Partial Order in Chaos: Consensus on Feature Attributions in the Rashomon Set

Gabriel Laberge, Yann Pequignot, Alexandre Mathieu, Foutse Khomh, Mario Marchand

Journal of Machine Learning Research 24 (2023) 1-50 Submitted 2/23; Revised 8/23; Published 12/23

Partial Order in Chaos: Consensus on Feature Attributions in the Rashomon Set.

Gabriel Laberge[1] GABRIEL.LABERGE@POLYMTL.CA
Yann Pequignot[2] YANN.PEQUIGNOT@IID.ULAVAL.CA
Alexandre Mathieu[2] ALEXANDRE.MATHIEU.7@ULAVAL.CA
Foutse Khomh[1] FOUTSE.KHOMH@POLYMTL.CA
Mario Marchand[2] MARIO.MARCHAND@IFT.ULAVAL.CA
[1]Génie Informatique et Génie Logiciel, Polytechnique Montréal
[2]Institut Intelligence et Données, Université de Laval à Québec

Editor: Pradeep Ravikumar

Abstract

Post-hoc global/local feature attribution methods are progressively being employed to understand the decisions of complex machine learning models. Yet, because of limited amounts of data, it is possible to obtain a diversity of models with good empirical performance but that provide very different explanations for the same prediction, making it hard to derive insight from them. In this work, instead of aiming at reducing the under-specification of model explanations, we fully embrace it and extract logical statements about feature attributions that are consistent across all models with good empirical performance (*i.e.* all models in the Rashomon Set). We show that **partial** orders of local/global feature importance arise from this methodology enabling more nuanced interpretations by allowing pairs of features to be incomparable when there is no consensus on their relative importance. We prove that every relation among features present in these partial orders also holds in the rankings provided by existing approaches. Finally, we present three use cases employing hypothesis spaces with tractable Rashomon Sets (Additive models, Kernel Ridge, and Random Forests) and show that partial orders allow one to extract consistent local and global interpretations of models despite their under-specification.

Keywords: XAI, Feature Attribution, Under-Specification, Rashomon Set, Uncertainty

1. Introduction

The Machine Learning (ML) framework has proven to be an essential tool in many data-intensive domains such as software engineering, medicine, and cybersecurity (Esteves et al., 2020; Kaieski et al., 2020; Salih et al., 2021). However, the lack of interpretability of complex models is still an important hurdle to their applicability. For this reason, various model-agnostic techniques such as LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), and Integrated/Expected Gradient (IG/EG) (Sundararajan et al., 2017; Erion et al., 2021) have recently been developed to provide explanations of model decisions in the form of local feature attributions. These attributions are meant to indicate the contribution (positive/negative/null) of individual features toward a model prediction, and their magnitudes (positive/null) can be used to rank features in order of importance. As researchers and practitioners have started to apply these model-agnostic explanations to real-world set-

tings, it has become apparent that they are subject to variability. First, given a fixed model, re-running the explainer can yield different local feature attributions (Visani et al., 2020; Slack et al., 2021; Zhou et al., 2021). Second, retraining the model can induce different local explanations for the same decisions (Fel et al., 2021; Shaikhina et al., 2021; Schulz et al., 2021). This phenomenon, known as under-specification, arises when one employs a rich hypothesis space containing various models that all fit the data while having very different behaviors (D'Amour et al., 2020).

In this work, we focus on uncertainty induced by the model under-specification, while controlling the variability arising from the explainer. Current literature addresses this uncertainty by aggregating local explanations from an ensemble of independently trained models. The aggregation is either conducted by averaging the models (Shaikhina et al., 2021), or averaging the local feature importance ranks (Schulz et al., 2021). We find that, although these methods provide a single local feature attribution to explain all models, it is unclear what statements practitioners are allowed to make with confidence using said explanation.

