x})h_{v,\mathcal{B}}(\boldsymbol{x})\Bigg] = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\Bigg[\sum_{\substack{u,v\subseteq[d]\setminus\{j\}\\|u|\geq 2,|v|\geq 2}} a(u,v)h_{u,\mathcal{B}}(\boldsymbol{x})h_{v,\mathcal{B}}(\boldsymbol{x})\Bigg] \\
&= \int_{\mathcal{X}} \sum_{\substack{u,v\subseteq[d]\setminus\{j\}\\|u|\geq 2,|v|\geq 2}} a(u,v)h_{u,\mathcal{B}}(\boldsymbol{x})h_{v,\mathcal{B}}(\boldsymbol{x})\,\rho(\boldsymbol{x})d\boldsymbol{x} \\
&= \int_{\mathcal{X}} \sum_{\substack{u,v\subseteq[d]\setminus\{j\}\\|u|\geq 2,|v|\geq 2}} a(u,v)h_{u,\mathcal{B}}(\boldsymbol{x})h_{v,\mathcal{B}}(\boldsymbol{x})\,\rho_{-j}(\boldsymbol{x}_{-j})\rho_j(x_j)\,dx_j \\
&= \int_{\mathcal{X}_{-j}} \sum_{\substack{u,v\subseteq[d]\setminus\{j\}\\|u|\geq 2,|v|\geq 2}} a(u,v)h_{u,\mathcal{B}}(\boldsymbol{x})h_{v,\mathcal{B}}(\boldsymbol{x})\,\rho_{-j}(\boldsymbol{x}_{-j})d\boldsymbol{x}_{-j}\int \rho_j(x_j)dx_j \\
&= \int_{\mathcal{X}_{-j}} \sum_{\substack{u,v\subseteq[d]\setminus\{j\}\\|u|\geq 2,|v|\geq 2}} a(u,v)h_{u,\mathcal{B}_j\times\mathcal{B}_{-j}}(\boldsymbol{x})h_{v,\mathcal{B}_j\times\mathcal{B}_{-j}}(\boldsymbol{x})\,\rho_{-j}(\boldsymbol{x}_{-j})d\boldsymbol{x}_{-j}.
\end{aligned}$$

By **Lemma A.1**, for any u not containing j, the subfunction $h_{u,\mathcal{B}_j\times\mathcal{B}_{-j}}$ does not depend on the choice of $\mathcal{B}_j$. Thus, we have proven $L_h(\mathcal{B}_j \times \mathcal{B}_{-j}) = L_h(\mathcal{B}'_j \times \mathcal{B}_{-j})$ for any alternative distribution $\mathcal{B}'_j$ on feature j. □

Corollary A.2 (**Corollary 3.1**). *The distances $D(\boldsymbol{\phi}, \boldsymbol{\phi}')$ between local PFI/SHAP/PFI explainers, as well as the L_2 CoE are all LoA functions.*

Proof. We prove that these various functions take the form of Equation 26. First comparing local PDP and SHAP, we have

$$\begin{aligned}
D(\boldsymbol{\phi}^{\mathrm{PDP}}, \boldsymbol{\phi}^{\mathrm{SHAP}}) &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{k=1}^{d}\left(h_k(\boldsymbol{x}) - \sum_{u\subseteq[d]:k\in u}\frac{h_u(\boldsymbol{x})}{|u|}\right)^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{k=1}^{d}\left(\sum_{u\subseteq[d]:k\in u,|u|\geq 2}\frac{h_u(\boldsymbol{x})}{|u|}\right)^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{k=1}^{d}\sum_{\substack{u,v\subseteq[d]\\ k\in u,k\in v\\ |u|\geq 2,|v|\geq 2}}\frac{h_u(\boldsymbol{x})\,h_v(\boldsymbol{x})}{|u|\,|v|}\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{u,v\subseteq[d]:|u|\geq 2,|v|\geq 2}\frac{|u\cap v|}{|u|\,|v|}\,h_u(\boldsymbol{x})\,h_v(\boldsymbol{x})\right].
\end{aligned}$$

The corresponding interaction penalization is $w(u) = a(u,u) = \frac{1}{|u|}$. Now comparing local PDP and PFI, we obtain

$$\begin{aligned}
D(\boldsymbol{\phi}^{\mathrm{PDP}}, \boldsymbol{\phi}^{\mathrm{PFI}}) &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{k=1}^{d}\left(h_k(\boldsymbol{x}) - \sum_{u\subseteq[d]:k\in u}h_u(\boldsymbol{x})\right)^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{k=1}^{d}\left(\sum_{u\subseteq[d]:k\in u,|u|\geq 2}h_u(\boldsymbol{x})\right)^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{k=1}^{d}\sum_{\substack{u,v\subseteq[d]\\ k\in u,k\in v\\ |u|\geq 2,|v|\geq 2}}h_u(\boldsymbol{x})\,h_v(\boldsymbol{x})\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{u,v\subseteq[d]:|u|\geq 2,|v|\geq 2}|u\cap v|\,h_u(\boldsymbol{x})\,h_v(\boldsymbol{x})\right].
\end{aligned}$$

The corresponding interaction penalization is $w(u) = a(u,u) = |u|$. The disagreement between local SHAP and PFI yields

$$\begin{aligned}
D(\boldsymbol{\phi}^{\mathrm{SHAP}}, \boldsymbol{\phi}^{\mathrm{PFI}}) &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{k=1}^{d}\left(\sum_{u\subseteq[d]:k\in u}\frac{h_u(\boldsymbol{x})}{|u|} - h_u(\boldsymbol{x})\right)^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{k=1}^{d}\left(\sum_{u\subseteq[d]:k\in u,|u|\geq 2}\frac{(1-|u|)\,h_u(\boldsymbol{x})}{|u|}\right)^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{k=1}^{d}\sum_{\substack{u,v\subseteq[d]\\ k\in u,k\in v\\ |u|\geq 2,|v|\geq 2}}\frac{(1-|u|)\,(1-|v|)\,h_u(\boldsymbol{x})\,h_v(\boldsymbol{x})}{|u|\,|v|}\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{u,v\subseteq[d]:|u|\geq 2,|v|\geq 2}\frac{|u\cap v|\,(|u|-1)\,(|v|-1)}{|u|\,|v|}\,h_u(\boldsymbol{x})\,h_v(\boldsymbol{x})\right].
\end{aligned}$$

The corresponding interaction penalization is $w(u) = a(u,u) = (|u|-1)^2/|u|$.

The L_2 Cost of Exclusion (CoE) is also a LoA

$$\begin{aligned} L_h^{\text{CoE}}(\mathcal{B}) &:= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\left(h(\boldsymbol{x}) - \sum_{u\subseteq[d]:|u|\leq 1} h_u(\boldsymbol{x})\right)^2\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\left(\sum_{u\subseteq[d]:|u|\geq 2} h_u(\boldsymbol{x})\right)^2\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\left[\sum_{\substack{u,v\subseteq[d] \\ |u|\geq 2,|v|\geq 2}} h_u(\boldsymbol{x})\, h_v(\boldsymbol{x})\right]. \end{aligned}$$

The corresponding interaction penalization is $w(u) = a(u,u) = 1$. □

We end this section by presenting the various penalization functions $w(u)$ implicit to each LoA when features are independent, see Figure 6. The weights w were normalized so that $w(u) = 1$ when $|u| = 2$. Note that any LoA that involves a disagreement with the PFI explainer will penalize the higher-order interactions to a greater extent. This is because the PFI counts the interaction h_u several times. In opposition, the disagreements between PDP and SHAP put more weight on low-order interactions.

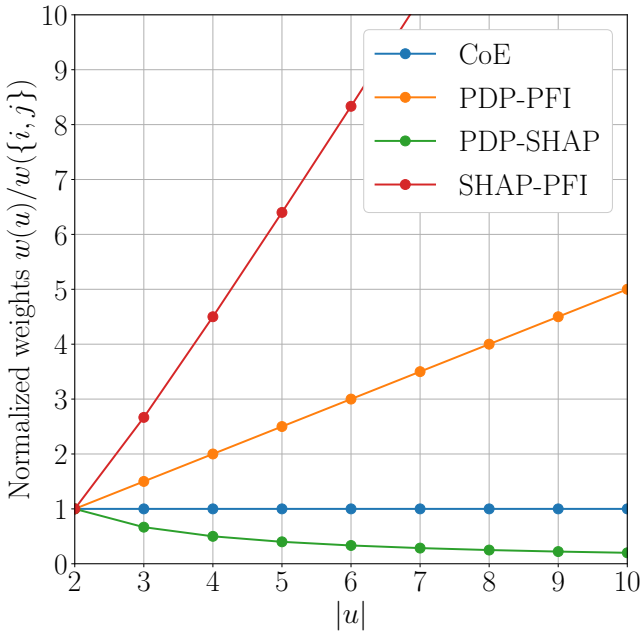


Figure 6: How various LoA penalize interaction orders differently.

B PERMUTATION FEATURE IMPORTANCE

The original definition of the PFI is (Breiman, 2001; Bénard et al., 2022)

$$\Phi_i^{\text{PFI-O}}(h) := \mathop{\mathbb{E}}_{\substack{(\boldsymbol{x},y)\sim\mathcal{B}\\ \boldsymbol{z}\sim\mathcal{B}}}\big[\big(h(\boldsymbol{x}_{-i},\boldsymbol{z}_i)-y\big)^2\big] - \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{B}}\big[\big(h(\boldsymbol{x})-y\big)^2\big], \tag{29}$$

which compares the model performance on the original data and on synthetic data where feature i is replaced by a sample from the marginal. This replacement is typically done by permuting the ith column of the data matrix $\boldsymbol{X}\in\mathbb{R}^{N\times d}$, which justifies the terminology Permutation Feature Importance. However, to establish a natural link between PFI and the Interventional Decomposition of h (and so with the explainers PDP and SHAP), we will slightly change the expression

$$\Phi_i^{\text{PFI}}(h) := \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{B}}\big[\big(\mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{x}_{-i},\boldsymbol{z}_i)]-y\big)^2\big] - \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{B}}\big[\big(h(\boldsymbol{x})-y\big)^2\big]. \tag{30}$$

Crucially, we moved the expectation w.r.t the noise sample $\boldsymbol{z}$ inside the square $(\cdot)^2$. This definition is still in line with the high-level idea of *replacing a feature with noise and reporting the impact on performance*. However, we now average the model predictions on noisy samples before comparing them to the labels. To recover the definition of the PFI used in the paper, we must eliminate the labels y and replace them with model predictions $h(\boldsymbol{x})$. To use y and $h(\boldsymbol{x})$ interchangeably, we need to make the following assumption : $h(\boldsymbol{x})-y=\epsilon$ where ϵ is a random variable that is independent of $\boldsymbol{x}$, $\mathbb{E}[\epsilon]=0$, and $\mathbb{E}[\epsilon^2]=\sigma^2$. Given these assumptions

$$\mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{B}}\big[\big(h(\boldsymbol{x})-y\big)^2\big] = \mathbb{E}[\epsilon^2]=\sigma^2.$$