Our characterization of explanations uncertainty departs from the current ones by focusing on *statements* about local/global feature attribution. Our motto in this context of model under-specification is: *only consider statements on which all models with good performance agree* Concretely, we are going to work with the set of all models with an empirical loss at most ϵ, or equivalently, with all models in the Rashomon Set (Fisher et al., 2019). At a fixed tolerance ϵ, local/global feature attribution statements on which there is a consensus in the Rashomon Set form **partial** orders, instead of the total orders typically used to rank features. Partial orders have the advantage to enable safer interpretation by allowing two features to be incomparable, which occurs when two models in the Rashomon Set disagree on their relative importance. In such cases, we abstain from claiming that one feature is more important than the other and let practitioners study both features and decide to modify whichever feature is most actionable. Here is a brief summary of the contributions of this work:

1. We identify local/global feature attribution *statements* on which there is a perfect consensus across all models with an empirical loss at most ϵ (*i.e.* all models in the Rashomon Set). These statements result in partial orders, which differ from the total orders commonly used to visualize feature attributions. Our methodology currently supports the Rashomon Sets of Additive Regression, Kernel Ridge Regression, and Random Forests.

2. We prove that if feature i is locally more important than feature j according to our partial orders, then the same relation holds in the total rankings proposed by Shaikhina et al. (2021); Schulz et al. (2021). This property establishes that our approach based on partial order is conservative over these total ranking approaches in the sense that it differs from them only by dismissing some relations among features that are deemed uncertain. This is a desirable property given the lack of ground truth in explainability, which restricts quantitative comparisons between competing techniques.

3. We finally present empirical evidence on three open-source datasets that our partial orders are indeed more cautious than total orders, while still conveying important

information about the predictions. Each use-case employs a different class of models to better highlight the versatility of our framework.

The rest of the paper is structured as follows: **Section 2** introduces Machine Learning notation, local/global feature attributions, and the problem of model under-specification, **Section 3** presents a toy-example that serves as the motivation behind our method, **Section 4** discusses our methodology for asserting consensus in the Rashomon Set, while **Sections 5, 6, & 7** apply the methodology to Additive models, Kernel Ridge Regression, and Random Forests respectively. Finally, **Section 8** discusses the results and **Section 9** concludes the paper.

2. Background & Related Work

2.1 Machine Learning Notation

In supervised Machine Learning (ML) settings, we work with an input space $\mathcal{X} \subseteq \mathbb{R}^d$, a target space $\mathcal{Y}$, an hypothesis space $\mathcal{H} : \mathcal{X} \to \mathcal{Y}'$, and a loss function $\ell : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_+$. We shall refer to each individual function $h \in \mathcal{H}$ as a model or a hypothesis. For parametric hypothesis spaces $\mathcal{H} = \{h_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^p\}$, each realization of the parameters $\boldsymbol{\theta}$ is a different model/hypothesis. We suppose there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ from which examples from a dataset $S = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^N \sim \mathcal{D}^N$ are sampled iid. The ultimate goal of supervised ML is to find a model $h^\star \in \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h)$, with minimal population loss $\mathcal{L}_{\mathcal{D}}(h) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(h(\boldsymbol{x}), y)]$. However, since the data-generating distribution $\mathcal{D}$ is unknown, we cannot compute the population loss $\mathcal{L}_{\mathcal{D}}(h)$ and must resort to studying the empirical loss on the dataset S

$$\widehat{\mathcal{L}}_S(h) := \frac{1}{N} \sum_{i=1}^{N} \ell(h(\boldsymbol{x}^{(i)}), y^{(i)}), \tag{1}$$

which can be minimized over $\mathcal{H}$ to get an estimate $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} \widehat{\mathcal{L}}_S(h)$ of $h^\star$. In this work, we study the hypothesis spaces $\mathcal{H}$ of Additive Splines (Hastie et al., 2009, Chapter 5), Kernel Ridge Regression (Mohri et al., 2018, Chapters 6 & 11) and Random Forests (Breiman, 2001a). The two loss functions ℓ that are considered are the squared loss $\ell(y', y) = (y' - y)^2$ for a continuous target $\mathcal{Y} \subseteq \mathbb{R}$ and the 0-1 loss $\ell(y', y) = \mathbb{1}(y' \neq y)$ for a binary target $\mathcal{Y} = \{0, 1\}$.

2.2 Feature Attribution

The ML paradigm has been successful in tackling tasks where traditional programming methods fail. Still, the lack of transparency of some state-of-the-art models such as Random Forests and Multi-Layered Perceptrons prohibits their wide-spread application (Arrieta et al., 2020). To meet this novel challenge, the community of eXplainable Artificial Intelligence (XAI) has recently been growing with the ambition of *explaining* black box models. In this paper, by *explaining*, we mean asking a *contrastive question* about the model and then *answering* said question.

A *contrastive question* takes the form : why is the model output $h(\boldsymbol{x})$ so high/low compared to a baseline value? The baseline value is commonly chosen to be the average

model output $\mathbb{E}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{z})]$ over a distribution $\mathcal{B}$ called the background. At the heart of any contrastive question is a quantity called the Gap

$$G(h, \boldsymbol{x}) := h(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{z})]. \tag{2}$$

Therefore, asking a contrastive question amounts to measuring a Gap $G(h, \boldsymbol{x}) \neq 0$ and wondering why it is strongly positive or negative. Examples of contrastive questions include:

1. Why is individual $\boldsymbol{x}$ predicted to have a higher-than-average risk of heart disease? Here, the Gap is positive and the background $\mathcal{B}$ is the empirical distribution over the whole dataset.

2. Why is house $\boldsymbol{x}$ predicted to have a lower price than house $\boldsymbol{z}$? In that case, the Gap is negative and the background $\mathcal{B}$ is the Dirac measure $\delta_{\boldsymbol{z}}$.

Now, to *answer* a contrastive question, we need mathematical tools to probe the model and extract information from it. Examples of such techniques are local feature attributions, which are vector-valued functionals $\boldsymbol{\phi} : \mathcal{H} \times \mathcal{X} \to \mathbb{R}^d$ whose vector output represents the contribution of each feature towards the Gap

$$\sum_{i=1}^{d} \phi_i(h, \boldsymbol{x}) = G(h, \boldsymbol{x}). \tag{3}$$

Since the feature attributions sum up to the Gap, a large positive attribution for feature i is interpreted as stating that the input component x_i increased the model output relative to the baseline. The amplitude of the score $|\phi_i(h, \boldsymbol{x})|$ is called the local feature importance and it is often used to rank features. That is, we interpret $|\phi_i(h, \boldsymbol{x})| < |\phi_j(h, \boldsymbol{x})|$ as stating that feature i is locally less important than feature j for explaining the Gap. Returning to the contrastive question on house prices, a negative attribution with maximal local importance may be given to the size $x_i = \texttt{small}$ of house $\boldsymbol{x}$. This would mean that the small size of the house is the main factor driving its price down relative to house $\boldsymbol{z}$. In this work, we are only going to consider local feature attributions that are linear w.r.t the model:

$$\boldsymbol{\phi}(h_1 + \alpha h_2, \boldsymbol{x}) = \boldsymbol{\phi}(h_1, \boldsymbol{x}) + \alpha\boldsymbol{\phi}(h_2, \boldsymbol{x}), \tag{4}$$

for any hypotheses $h_1, h_2 \in \mathcal{H}$, and $\alpha \in \mathbb{R}$. The principal reason for this restriction is that it will render the optimization problems described in **Section 4** tractable. We now present two linear local feature attributions methods that have previously been used to answer contrastive questions about black boxes: SHAP (Lundberg and Lee, 2017), and Expected Gradient (EG) (Erion et al., 2021).

2.2.1 Shapley Values

The Shapley values are a fundamental concept from cooperative game theory (Shapley, 1953). Letting $[d] = \{1, 2, \ldots, d\}$ be the set of all d features, and given a subset $P \subseteq [d]$ of features, we define the replace function $\boldsymbol{r}_P : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ as

$$r_P(\boldsymbol{z}, \boldsymbol{x})_i = \begin{cases} x_i & \text{if } i \in P \\ z_i & \text{otherwise.} \end{cases} \tag{5}$$