Letting $g(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{x}_{-i},\boldsymbol{z}_i)]$, the left term in the PFI can be expressed

$$\begin{aligned}
\mathop{\mathbb{E}}_{(\boldsymbol{x},\epsilon)\sim\mathcal{B}}\big[\big(g(\boldsymbol{x})-y\big)^2\big] &= \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{B}}\big[\big(g(\boldsymbol{x})-h(\boldsymbol{x})+\epsilon\big)^2\big]\\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\big[\big(g(\boldsymbol{x})-h(\boldsymbol{x})\big)^2\big] + \mathop{\mathbb{E}}_{(\boldsymbol{x},\epsilon)\sim\mathcal{B}}\big[\epsilon\big(g(\boldsymbol{x})-h(\boldsymbol{x})\big)\big] + \sigma^2\\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\big[\big(g(\boldsymbol{x})-h(\boldsymbol{x})\big)^2\big] + \mathbb{E}[\epsilon]\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\big[\big(g(\boldsymbol{x})-h(\boldsymbol{x})\big)\big] + \sigma^2\\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\big[\big(g(\boldsymbol{x})-h(\boldsymbol{x})\big)^2\big] + \sigma^2
\end{aligned}$$

and so the noise σ^2 computed previously will cancel out. We are left with

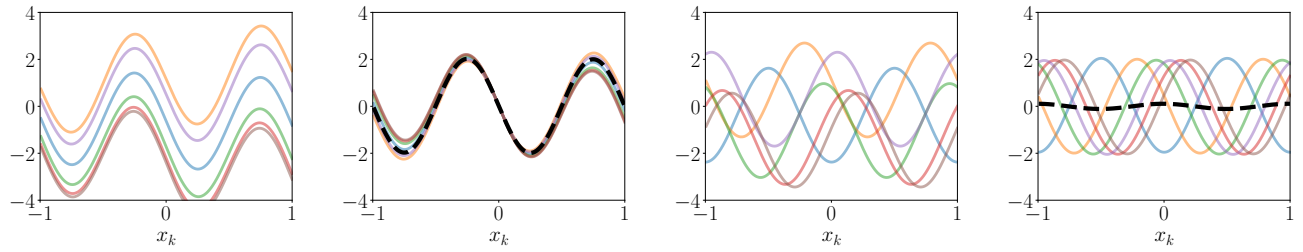
$$\Phi_i^{\text{PFI}}(h) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\big[\big(g(\boldsymbol{x})-h(\boldsymbol{x})\big)^2\big] = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\big[\big(\mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}\big[h(\boldsymbol{x}_{-i},\boldsymbol{z}_i)\big]-h(\boldsymbol{x})\big)^2\big]. \tag{31}$$

This definition is desirable because it can easily be expressed in the Interventional Decomposition. Indeed, from Equation 4 we can deduce that

$$\mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{x}_u,\boldsymbol{z}_{-u})] = \sum_{v\subseteq u} h_{v,\mathcal{B}}(\boldsymbol{x}) \tag{32}$$

and so

$$\begin{aligned}
\Phi_i^{\text{PFI}}(h) &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\big[\big(\mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}\big[h(\boldsymbol{x}_{-i},\boldsymbol{z}_i)\big]-h(\boldsymbol{x})\big)^2\big]\\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\big[\big(\sum_{u\subseteq[d]\setminus\{i\}} h_{u,\mathcal{B}}(\boldsymbol{x}) - \sum_{u\subseteq[d]} h_{u,\mathcal{B}}(\boldsymbol{x})\big)^2\big]\\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}\big[\big(\sum_{u\subseteq[d]:i\in u} h_{u,\mathcal{B}}(\boldsymbol{x})\big)^2\big].
\end{aligned} \tag{33}$$



(a) ICE curve (weak interactions)

(b) Centered ICE curve (weak interactions)

(c) ICE curve (strong interactions)

(d) Centered ICE curve (strong interactions)

Figure 7: Intuition behind the GADGET-PDP. The colored lines are the (centered or uncentered) ICE curves for various values of $\boldsymbol{z}_{-k}$. The dashed dark line is the centered PDP. (a) There are weak interactions involving feature k so the ICE curves for various $\boldsymbol{z}_{-k}$ are nearly parallel. (b) After centering the ICE curves, the centered PDP is computed and is a good estimate of the centered ICEs. Thus the GADGET-PDP loss is very low. (c) There are strong interactions involving feature k and the ICE curves are not parallel. (b) After centering, the PDP is a poor estimate of the ICEs and the GADGET-PDP loss is large.

C GADGET

GADGET (Herbinger et al., 2023) is a similar technique to FD-Trees and so we must take the time to discuss it in detail. Like FD-Trees, GADGET partitions the input space by fitting a decision tree. However, the losses minimized during the tree-growth are different from the ones we propose. GADGET employs two distinct losses depending on whether the user wants to compute regional PDPs or regional SHAP values.

C.1 GADGET-SHAP

We will not discuss in too much detail the loss employed to compute regional SHAP explanations because its computation is suboptimal. Indeed, computing the loss requires regressing the SHAP value $\phi_i^{\text{SHAP}}(h, \boldsymbol{x})$ onto feature x_i with a spline model. The larger the error of this regression, the more interactions involve feature i in h. Crucially, a **different** spline model must be fitted for each split candidate along a feature j. Additionally, GADGET-SHAP requires recomputing SHAP values after each node split because the background distribution has changed. Seeing as running SHAP can take several minutes, growing the full tree can be unnecessarily long. On the other hand, as we are about to prove, the GADGET-PDP loss can be efficiently optimized since it only requires pre-computing a $N \times N \times d$ tensor $\boldsymbol{R}$.

C.2 GADGET-PDP

In this section, we will use a different notation $h(x_k, \boldsymbol{z}_{-k}) \equiv h(\boldsymbol{x}_k, \boldsymbol{z}_{-k})$ to accentuate that x_k is a scalar dimension. Also, for a given background distribution $\mathcal{B}$, the marginal along the subset of feature $S \subset [d]$ is $\mathcal{B}_S$.

GADGET-PDP computes the ICE curves (Goldstein et al., 2015)

$$\phi_k^{\text{ICE}}(h, x_k, \boldsymbol{z}_{-k}) := h(x_k, \boldsymbol{z}_{-k}) \tag{34}$$

which are then centered with respect to x_k to obtain the mean-centered ICE curve

$$\phi_k^{\text{ICE-c}}(h, x_k, \boldsymbol{z}_{-k}) := h(x_k, \boldsymbol{z}_{-k}) - \mathop{\mathbb{E}}_{x_k \sim \mathcal{B}_k}\big[h(x_k, \boldsymbol{z}_{-k})\big]. \tag{35}$$

These curves can be visualized as a function of x_k to understand the effect of feature k when the other features are set to $\boldsymbol{z}_{-k}$. See Figure 7 for a toy example of how centered (and uncentered) ICE curves are typically visualized. Subsequently, by averaging the mean-centered ICE curves w.r.t $\boldsymbol{z}_{-k}$, we obtain the mean-centered PDP

$$\phi_k^{\text{PDP-c}}(h, x_k) := \mathop{\mathbb{E}}_{\boldsymbol{z}_{-k} \sim \mathcal{B}_{-k}}\big[\phi_k^{\text{ICE-c}}(h, x_k, \boldsymbol{z}_{-k})\big]. \tag{36}$$

Like ICEs, the centered PDP is visualized as a function of x_k but it now represents the **average** effect of setting feature k to x_k. If there are no interactions in h involving feature k, then the ICE and PDP curves are parallel

when plotted as functions of x_k. Consequently, the *centered* ICEs and PDP should be identical. The loss employed in GADGET is

$$\begin{aligned} L_h^{\text{GADGET-PDP}}(\mathcal{B}) &:= \sum_{k=1}^{d} \mathop{\mathbb{E}}_{\substack{x_k \sim \mathcal{B}_k \\ \boldsymbol{z}_{-k} \sim \mathcal{B}_{-k}}} \left[\left(\phi_k^{\text{ICE-c}}(h, x_k, \boldsymbol{z}_{-k}) - \phi_k^{\text{PDP-c}}(h, x_k) \right)^2 \right] \\ &= \sum_{k=1}^{d} \mathop{\mathbb{E}}_{x_k \sim \mathcal{B}_k} \left[\mathbb{V}_{\boldsymbol{z}_{-k} \sim \mathcal{B}_{-k}} \left[\phi_k^{\text{ICE-c}}(h, x_k, \boldsymbol{z}_{-k}) \right] \right]. \end{aligned} \tag{37}$$

The intuition behind this loss is illustrated in Figure 7.

Proposition C.1 (**Proposition 3.1**). *The loss employed inside GADGET-PDP is a valid LoA.*

Proof. The function $L_h^{\text{GADGET-PDP}}(\mathcal{B})$ respects the three properties of **Definition 3.1**.

Property 1 When the model is additive, all ICE curves of a given feature are parallel, which implies that the GAGDET-PDP loss is zero.

Property 2 We first express the centered ICE curves in the Interventional Decomposition with $\mathcal{B}_{\text{ind}}$. Our derivation employs the following "annihilation" property (Kuo et al., 2010)

$$v \cap u \neq \emptyset \Rightarrow \mathop{\mathbb{E}}_{\boldsymbol{x}_v \sim \mathcal{B}_{\text{ind},v}} \left[h_{u,\mathcal{B}_{\text{ind}}}(\boldsymbol{x}) \right] = 0. \tag{38}$$

We have

$$\begin{aligned} \phi_k^{\text{ICE-c}}(h, x_k, \boldsymbol{z}_{-k}) &:= h(x_k, \boldsymbol{z}_{-k}) - \mathop{\mathbb{E}}_{x_k \sim \mathcal{B}_{\text{ind},k}} \left[h(x_k, \boldsymbol{z}_{-k}) \right] \\ &= \sum_{u \subseteq [d]} h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k}) - \mathop{\mathbb{E}}_{x_k \sim \mathcal{B}_{\text{ind},k}} \left[\sum_{u \subseteq [d]} h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k}) \right] \\ &= \sum_{u \subseteq [d]} h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k}) - \sum_{u \subseteq [d]} \mathop{\mathbb{E}}_{x_k \sim \mathcal{B}_{\text{ind},k}} [h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k})] \\ &= \sum_{u \subseteq [d]} h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k}) - \sum_{u \subseteq [d] \setminus \{k\}} \mathop{\mathbb{E}}_{x_k \sim \mathcal{B}_{\text{ind},k}} [h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k})] \qquad \text{(Annihilation Property)} \\ &= \sum_{u \subseteq [d]} h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k}) - \sum_{u \subseteq [d] \setminus \{k\}} h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k}) \\ &= \sum_{u \subseteq [d] : k \in u} h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k}). \end{aligned}$$

The centered-PDP can also be expressed in terms of the FD

$$\begin{aligned} \phi_k^{\text{PDP-c}}(h, x_k) &:= \mathop{\mathbb{E}}_{\boldsymbol{z}_{-k} \sim \mathcal{B}_{\text{ind},-k}} \left[\phi_k^{\text{ICE-c}}(h, x_k, \boldsymbol{z}_{-k}) \right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{z}_{-k} \sim \mathcal{B}_{\text{ind},-k}} \left[\sum_{u \subseteq [d] : k \in u} h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k}) \right] \\ &= \sum_{u \subseteq [d] : k \in u} \mathop{\mathbb{E}}_{\boldsymbol{z}_{-k} \sim \mathcal{B}_{\text{ind},-k}} [h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k})] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{z}_{-k} \sim \mathcal{B}_{\text{ind},-k}} [h_{k,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k})] \qquad \text{(Annihilation property)} \\ &= h_{k,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k}). \end{aligned}$$