Moreover, let π be a permutation of $[d]$, $\pi(i)$ be the position of the feature i in π, and $\pi_{:i} = \{j \in [d] : \pi(j) < \pi(i)\}$. The Shapley values, as defined in the library SHAP (Lundberg and Lee, 2017), are the average marginal contributions of specifying the ith feature from the background distribution across all coalitions

$$\phi_i^{\text{SHAP}}(h, \boldsymbol{x}) := \mathop{\mathbb{E}}_{\substack{\pi \sim \Omega \\ \boldsymbol{z} \sim \mathcal{B}}} \left[h(\, \boldsymbol{r}_{\pi_{:i} \cup \{i\}}(\boldsymbol{z}, \boldsymbol{x})\,) - h(\, \boldsymbol{r}_{\pi_{:i}}(\boldsymbol{z}, \boldsymbol{x})\,) \right], \tag{6}$$

where Ω is the uniform distribution over all $d!$ permutations of the $[d]$. Because they involve an expectation over all permutations, the Shapley values scale poorly w.r.t the number of features, although a method called TreeSHAP was recently developed to reduce the complexity to polynomial assuming the model being explained is an ensemble of decision trees (Lundberg et al., 2020; Laberge and Pequignot, 2022).

2.2.2 Integrated/Expected Gradient

The Integrated/Expected Gradient (IG/EG) originates from a different background: cost-sharing in economics. It is also known as the Aumann-Shapley value and has been previously used to compute saliency maps of Convolutional Neural Networks (Sundararajan et al., 2017; Erion et al., 2021). The general definition of EG is

$$\phi_i^{\text{EG}}(h, \boldsymbol{x}) := \mathop{\mathbb{E}}_{\substack{\boldsymbol{z} \sim \mathcal{B}, \\ t \sim U(0,1)}} \left[(x_i - z_i) \frac{\partial h}{\partial x_i}\bigg|_{t\boldsymbol{x} + (1-t)\boldsymbol{z}} \right]. \tag{7}$$

The main idea of this approach is to average the gradient along linear paths between reference inputs sampled from the background and the input $\boldsymbol{x}$ of interest. When the background distribution degenerates to a single atom at input $\boldsymbol{z}$ ($\mathcal{B} = \delta_{\boldsymbol{z}}$), the Expected Gradient falls back the so-called Integrated Gradient.

2.2.3 Global Feature Attribution

As a complement to local feature attributions, global feature attributions are vector-valued functionals $\boldsymbol{\Phi} : \mathcal{H} \to \mathbb{R}_+^d$ that aim to highlight which features are globally most used by the model. Unlike local explanations, these functionals $\boldsymbol{\Phi}$ are not specific to a given input, and the values of the attributions are restricted to be positive. Hence, we will often refer to them as global feature importance. A straightforward way to extract global insight from local feature attributions is to average their absolute value across the data

$$\Phi_i^{[1]}(h) := \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}} [\, |\phi_i(h, \boldsymbol{x})| \,] \tag{8}$$

which is the by-default scheme in the Python libraries `SHAP` (Lundberg and Lee, 2017) and `InterpretML` (Nori et al., 2019). Another way to combine local attributions into global ones is to average their squared amplitude

$$\Phi_i^{[2]}(h) := \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}} [\, \phi_i(h, \boldsymbol{x})^2 \,]. \tag{9}$$

Although this functional has not yet been proposed, it is a natural measure of importance for linear models $h(\boldsymbol{x}) = \omega_0 + \sum_{i=1}^d \omega_i x_i$ whose local feature attributions (using $\mathcal{B} = \mathcal{D}$) are

$\phi_i(h, \boldsymbol{x}) = \omega_i(x_i - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}[z_i])$ (Lundberg and Lee, 2017). In that case, the global importance $\Phi_i^{[2]} = \omega_i^2 \mathbb{V}_{\boldsymbol{z}\sim\mathcal{D}}(z_i)$ correspond to the standardized coefficients. We will also see in **Section 5** that $\boldsymbol{\Phi}^{[2]}$ presents computational advantages over $\boldsymbol{\Phi}^{[1]}$ in the case of Additive Regression.

This work focuses on SHAP and EG local feature attributions and their global counterpart $\boldsymbol{\Phi}^{[1]}$ or $\boldsymbol{\Phi}^{[2]}$. Still, but there exist many more post-hoc methods for local/global feature attributions. For instance, LIME (Ribeiro et al., 2016) computes local feature attributions by training a linear model to mimic the behavior of h around $\boldsymbol{x}$. This local explainer was not used because it does not respect Equation 3. Moreover, Permutation Feature Importance (Breiman, 2001a) and SAGE (Covert et al., 2020) extract global feature importance by perturbing a feature and reporting the impact on model performance. Studying these two global importance techniques is part of our future work.

2.3 Under-Specification and Rashomon Set

The Rashomon Effect (Breiman, 2001b), also known as model under-specification (D'Amour et al., 2020) or model multiplicity (Marx et al., 2020) refers to the observation that there often exists a large diversity of models that fit the data well. This is especially true when one is employing a hypothesis space with a large capacity. Formally, model under-specification can be characterized via the Rashomon Set (Fisher et al., 2019)

Definition 1 (Rashomon Set) *Given a hypothesis space $\mathcal{H}$, a loss function ℓ, a data set S, and a tolerance threshold $\epsilon > 0$, the Rashomon set is defined as*

$$\mathcal{R}(\mathcal{H}, \epsilon) := \left\{h \in \mathcal{H} : \widehat{\mathcal{L}}_S(h) \leq \epsilon\right\}, \tag{10}$$

where we leave the dependence in S and ℓ implicit from the context.

Although Rashomon Sets have an appealing and simple interpretation, their computation is intractable unless $|\mathcal{H}|$ is small or unless $\mathcal{H}$ is the set of linear hypotheses fitted with squared loss. Hence, in general settings, the Rashomon Sets have to be estimated, which can be done by sampling models and keeping the ones with satisfactory performance (Dong and Rudin, 2019; Semenova et al., 2022). However, this method can be time-consuming and requires extensive memory to store thousands of models. Alternatively, by relaxing the notion of "model" to include all possible feature selections (*i.e.* $\mathcal{H} = \cup_{D\subseteq[d]}\mathcal{H}_D$ where $\mathcal{H}_D$ only relies on features in D), forward selection strategies can enumerate good models more efficiently (Kissel and Mentch, 2021).

Other approaches work implicitly with the Rashomon Set by solving optimization problems over $\mathcal{H}$ under the constraint that $\widehat{\mathcal{L}}_S(h) \leq \epsilon$. In doing so, one can explore the different characteristics of models in the Rashomon Set without ever needing to represent the set explicitly. Such optimization problems have been studied to characterize the under-specification of model predictions (Marx et al., 2020; Coker et al., 2021; Hsu and Calmon, 2022), and global feature importance (Fisher et al., 2019). However, to the best of our knowledge, this is the first work that explores the range of possible local feature attributions $\boldsymbol{\phi}(h, \boldsymbol{x})$ across all models from the Rashomon Set.

2.4 Under-Specification of Feature Attributions.

In the words of Leo Breiman (2001b) "*The multiplicity problem and its effect on conclusions drawn from models needs serious attention.*" Indeed, since models are under-specified, so are their interpretations via local/global feature attributions. In practice, this translates to situations where a large set of independently trained models all yield different local explanations for the same Gap, or different rankings of global feature importance. If our goal is to not just understand one hypothesis h, but also to provide interpretations that are robust to the inherent under-specification of the ML pipeline, then contradicting explanations are problematic. Previous work tackles this uncertainty by aggregating the feature attributions of multiple independently trained models. They both consider an ensemble $E = \{h_k\}_{k=1}^M$ of M models trained independently via a stochastic learning algorithm $h_k \sim \mathcal{A}(S)$. The local feature attributions of each of these models are computed $\{\boldsymbol{\phi}(h_k, \boldsymbol{x})\}_{k=1}^M$ and aggregated. Uncertainty scores are provided in tandem with the aggregated attributions as means to convey how "confident" the local attributions are.

For instance, Shaikhina et al. (2021) aggregate local feature attributions by explaining the average model and the uncertainty scores are the variances of feature attributions among models. That is, they define the average model $h_E = \frac{1}{M}\sum_{i=1}^M h_k$ and compute its corresponding feature attributions $\boldsymbol{\phi}(h_E, \boldsymbol{x})$, which the authors show to be equivalent to averaging the local feature attributions of each individual model when attribution is a linear functional. The uncertainty score for the attribution of feature i is the variance $\frac{1}{M}\sum_{k=1}^M(\,\phi_i(h_k, \boldsymbol{x}) - \phi_i(h_E, \boldsymbol{x})\,)^2$.