Thus the GADGET-PDP loss simplifies

$$
\begin{aligned}
L_h^{\text{GADGET-PDP}}(\mathcal{B}) &:= \sum_{k=1}^{d} \mathop{\mathbb{E}}_{\substack{x_k \sim \mathcal{B}_{\text{ind},k} \\ \boldsymbol{z}_{-k} \sim \mathcal{B}_{\text{ind},-k}}} \left[\left(\phi_k^{\text{ICE-c}}(h, x_k, \boldsymbol{z}_{-k}) - \phi_k^{\text{PDP-c}}(h, x_k) \right)^2 \right] \\
&= \sum_{k=1}^{d} \mathop{\mathbb{E}}_{\substack{x_k \sim \mathcal{B}_{\text{ind},k} \\ \boldsymbol{z}_{-k} \sim \mathcal{B}_{\text{ind},-k}}} \left[\left(\sum_{u \subseteq [d] : k \in u} h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k}) - h_{k,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k}) \right)^2 \right] \\
&= \sum_{k=1}^{d} \mathop{\mathbb{E}}_{\substack{x_k \sim \mathcal{B}_{\text{ind},k} \\ \boldsymbol{z}_{-k} \sim \mathcal{B}_{\text{ind},-k}}} \left[\left(\sum_{\substack{u \subseteq [d] : k \in u \\ |u| \geq 2}} h_{u,\mathcal{B}_{\text{ind}}}(x_k, \boldsymbol{z}_{-k}) \right)^2 \right] \\
&= \sum_{k=1}^{d} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{B}_{\text{ind}}} \left[\left(\sum_{\substack{u \subseteq [d] : k \in u \\ |u| \geq 2}} h_{u,\mathcal{B}_{\text{ind}}}(\boldsymbol{x}) \right)^2 \right] && \text{(By feature independence)} \\
&= \sum_{k=1}^{d} \sum_{\substack{u \subseteq [d] : k \in u \\ |u| \geq 2}} \sigma_u^2 = \sum_{\substack{u \subseteq [d] \\ |u| \geq 2}} |u| \sigma_u^2. && \text{(By feature independence and Equation 27)}
\end{aligned}
$$

Thus, when features are independent, GADGET-PDP penalizes $|u|$-way interactions with a weight $w(u) = |u|$. Like the LoA $D(\boldsymbol{\phi}^{\text{PDP}}, \boldsymbol{\phi}^{\text{PFI}})$, higher order interactions are penalized more than low order interactions.

Property 3 We assume that h is additive in feature j and must prove that $L_h^{\text{GADGET-PDP}}(\mathcal{B}_j \times \mathcal{B}_{-j}) = L_h^{\text{GADGET-PDP}}(\mathcal{B}'_j \times \mathcal{B}_{-j})$ for any two distributions $\mathcal{B}_j$ and $\mathcal{B}'_j$ on feature j. We will study the ICE curves of feature j and features $k \neq j$ separately.

Feature j : Because the model is additive in j, the ICE curves and the PDP of feature j will be parallel. Hence the contribution of feature j to the GADGET-PDP loss is null :

$$
\mathop{\mathbb{E}}_{\substack{x_j \sim \mathcal{B}_j \\ \boldsymbol{z}_{-j} \sim \mathcal{B}_{-j}}} \left[\left(\phi_j^{\text{ICE-c}}(h, x_j, \boldsymbol{z}_{-j}) - \phi_j^{\text{PDP-c}}(h, x_j) \right)^2 \right] = 0. \tag{39}
$$

Feature $k \neq j$: Since the Interventional Decomposition is minimal, there will not be any interaction terms h_u with $j \in u$. In that case, the Centered ICE curves of feature $k \neq j$ are

$$
\begin{aligned}
\phi_k^{\text{ICE-c}}(h, x_k, \boldsymbol{z}_{-k}) &:= h(x_k, \boldsymbol{z}_{-k}) - \mathop{\mathbb{E}}_{x_k \sim \mathcal{B}_k} [h(x_k, \boldsymbol{z}_{-k})] \\
&= \sum_{u \subseteq [d]} h_{u,\mathcal{B}}(x_k, \boldsymbol{z}_{-k}) - \mathop{\mathbb{E}}_{x_k \sim \mathcal{B}_k} \left[h_{u,\mathcal{B}}(x_k, \boldsymbol{z}_{-k}) \right] \\
&= h_{j,\mathcal{B}}(x_k, \boldsymbol{z}_{-k}) - \mathop{\mathbb{E}}_{x_k \sim \mathcal{B}_k} \left[h_{j,\mathcal{B}}(x_k, \boldsymbol{z}_{-k}) \right] + \sum_{u \subseteq [d] \setminus \{j\}} h_{u,\mathcal{B}}(x_k, \boldsymbol{z}_{-k}) - \mathop{\mathbb{E}}_{x_k \sim \mathcal{B}_k} \left[h_{u,\mathcal{B}}(x_k, \boldsymbol{z}_{-k}) \right] \\
&= h_{j,\mathcal{B}}(z_j) - h_{j,\mathcal{B}}(z_j) + \sum_{u \subseteq [d] \setminus \{j\}} h_{u,\mathcal{B}}(x_k, \boldsymbol{z}_{-k}) - \mathop{\mathbb{E}}_{x_k \sim \mathcal{B}_k} \left[h_{u,\mathcal{B}}(x_k, \boldsymbol{z}_{-k}) \right] && \text{(Since } k \neq j\text{)} \\
&= \sum_{u \subseteq [d] \setminus \{j\}} h_{u,\mathcal{B}}(x_k, \boldsymbol{z}_{-k}) - \mathop{\mathbb{E}}_{x_k \sim \mathcal{B}_k} \left[h_{u,\mathcal{B}}(x_k, \boldsymbol{z}_{-k}) \right].
\end{aligned}
$$

We have just proven that

$$
\phi_k^{\text{ICE-c}}(h, x_k, \boldsymbol{z}_{-k}) \text{ does } \textbf{not} \text{ depend on } z_j \tag{40}
$$

since we sum over all interactions h_u that exclude feature j. Finally, the total GADGET-PDP loss is

$$\begin{aligned} L_h^{\text{GADGET-PDP}}(\mathcal{B}) &:= \sum_{k\in[d]} \mathop{\mathbb{E}}_{x_k\sim\mathcal{B}_k} \left[\mathbb{V}_{\boldsymbol{z}_{-k}\sim\mathcal{B}_{-k}} \left[\phi_k^{\text{ICE-c}}(h, x_k, \boldsymbol{z}_{-k}) \right] \right] \\ &= \sum_{k\in[d]\setminus\{j\}} \mathop{\mathbb{E}}_{x_k\sim\mathcal{B}_k} \left[\mathbb{V}_{\boldsymbol{z}_{-k}\sim\mathcal{B}_{-k}} \left[\phi_k^{\text{ICE-c}}(h, x_k, \boldsymbol{z}_{-k}) \right] \right] && \text{(Equation 39)} \\ &= \sum_{k\in[d]\setminus\{j\}} \mathop{\mathbb{E}}_{x_k\sim\mathcal{B}_k} \left[\mathbb{V}_{\boldsymbol{z}_{-\{j,k\}}\sim\mathcal{B}_{-\{j,k\}}} \left[\phi_k^{\text{ICE-c}}(h, x_k, \boldsymbol{z}_{-k}) \right] \right]. && \text{(Equation 40)} \end{aligned}$$

This expression does not involve $\mathcal{B}_j$ so it is independent of the manner in which feature j is distributed. The direct implication is that $L_h^{\text{GADGET-PDP}}(\mathcal{B}_j \times \mathcal{B}_{-j}) = L_h^{\text{GADGET-PDP}}(\mathcal{B}'_j \times \mathcal{B}_{-j})$. □

The fact that the GADGET-PDP loss is a LoA implies that it penalizes interactions and so it can potentially reduce the disagreements between PDP/SHAP/PFI explanations. This an interesting result since GADGET-PDP was initially designed with PDP and ICE in mind. If we let $\boldsymbol{R}$ be a $N \times N \times d$ Tensor with components

$$R_{ijk} := h(x_k^{(i)}, \boldsymbol{x}_{-k}^{(j)}), \tag{41}$$

then the empirical counterpart of the GADGET-PDP loss is

$$\widehat{L}_h^{\text{GADGET-PDP}}(S_\Omega) = \frac{1}{|S_\Omega|} \sum_{k=1}^{d} \sum_{i,j\in S_\Omega} \left(R_{ijk} - \frac{1}{|S_\Omega|} \sum_{\ell\in S_\Omega} R_{\ell jk} - \frac{1}{|S_\Omega|} \sum_{m\in S_\Omega} R_{imk} + \frac{1}{|S_\Omega|^2} \sum_{\ell,m\in S_\Omega} R_{\ell mk} \right)^2. \tag{42}$$

Name	Local	Global
H^2_{ij}	N/A	$\Phi^{\text{Inter}}_{ij}(h) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{B}}[\, h_{\{i,j\},\mathcal{B}}(\boldsymbol{x})^2\,]$
SHAP-T	$\phi^{\text{SHAP-T}}_{ij}(h,\boldsymbol{x}) = \begin{cases} h_{i,\mathcal{B}}(\boldsymbol{x}) & \text{if } i=j \\ \sum_{\substack{u\subseteq[d]\\ i,j\in u}} \frac{h_{u,\mathcal{B}}(\boldsymbol{x})}{\binom{\lvert u\rvert}{2}} & \text{if } i\neq j. \end{cases}$	$\Phi^{\text{SHAP-T}}_{ij}(h) = \begin{cases} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{B}}[h_{i,\mathcal{B}}(\boldsymbol{x})^2] & \text{if } i=j \\ \mathbb{E}_{\boldsymbol{x}\sim\mathcal{B}}[\big(\sum_{\substack{u\subseteq[d]\\ i,j\in u}} \frac{h_{u,\mathcal{B}}(\boldsymbol{x})}{\binom{\lvert u\rvert}{2}}\big)^2] & \text{if } i\neq j. \end{cases}$

Table 3: Expressing the various interactions indices in terms of the Interventional Decomposition.

D ADDITIONAL EXPERIMENTS

D.1 Interaction Detection

As was discussed in the main text, FD-Trees should only split along features that interact. But how do we discover features that interact? Several options have been proposed in the literature. For instance, the H^2_{ij} statistic (Friedman and Popescu, 2008) quantifies the interaction between features i and j via

$$\Phi^{\text{Inter}}_{ij}(h) := \mathbb{V}_{\boldsymbol{x}\sim\mathcal{B}}\Big[\, \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{x}_{\{i,j\}}, \boldsymbol{z}_{-\{i,j\}})] - \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{x}_i, \boldsymbol{z}_{-i})] - \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{x}_j, \boldsymbol{z}_{-j})]\,\Big]. \tag{43}$$

Moreover, Shapley-Taylor indices (Sundararajan et al., 2020) generalize SHAP values by providing local attributions $\phi_{ij}(h,\boldsymbol{x})$ for $i,j\in[d]$ that still respect Efficiency

$$\sum_{i=1}^{d}\sum_{j=1}^{d} \phi^{\text{SHAP-T}}_{ij}(h,\boldsymbol{x}) = h(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{z})]. \tag{44}$$

As was done with SHAP, one can compute global importance from Shapley-Taylor indices by averaging the squared amplitudes

$$\Phi^{\text{SHAP-T}}_{ij}(h) := \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{B}}[\, \phi^{\text{SHAP-T}}_{ij}(h,\boldsymbol{x})^2\,]. \tag{45}$$

In Table 3 we express these interaction indices in the Interventional Decomposition. We see that the H^2_{ij} statistic is the variance of the $\{i,j\}$-interaction, and Shapley-Taylor attributes importance to the $\{i,j\}$ interaction by identifying all M-way interactions that involve $\{i,j\}$ and sharing them evenly between the $\binom{M}{2}$ pairs of features involved (Sundararajan et al., 2020).