In a similar effort, Schulz et al. (2021) obtain aggregated local explanations by averaging the ranks of the feature importance across models $\frac{1}{M}\sum_{k=1}^M \mathbf{r}[\,|\boldsymbol{\phi}(h_k, \boldsymbol{x})|\,]$, where $\mathbf{r} : \mathbb{R}_+^d \to [d]$ is the rank function that maps each component of a vector to its rank among the other components. For the uncertainty score for feature i they suggest using the ordinal consensus metric, which takes values between 0 and 1 and measures the consistency between the rankings. As we shall see in **Section 3**, both of these approaches share the same limitations: it is unclear what statements we can/cannot make with confidence when analyzing the resulting local feature attributions. Indeed, they both end up providing a total order of local feature importance which suggests that every feature is either more important or less important than any other feature, irrespective of the explanation uncertainty. Moreover, the uncertainty scores shown in tandem with the explanations do not easily translate to confidence scores about statements of the form "feature i is locally more important than feature j". Finally, they do not consider the whole Rashomon Set but rather employ ensembles of M independently trained models, which may underestimate the true under-specification of the ML task.

3. Motivation

We illustrate the limitations of current methods and motivate our own with a toy regression problem. We sampled 1000 4-dimensional points $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is identity, except for $\Sigma_{1,2} = \Sigma_{2,1} = 0.75$, labelled them via $y := f(\boldsymbol{x}) + \Delta$, with $f(\boldsymbol{x}) = -8\cos(x_1 - x_2)\cos(x_1 + x_2) + 1.5x_3$ (x_4 is a dummy variable) and Δ is Gaussian noise with standard deviation $\sigma = 0.1$. We then independently trained five Multi-Layered Perceptrons (MLP) with layerwidths$= 50, 20, 10$ and ReLU activations. All models ended up

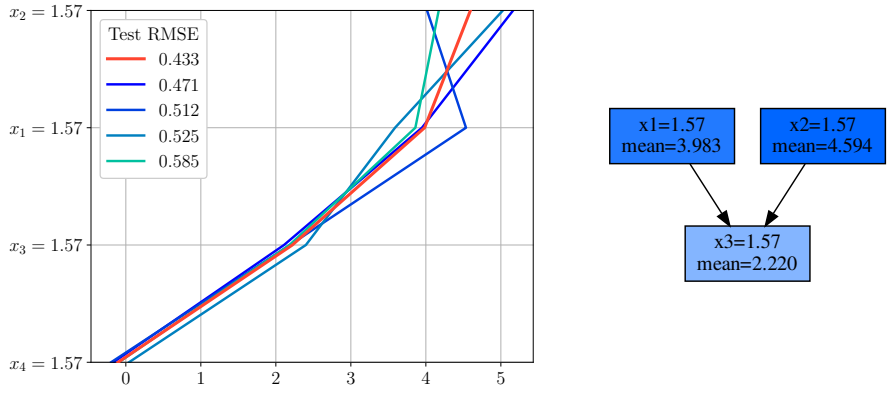


Figure 1: Left: local feature attributions for the average model h_E (orange line) and each individual model (blue lines). Right: Partial order of local feature importance. There is a directed path from feature x_i to feature x_j if **all good models** agree that feature x_i is more important than x_j.

having test set Root-Mean-Squared-Error (RMSE) between 0.47 and 0.62, while the target had a standard deviation of 4.91. After conducting paired Student-t tests between the model with RMSE 0.47 and the four others, we concluded that the one with error 0.62 was significantly worst and should be discarded. The other three models did not have a significantly worst test RMSE and so we kept them.

We analyzed the predictions of the four remaining models at the input $\boldsymbol{x} = (\frac{\pi}{2}, \frac{\pi}{2}, \frac{\pi}{2}, \frac{\pi}{2})$ which ranged from 9.05 to 10.05 (the ground truth being $f(\boldsymbol{x}) = 0.75\pi + 8 \approx 10.36$). Specifying the background distribution $\mathcal{B}$ to be the whole training set, we computed the output baselines $\mathbb{E}_{\boldsymbol{z}\sim\mathcal{B}}[\, h_k(\boldsymbol{z})\,]$ which ranged from -1.00 to -1.07 across the four models. Therefore, for all four models, the prediction Gap $G(h_k, \boldsymbol{x})$ was positive meaning that running SHAP or EG on all models would answer the same contrastive question: why is the prediction at $\boldsymbol{x}$ so much higher-than-average? To provide insight into why the Gaps at $\boldsymbol{x}$ are positive, Figure 1 (Left) presents the SHAP local feature attributions for all four models as blue lines. We see that the various MLPs lead to different interpretations. To make sense of these, we used the two state-of-the-art methods for local feature attribution aggregation.

Following Shaikhina et al. (2021), we average the predictions of our four models, leading to a single predictor h_E with a test RMSE of 0.43. The resulting SHAP feature attribution is shown as an orange line in Figure 1 (Left). The total order of local feature importance for this average model is represented in the first column of Table 1. In particular, this explanation suggests that x_2 is more important than x_1, which given our knowledge of the symmetry of the ground truth seems somewhat spurious. Indeed, since the underlying data-generating distribution, the target function f, and the point $\boldsymbol{x}$ to explain are all symmetric w.r.t x_1 and x_2, an ideal explanation would certainly not support that x_2 is more important than x_1. The uncertainty of the local feature attribution is characterized via the variance across the five models, see the second column of Table 1. We note that variance is higher for the attributions of features x_1 and x_2, suggesting that their contribution toward the output is more uncertain. Still, it is unclear what variance values are low/high enough to

Feature	Attribution h_E	Variance	Mean rank	Ordinal Consensus
$x_2 = 1.57$	4.59	0.50	2.75	0.83
$x_1 = 1.57$	3.98	0.35	2.25	0.83
$x_3 = 1.57$	2.22	0.10	1.0	1.00
$x_4 = 1.57$	-0.10	0.09	0.0	1.00

Table 1: Aggregated feature attributions and uncertainty scores following previous methods.

label attributions as trustworthy/untrustworthy. Moreover, despite their higher variance, features x_1 and x_2 are locally more important than features x_3 and x_4 for all models. Thus, the variance can lead to an overly pessimistic picture of the insights one can gather from feature attributions of multiple models

Following Schulz et al. (2021), we averaged the ranks of the SHAP local feature importance, see the third column of Table 1. This method also suggests that x_2 is locally more important than x_1, which is again spurious. Using the Ordinal Consensus as an uncertainty metric (the fourth column of Table 1) suggests that all feature importance ranks are confident. Indeed, both x_1 and x_2 have an Ordinal Consensus of 0.83 seeing that there is only a single model for which the ranks of these two features are switched. Nonetheless, looking at Figure 1 (Left), the model that contradicts all others has a test RMSE of 0.512, which is the second best of the whole ensemble. Simply put, this model offers a different but still valid perspective on the data. However, its opinion is "washed out" by the other three models in the computation of the Ordinal Consensus. Hence, we argue that the Ordinal Consensus offers a view of uncertainty that is too optimistic.

As we have just highlighted, the methods of Shaikhina et al. (2021) and Schulz et al. (2021) share the same limitations:

- It is unclear what statements one can/cannot make using these frameworks. For instance, is x_2 really more important than x_1 for explaining the gap? Both approaches return a total order of local feature importance, which suggests one statement of relative importance for every pair of features *i.e.* feature i is locally less/more important than feature j. As we have seen, the uncertainty metrics provided in tandem with the total orders (Variance or Ordinal Consensus) do not help to decide what statements on relative importance are trustworthy.

- It is unclear what is the impact of model performances on the insights provided by these two methods. For instance, the second-best model in the ensemble contradicts all others regarding the relative importance of x_1 and x_2. However, its opinions are diluted when aggregating all explanations.