In our experiments, we detected interactions using the Shapley-Taylor global indices (cf. Equation 45) and not the H^2_{ij} statistic. The reason is that Shapley-Taylor indices $\Phi_{ij}(h)$ quantify the interactions $\{i,j\}$, but also the higher degree interactions u such that $\{i,j\}\subset u$. To compute these indices quickly when explaining tree ensembles, we rely on the work of Laberge et Pequignot (Laberge and Pequignot, 2022) who generalized the Interventional TreeSHAP algorithm to the Shapley-Taylor index.

Adults Figure 8 shows the interactions in the Adults dataset. We note that the strongest interactions involve the features `age`, `capital-gain`, `marital-status`, and `relationship`. Therefore, only these four features were split candidates.

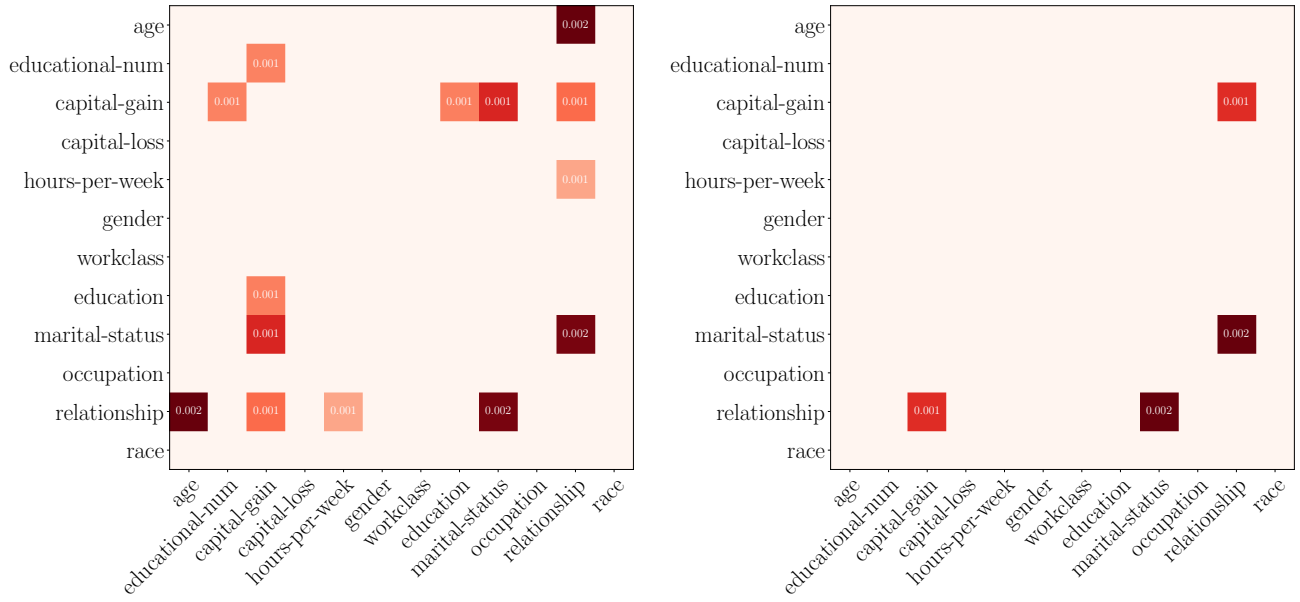
BikeSharing Figure 9 presents the interactions in the BikeSharing dataset. The dominating interactions are amongst the features `hr`, `workingday`, `year`, and `temp`. Hence, these four features were split upon by the FD-Trees.

Marketing Figure 10 highlights interactions in Marketing. The main interactions involve `month`, `day`, `contact`, `pdays`, which will be used by the FD-Trees.

Default-Credit Figure 11 presents the interactions in the Default-Credit dataset. Notable interactions involve `Delay-Sep`, `Delay-Aug`, `Bill-Sep`, and `Bill-Aug`. These features are the four split candidates in FD-Trees.

Kin8nm Figure 12 illustrates interactions in Kin8nm. All features interact with other features so they must all be considered when fitting FD-Trees.

California Figure 13 shows the interactions in the California Housing data. Strong interactions between `Lattitude` and `Longitude` are present. We also see weaker interactions between `Occupancy`, `MedInc`, and `HouseAge`. These five features were used by the FD-Trees to define the regions.

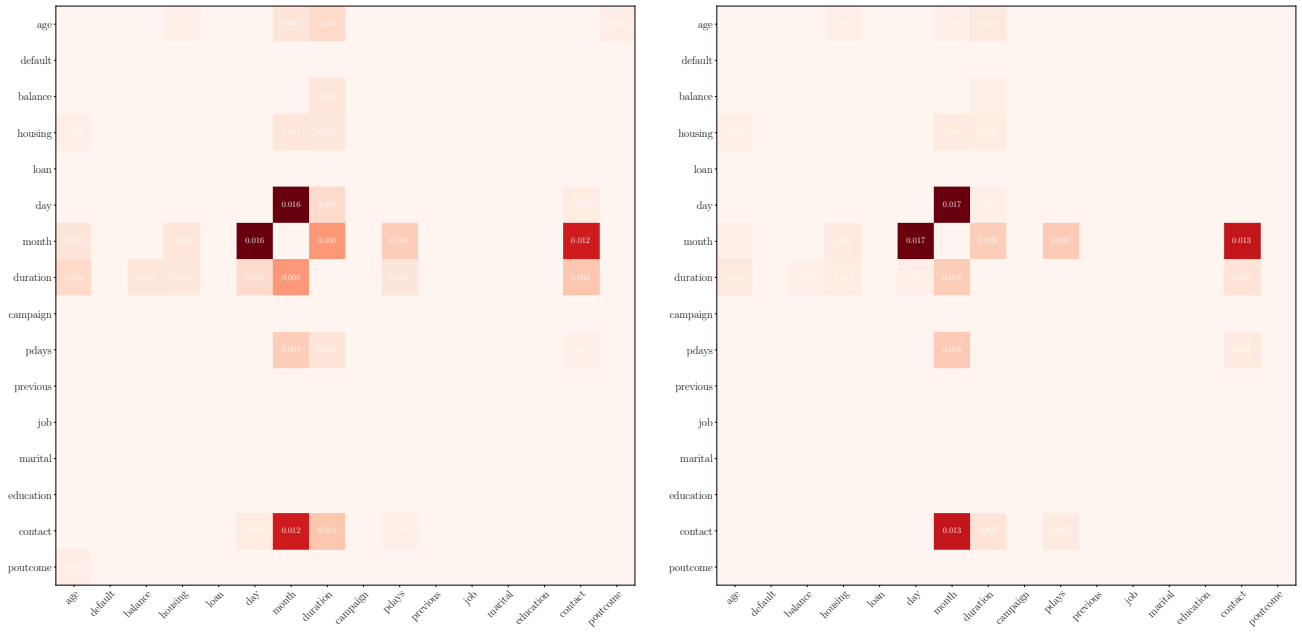


(a) RF (b) GBT

Figure 8: Interaction Indices on Adult.

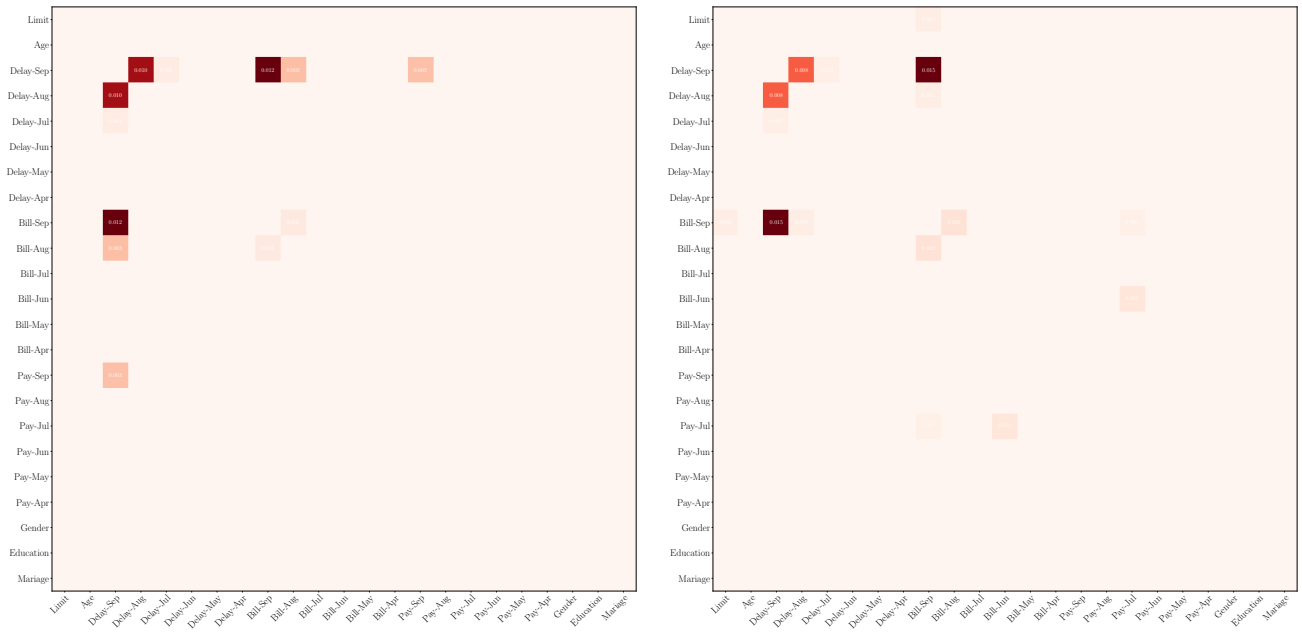
(a) RF (b) GBT

Figure 9: Interaction Indices on BikeSharing.



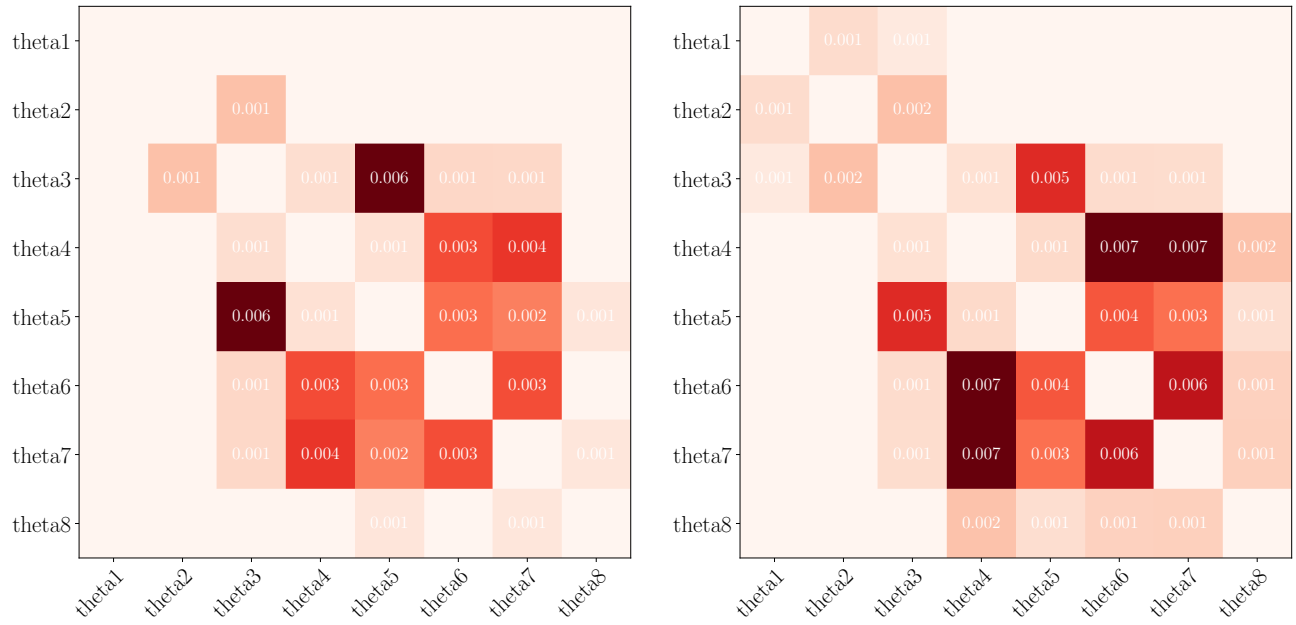
(a) RF (b) GBT

Figure 10: Interaction Indices on Marketing.



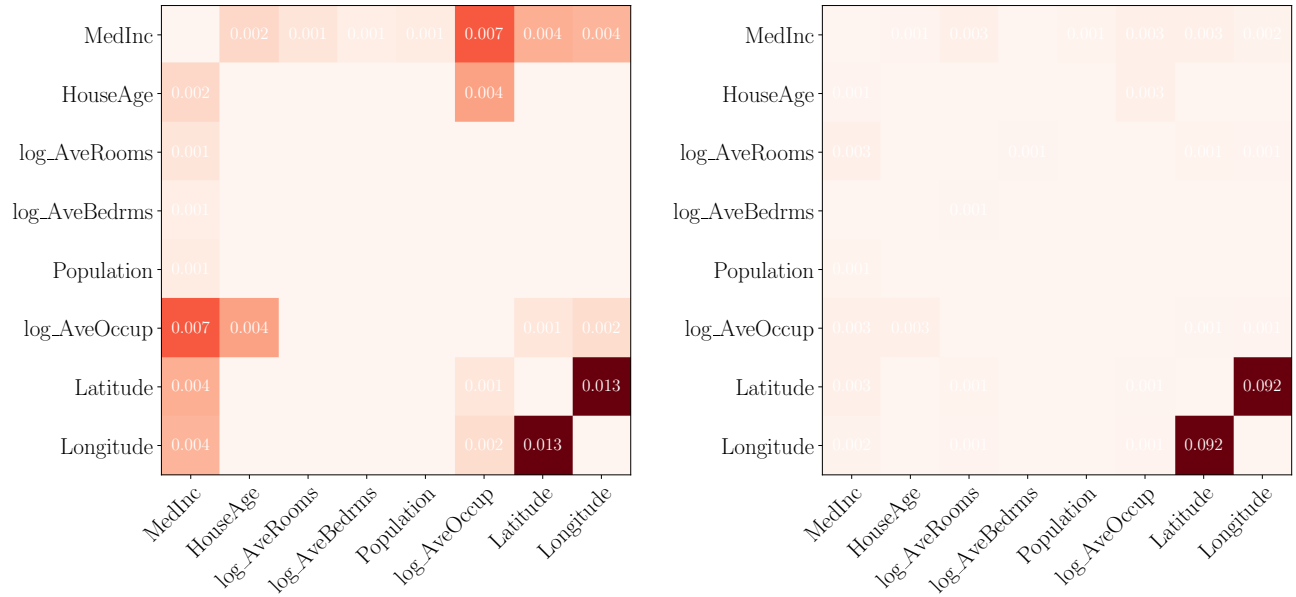
(a) RF (b) GBT

Figure 11: Interaction Indices on Default-Credit.



(a) RF (b) GBT

Figure 12: Interaction Indices on Kin8nm.



(a) RF (b) GBT

Figure 13: Interaction Indices on California.

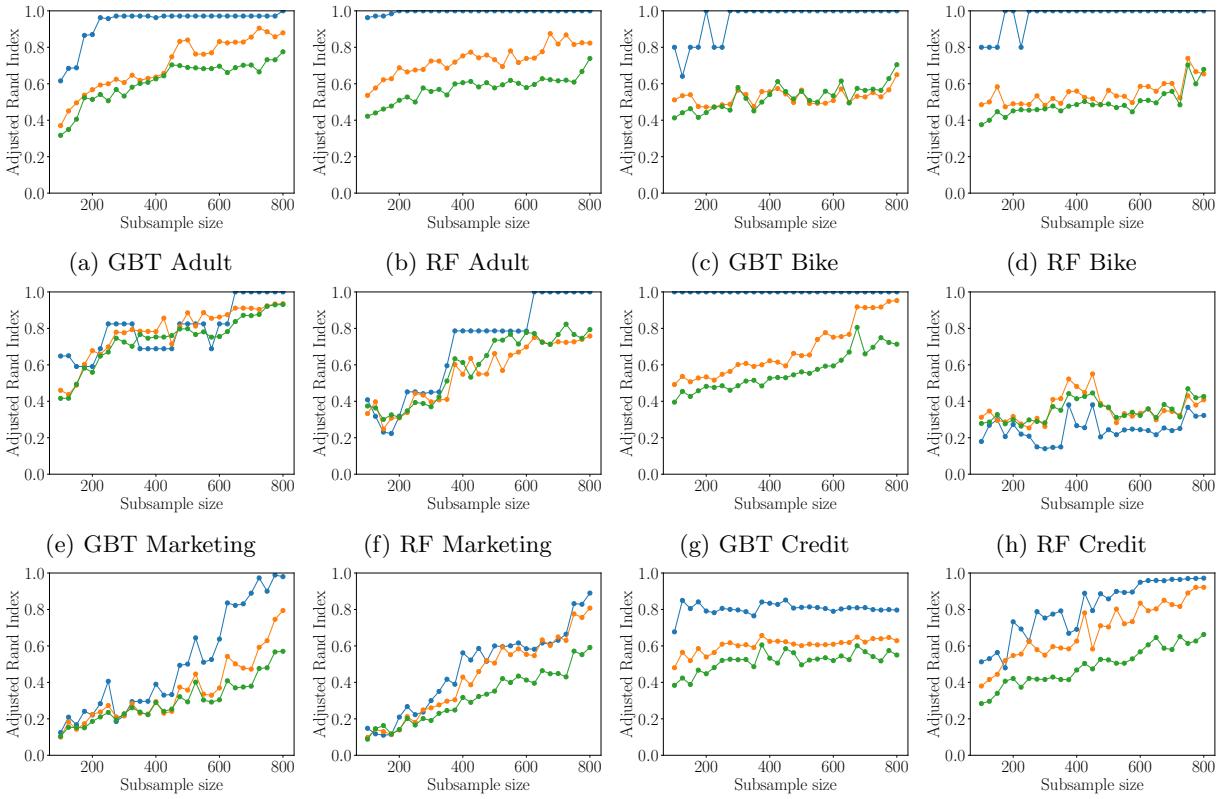


Figure 14: Stability of the Partitions given by FD-Trees as a function of the subsample size.

D.2 Stability of Partitions

The tensor $\boldsymbol{H}$ required to train FD-Trees has a size of $N \times N \times d$. Thus, on any realistic dataset it is crucial to subsample $N' \ll N$ data points and use them to approximate $\boldsymbol{H}$. However, this introduces stochasticity in the training of FD-Trees since different subsamples may lead to different trees. Ideally, N' should be large enough to lead to stable partitions $(\Omega_1, \Omega_2, \ldots, \Omega_M)$ on multiple reruns. Yet, it should also not be too large to avoid the $\mathcal{O}(N^2)$ time and space complexities.

To assess the stability of partitions, we repeat the following methodology 10 times: 1) subsample N' data points, 2) compute $\boldsymbol{H}$, 3) grow a full FD-Tree using $\boldsymbol{H}$, 4) return the associated partition. The Rand Index (Hubert and Arabie, 1985) can then be used to measure the *similarity* between any pairs of partitions resulting from these repeated reruns. The Rand Index considers all pairings of data points $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$. The two partitions are said to agree on the pair if either 1) $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$ are in the same group for both partitions, or 2) $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$ are in separate groups for both partitions. The Rand Index is then the ratio of agreeing pairs to the total number of pairs. However, it is best to employ the Adjusted Rand Index (Hubert and Arabie, 1985) which normalizes the metric to account for chance. Thus, random partitions will have an index close to zero.

In Figure 14, we present the Adjusted Rand Index as a function of subsample size. A general trend is that partitions stabilize when more samples are used to train FD-Trees. Importantly, the partitions of depth-1 FD-Trees are extremely stable on Adults and BikeSharing. For the other datasets, it takes multiple samples before the first split stabilizes. Based on these results, we advocate that subsampling $N' = 600$ is reasonable on all these datasets, although kin8nm could benefit from including more samples. Still, to keep the experimental setup simple, we subsample 600 data points for all datasets and models.

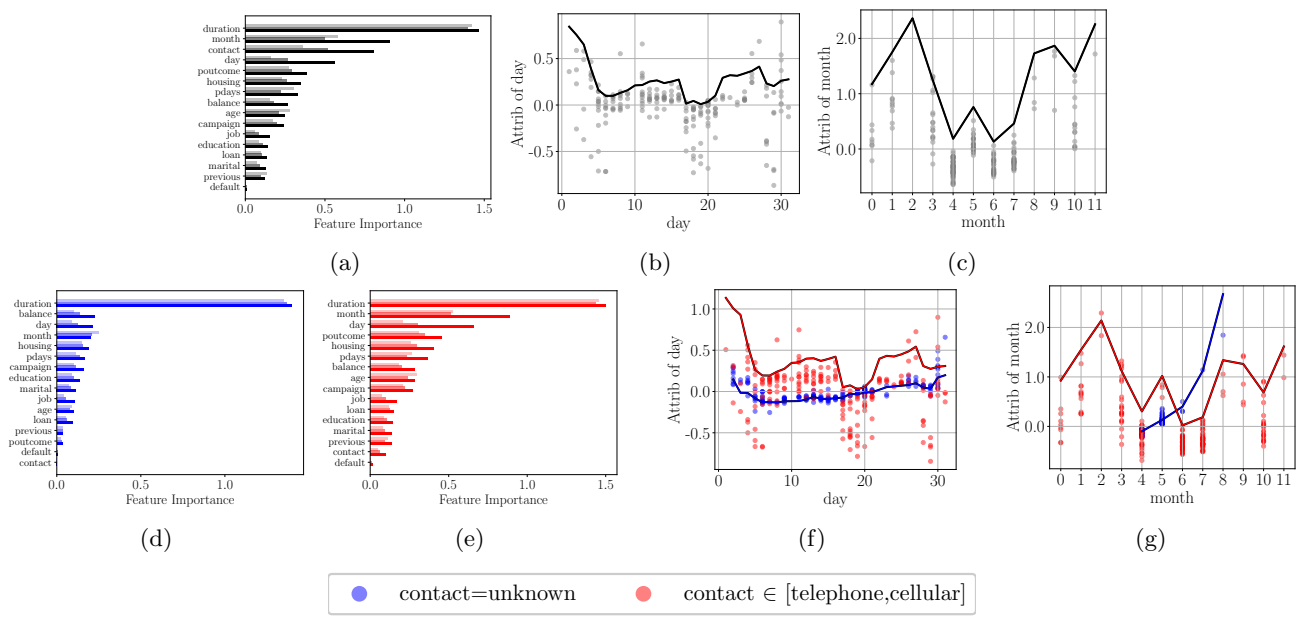


Figure 15: Marketing (Part 1). Lines are PDPs while points are SHAP values. The top row shows the explanations when the background is set to the whole data distribution. The bottom row presents the regional explanations extracted from a FD-Tree of depth 1.

D.3 Qualitative Results

D.3.1 Marketing

The Marketing dataset available on the UCI Repository[4] describes the marketing campaign of a Portuguese banking institution. Each instance corresponds to a distinct phone call and the binary label encodes whether or not the client subscribed to a term deposit.

Figure 15 (top row) shows the global/local explanations of a GBT when the whole dataset is used as the background distribution. We note strong disagreements between the PDP/SHAP/PFI global feature importance of `month`, `contact`, and `day`. Moreover, the PDP local attributions are poor estimates of the SHAP values. These results highlight the presence of strong interactions involving these three features. To reduce interactions and increase agreement between post-hoc explainers, we partition the input space with a FD-Tree trained with PDP-PFI loss. This loss was used since it dominates other losses on this dataset according to Table 2. The first split of the tree separates `contant=unknown` from `contact`$\in$[`telephone`, `cellular`]. This was systematically the first split for any FD-Tree. Figure 15 (bottom row) shows explanations on the two regions identified by the split. We note that there are no longer strong disagreements w.r.t the importance of `contact`. Still, there remain disagreements between the global feature importance of `month` and `day` when `contact`$\in$[`telephone`, `cellular`].

These two features are split upon when considering deeper FD-Trees, see Figure 16. Looking at the top-row reveals that the next splits were made w.r.t `month`. As a result, the local feature attributions of `day` are starting to highlight some interesting heterogeneity. The bottom row highlights more heterogeneity in the local feature attribution of `day` given different `month` intervals. There are even sub-regions of `contact`$\in$[`telephone`, `cellular`] where the PDP/SHAP local feature attributions have strong agreements: region $1.00 <$ `month` ≤ 5.00 and region $5.00 <$ `month` ≤ 7.00. For the two other sub-regions of `contact`$\in$[`telephone`, `cellular`] (region `month` ≤ 1.00 and region `month` > 7.00), the PDP remains a poor estimate of the SHAP values and hence it may be safer to abstain from explaining those regions.

[4] `https://archive.ics.uci.edu/dataset/222/bank+marketing`

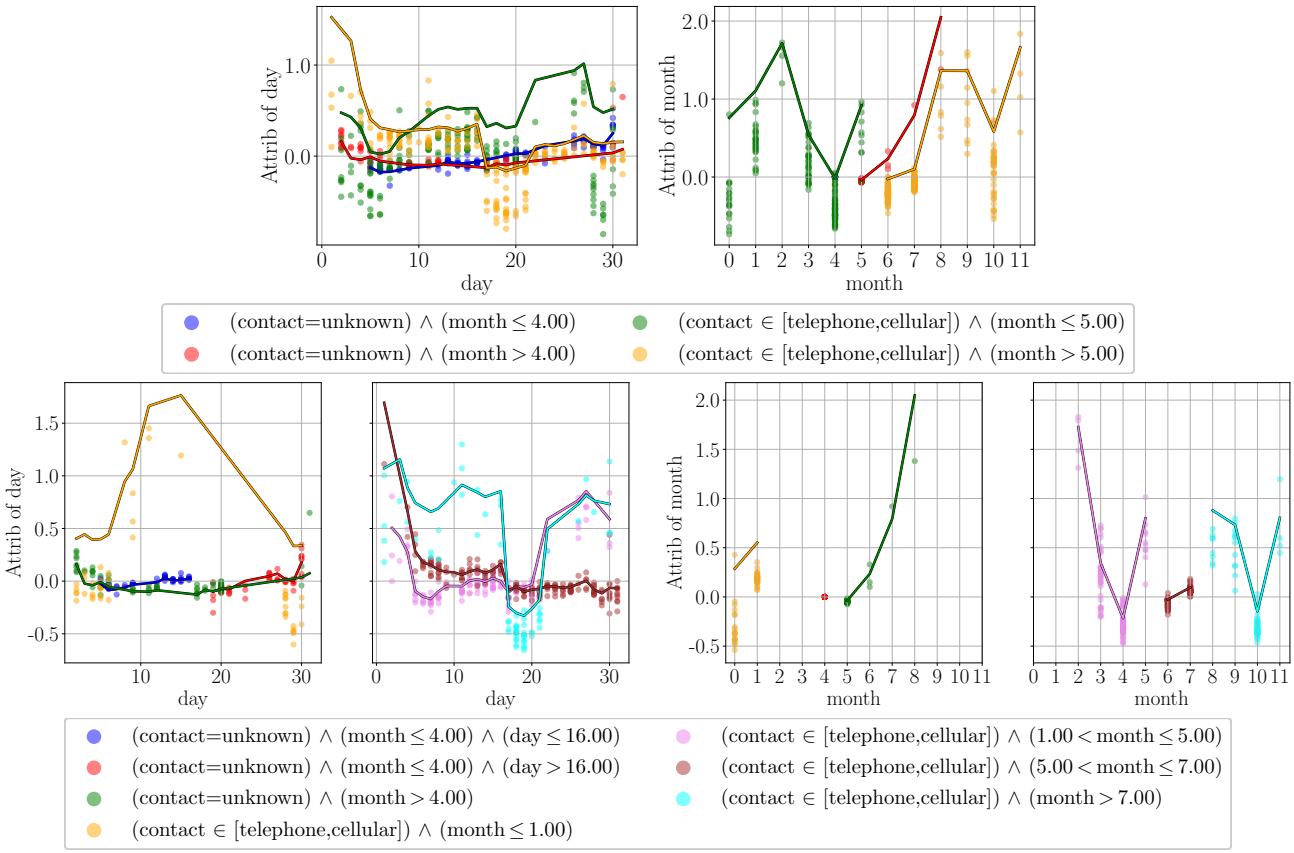


Figure 16: Marketing (Part 2). Lines are PDPs while points are SHAP values. From the top to the bottom row we show the regional explanations extracted from a FD-Tree of depth= 2, 3.

D.3.2 Default-Credit

The Default-Credit dataset available on the UCI Repository[5] aims at predicting if clients of a Taiwanese bank will default on their credit. The data contains records of 30K individuals and 23 features related to past payments/bills/delays and demographic characteristics.

Figure 17 (top row) shows the global/local explanations of a GBT when the whole dataset is used as the background distribution. According to Figure 17 (a), the global feature importance of `Delay-Sep`, `Bill-Sep`, and `Bill-Aug` disagree considerably. Additionally, as seen in Figure 17 (b) the PDP local attributions of `Delay-Aug` is a poor estimate of the SHAP values. Both of these observations are caused by strong feature interactions. To aim at reducing interactions, we partition the input space with a FD-Tree trained with CoE loss. The corresponding first split differentiates `not Delay-Sep` from `Delay-Sep`. This first split was consistent across all FD-Trees fitted on Default-Credit. Investigating Figure 17 (e), the regional feature attributions of `Delay-Aug` depend heavily on whether or not there was a delay in September payments. In fact, the PDP of `Delay-Aug` is an increasing function of `Delay-Aug` when there is no September delay. The PDP is a decreasing function when there is a September delay. From Figure 17 (f), the regional attributions of `Bill-Sep` behave differently near the origin depending on whether there was a delay in September payments. In fact, for individuals with a delayed payment in September, the attribution of having a small September bill is extremely negative.

The next splits separate individuals with positive August bills from individuals with null (or negative) bills, see Figure 17 (bottom row). We again note very diverse model behaviors depending on the region, especially for the feature `Delay-Aug`, see Figure 17 (h). Given our limited knowledge of the data, we cannot fully explain these complex model behaviors. Indeed, it is not clear to us why bills could be negative and what link that could have with delayed payments. Yet, our goal with this analysis is **not** to demystify the Default-Credit dataset. Rather,

[5] `https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients`

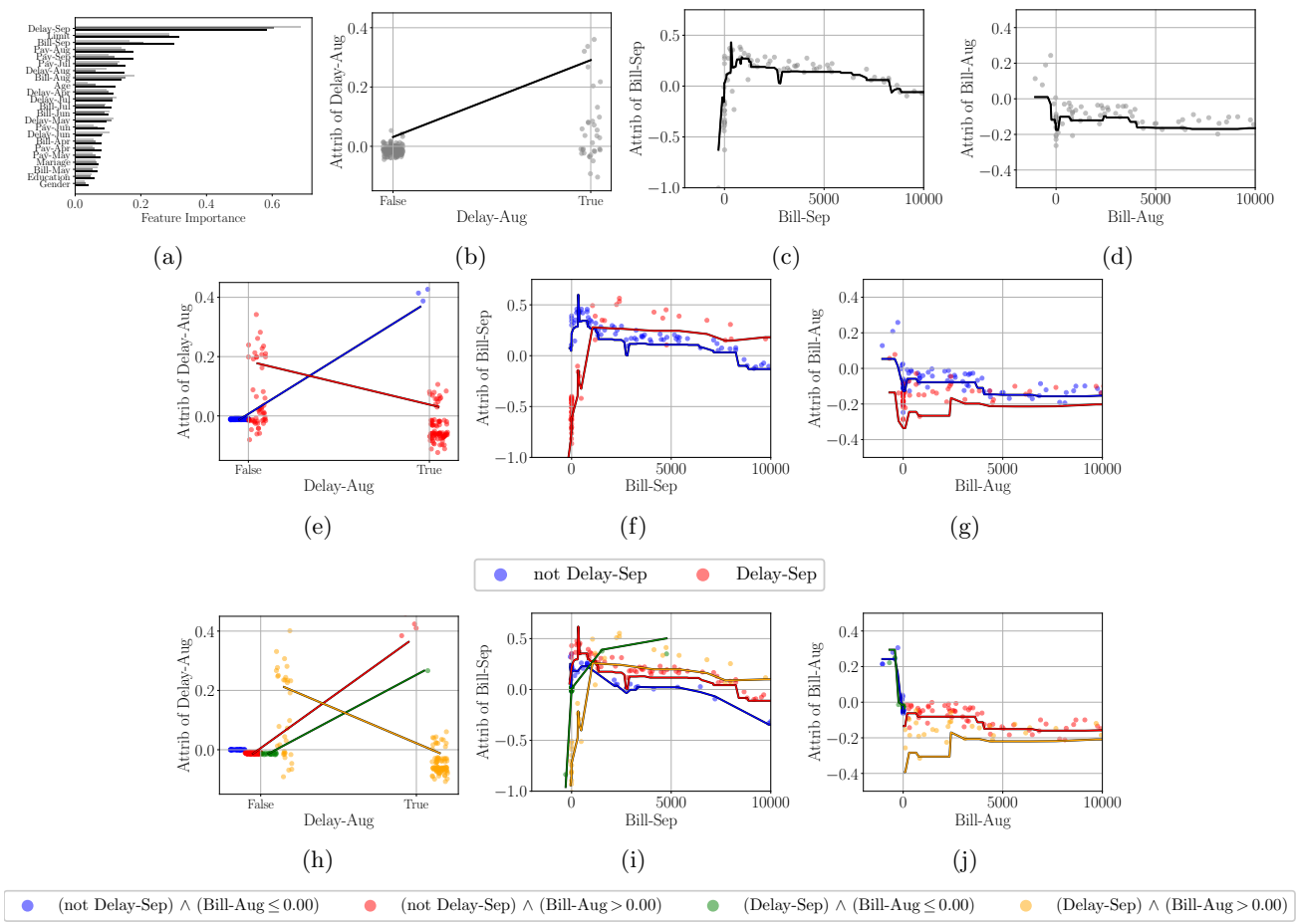


Figure 17: Default Credit. Lines are PDPs while points are SHAP values. The top row shows the explanations when the background is set to the whole data distribution. The middle row presents the regional explanations extracted from a FD-Tree of depth 1. The bottom row illustrates regional explanations extracted via a FD-Tree of depth 2.

it is simply to show that regional explanations extracted from FD-Tree leaves can highlight heterogeneous model behaviors.

D.3.3 Kin8nm

The kin8nm dataset is a realistic simulation of the forward dynamics of an 8-link robot arm. The regression task is to predict the distance of the end-effector from a target based on 8 input features representing angles from the robot arm joints. Our knowledge of this dataset is extremely limited since we cannot interpret what each angle represents. Nonetheless, this dataset was chosen because we hypothesize that it contains strong interactions since the effects of varying an angle should depend on the other angles. As evidenced by Figure 12, there are indeed strong interactions between certain angles. We go further and visualize the Shapley-Taylor attributions $\phi_{ij}^{\text{SHAP-T}}(h, \boldsymbol{x})$ in Figure 18. Most interactions ressemble XOR functions where the effect of one angle changes drastically if the other angle is positive or negative.

These strong interactions can also be observed if we compute global feature importance and local feature attributions using the whole dataset as background, see Figure 19. Indeed, according to Figure 19 (a), the global importance of theta4, 5, 6, and 7 are highly uncertain since various explainers give them drastically different importance. Strong disagreements are also apparent in the PDP/SHAP local feature attributions in Figure 19 (b)-(c)-(d).

To reduce these disagreements we fitted a FD-Tree with the GADGET-PDP objective. Figure 20 shows the

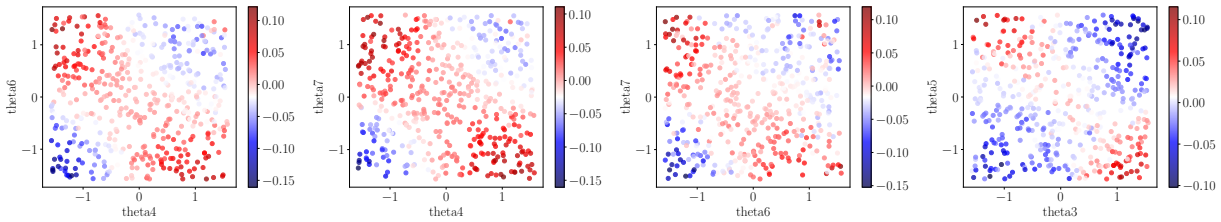
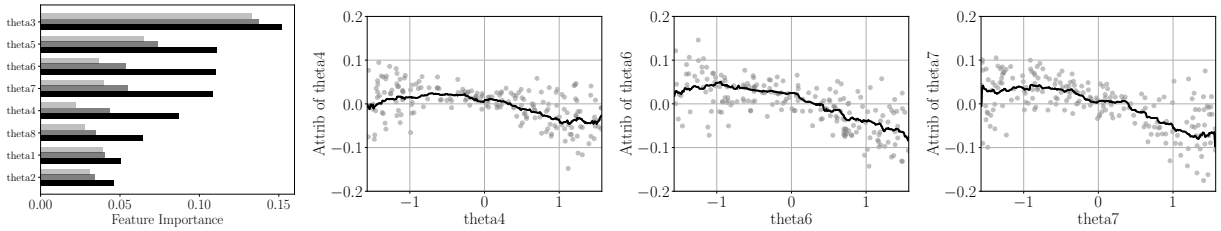


Figure 18: Kin8nm. Local Shapley-Taylor attribution for angles that interact most.



(a) (b) (c) (d)

Figure 19: Kin8nm part 2. (a) The global feature importance. (b)-(c)-(d) show the PDP (line) and SHAP (points) local feature attributions.

global feature importance in each of the 8 regions identified by a FD-Tree of depth 3. We first note increased agreement between the global feature importance in regions (b) and (h). However, there remain disagreements in other regions. For example, the importance of theta4 remains uncertain in regions (c) and (e). This means that 8 regions may not be enough for this dataset.

We also investigate local feature attributions in Figure 21. Crucially, the model explanations vary highly depending on the region. For instance, the attribution of theta4 depends on whether or not theta6 and theta7 are positive or negative. For the other features, it is more difficult to see such a general trend emerge. But we still observe very different behaviors depending on the region: sometimes the attribution increases, decreases, or remains constant. This diversity of trends is hidden when using the whole dataset as the background.

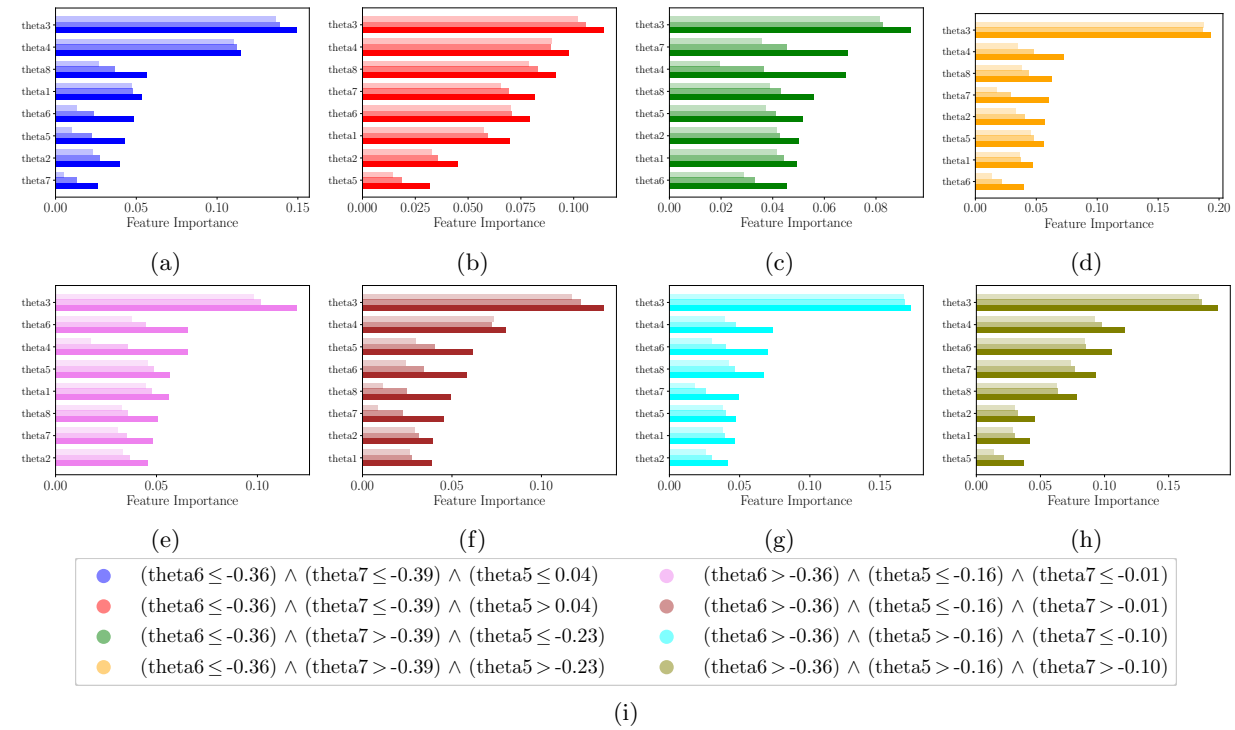


Figure 20: Kin8nm part 3. Feature Importance on the 8 regions identified by a FD-Tree.

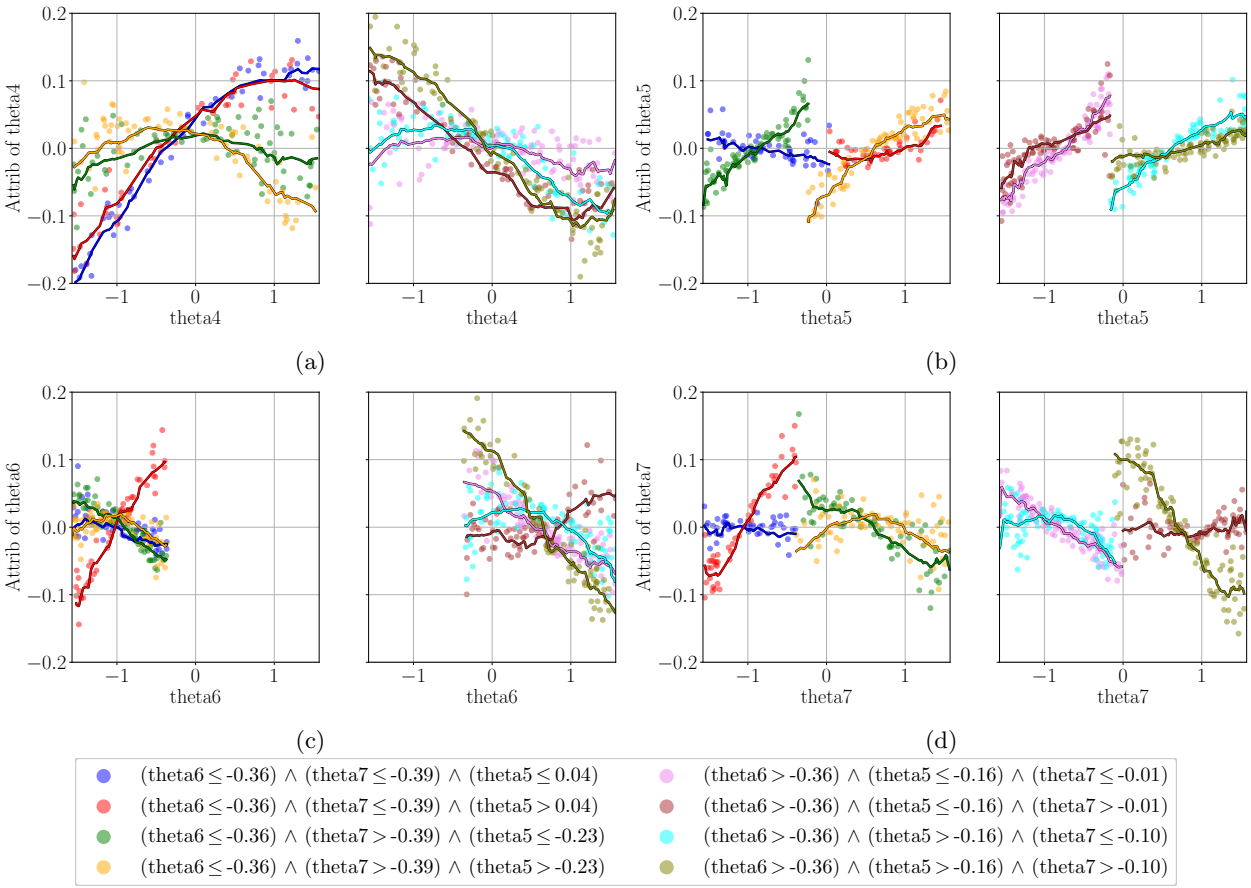


Figure 21: Kin8nm part 4. Local Feature Attribution on the 8 regions identified by a FD-Tree.

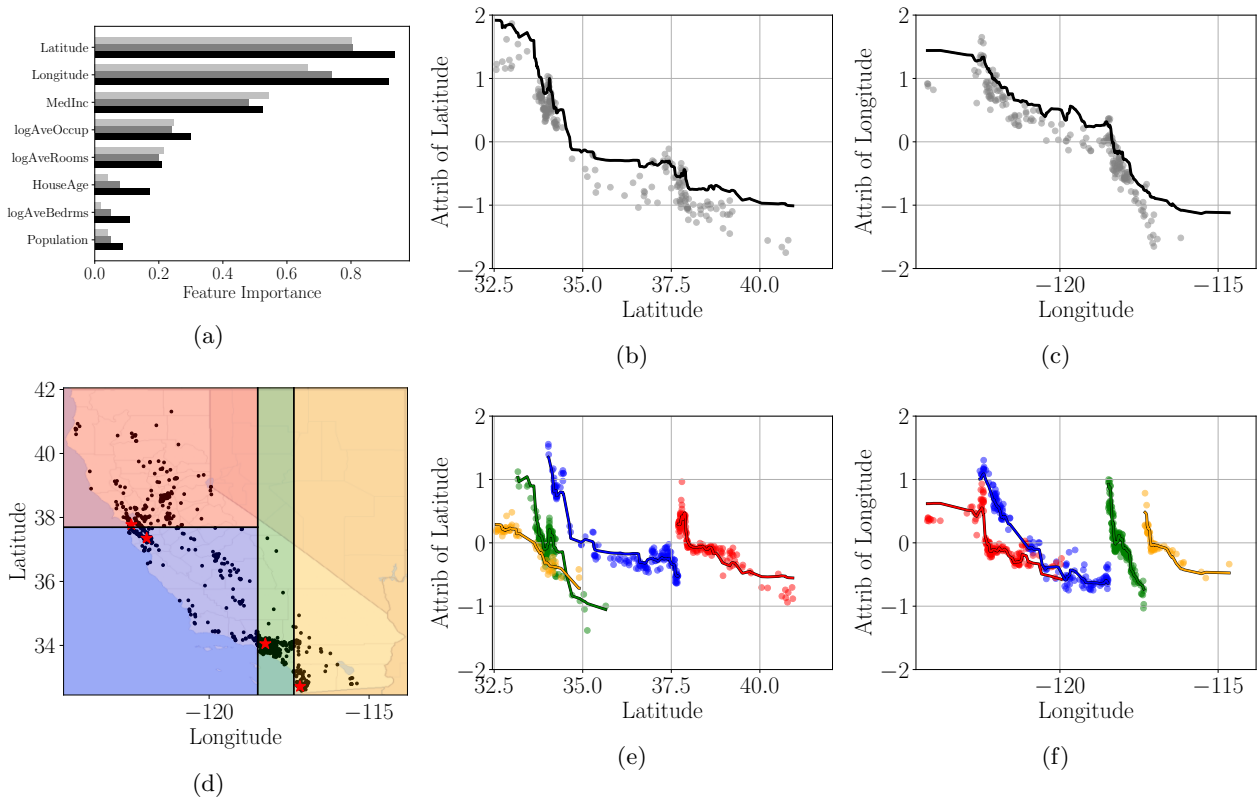


Figure 22: California. The top row shows the global (a) and local (b)&(c) explanations when the background is set to the whole data distribution. Lines are the local PDP while points are the local SHAP values. (d) The state of California is split by a FD-Tree into four regions shown in color. The major cities of Los Angeles, San Francisco, San Diego, and San Jose are shown as red stars. (e)&(f) The local PDP/SHAP explanations extracted from these four regions.

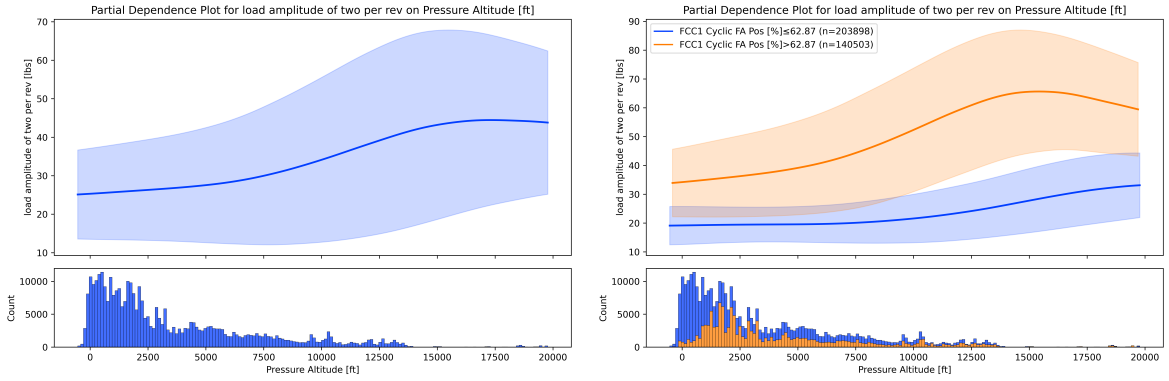
D.3.4 California

The California dataset available on the Statlib Repository[6] consists in predicting the median house value in a California block from 1990. The input features involve demographic characteristics aggregated over each block as well as the `longitude` and `lattitude` of the respective blocks. We present the results of fitting a GBT on this dataset.

As highlighted in Figure 13, the strongest interactions involve features `longitude` and `lattitude`. In Figure 22 (a), we observe stronger disagreements regarding the global importance of these two features. Additionally, according to Figure 22 (b)&(c), the PDP local explanation is a poor estimate of the SHAP values. Based on all of these observations, it is difficult to explain the effect of each separate coordinate on the model. The impact of varying `longitude` depends on `lattitude` and vice versa. Nonetheless, we hypothesize that FD-Trees can be used to split up California into regions where the role of both coordinates is more additive.

We fitted a depth-2 FD-Tree with the PDP-PFI objective because it dominates the other objectives (Table 2). All of the splits conducted by this tree were applied to the `longitude` and `lattitude` features. Consequently, the tree leaves can be visualized over a map of California, see Figure 22 (d). Figure 22 (e)&(f) presents the PDP and SHAP explanations whose background distribution is restricted to a single leaf. The disagreements between the two explainers are greatly reduced as a result. As a final note, the large cities of Los Angeles, San Francisco, San Diego, and San Jose are shown on the California map as red stars. Interestingly, each city belongs to a separate FD-Tree leaf. This result is not a coincidence since these splits were consistent across various FD-Trees trained on GBTs with the CoE and PDP-PFI objectives.

[6] `http://lib.stat.cmu.edu/datasets/`



(a) On the whole test set.

(b) On each of the two regions learned.

Figure 23: Partial dependence plots for load amplitude corresponding to two rpm on the feature Pressure altitude. The shaded area represents the standard deviation, while the solid line represents the mean, namely the PDP value. Histograms show the marginal distribution of the data for Pressure altitude.

E INDUSTRIAL USE-CASE: PITCHLINK LOAD PREDICTION

In this section, we provide a quick glimpse into the practical application of our methodology and techniques in an industrial context. As a collaborative effort within the DEEL[7] project, we worked in conjunction with experts from Bell Textron to develop a machine learning model geared towards the prediction of pitchlink load signals. These predictions are based on data obtained from the flight data recorder (FDR) during helicopter operations. Pitchlinks are critical components of helicopters that are subjected to significant stress and necessitate frequent inspections and replacements. Our primary objective throughout this project was to explore the potential use of machine learning for load monitoring in real-world maintenance procedures.

The proprietary data collected by Bell Textron consists of many hours (about 74 hours in total) of load measurements on a pitchlink along with 15 parameters typically collected by a flight data recorder (e.g. Pressure altitude, Computed Airspeed, cyclic positions, ...). This time series data underwent preprocessing steps, including data cleaning, normalization, and feature extraction, to ensure that it was in a suitable format for learning a model. While we cannot divulge at this stage the exact preprocessing steps, these were tailored to take advantage of the specific characteristics of the load signal for the purpose of machine learning.

Fully connected networks with ReLU activations were trained for a regression task using a proprietary labeled dataset pairing FDR parameters with corresponding load signals. We assessed the models' performance using industry-standard metrics, such as Root Mean Square Error (RMSE), and others suitable for regression tasks. These metrics helped us evaluate the models' accuracy and reliability in predicting load signals during flights.

Our models exhibited significant promise in predicting load signals during helicopter flights based on FDR parameters. This has substantial implications for the potential use of machine learning in real-time load monitoring for maintenance purposes. We emphasize that the feasibility of certification remains a critical step in confirming its practical application. In particular, the European Union Aviation Safety Agency (EASA) in the Issue 2 of its Concept Paper on Artificial Intelligence (AI) and Machine Learning (ML) (Agency., 2023) identifies the following objective "EXP-03: The applicant should identify and document the methods at AI/ML item and/or output level satisfying the specified AI explainability needs". Focusing on global explanations for domain experts, we find that our approach using FD-Trees provided useful insights into the model behavior that were not revealed by other available methods. In particular, FD-Trees revealed that our model exhibits several interactions among flight parameters. Moreover, FD-Trees led to the discovery certain flight conditions for which global explanations of our model are more reliable.

In Figure 23, we present results of the analysis our models pertaining to the prediction of the amplitude of a key frequency of the load signal. A feature importance analysis on the whole test set revealed that Computed Airspeed, FCC1 Cyclic FA Pos and Pressure Altitude are among the most important features for predicting this

[7] `https://deel.quebec`

amplitude. As an example, we fitted a depth-1 FD-Tree using the CoE objective on a test sample of size 2048 that resulted in a split on the forward position of the cyclic control (FCC1 Cyclic FA Pos) at the value 62.87%. We present the Partial Dependence Plots of Pressure Altitude in Figure 23. It appears that the dependence on Pressure Altitude is more important when the cyclic control is tilted forward (value above 62.87%). Given that FCC1 Cyclic FA Pos is strongly positively correlated with Computed Airspeed, this seems to suggest that the load amplitude of this key frequency is more affected by Pressure Altitude when flying at high speed, an interpretation that was validated by experts from Bell Textron.