In light of those takeaways, we decide to focus our method directly on statements about relative feature importance, and whether or not all good models agree on them. For instance, how can we decide if feature x_2 is locally more important than x_1? As noted earlier, one model considers, contrary to the other four, that x_1 is more important than x_2. Given that this model is as good as any other, we can simply decide to **abstain** from claiming any relation of importance between x_1 and x_2. In this case, abstention seems indeed a

cautious position given the symmetry of the ground truth. Following this logic, for every other pair of features, we check if all four models agree on their relative importance. For instance, all four models agree that x_1 is more important than x_3. We decide to record this consensus as a trustworthy statement and we represent it with an arrow from x_1 to x_3 in Figure 1 (Right). Furthermore, we observe that while all four models agree that x_1, x_2 and x_3 have a positive attribution, this is not the case for x_4 (our dummy variable). Based on this observation, we decide to keep only the variables for which all models agree on the sign and exclude x_4 from our final explanation.

All relations of importance among pairs of features for which there is consensus among the four models actually form a *partial order*, a generalization of total orderings which can be conveniently represented using a Directed Acyclic Graph called a Hasse diagram. The partial order of Figure 1 (Right) summarizes our explanation. Note that the partial order suggests that the only relative importance statements we can make are that features x_1 and x_2 are locally more important than x_3. These two statements are also supported by the total orders of Shaikhina et al. (2021) and Schulz et al. (2021), a fact that always holds as discussed in **Section 4.3**.

4. Methodology

4.1 Consensus on Statements about Feature Attributions

4.1.1 Local

Having introduced a basic motivation for considering the consensus among diverse models with good performance, we now present a formal description of the approach. First and foremost, our theory focuses on statements $s : \mathcal{H} \times \mathcal{X} \to \{0, 1\}$ about local feature attributions. Given a performance threshold $\epsilon > 0$, end-users will only be presented statements on which there is a perfect consensus for all models in the Rashomon Set

$$\forall h \in \mathcal{R}(\mathcal{H}, \epsilon) \quad s(h, \boldsymbol{x}) = 1. \tag{11}$$

We now present various statements about local feature attributions.

Definition 2 (Positive (Negative) Gap) *We say that the gap $G(h, \boldsymbol{x})$ is positive (resp. negative) according to h if $G(h, \boldsymbol{x}) > 0$ (resp. $G(h, \boldsymbol{x}) < 0$). Formally, the statements take the form $s(h, \boldsymbol{x}) = \mathbb{1}[G(h, \boldsymbol{x}) > 0]$ and $s(h, \boldsymbol{x}) = \mathbb{1}[G(h, \boldsymbol{x}) < 0]$.*

Before running SHAP or EG, it is primordial to understand the sign of the gap as it is the basis behind the contrastive question we attempt to answer. There may exist instances $\boldsymbol{x}^{(i)}$ in the data where there is no consensus on the sign of the gap. Therefore, we let

$$\mathrm{SG}(\epsilon) := \left\{ i \in [N] : \forall h_1, h_2 \in \mathcal{R}(\mathcal{H}, \epsilon) \;\; \texttt{sign}[G(h_1, \boldsymbol{x}^{(i)})] = \texttt{sign}[G(h_2, \boldsymbol{x}^{(i)})] \right\}, \tag{12}$$

be the sets of data instances on which a contrastive question makes sense. If two models disagree on the sign of the Gap, then it is useless to run SHAP or EG on them since these techniques would not end up answering the same contrastive question. If a contrastive question has been formulated without ambiguity, we can run SHAP or EG and analyze the local feature attributions.

Definition 3 (Positive (Negative) Attribution) *We say that feature i has positive (resp. negative) attribution according to h if $\phi_i(h, \boldsymbol{x}) > 0$ (resp. $\phi_i(h, \boldsymbol{x}) < 0$). More formally, the statements are $s(h, \boldsymbol{x}) = \mathbb{1}[\phi_i(h, \boldsymbol{x}) > 0]$ and $s(h, \boldsymbol{x}) = \mathbb{1}[\phi_i(h, \boldsymbol{x}) < 0]$.*

We can now define the sets

$$\mathrm{SA}(\epsilon, \boldsymbol{x}) := \left\{ i \in [d] : \forall h_1, h_2 \in \mathcal{R}(\mathcal{H}, \epsilon) \ \ \mathtt{sign}[\phi_i(h_1, \boldsymbol{x})] = \mathtt{sign}[\phi_i(h_2, \boldsymbol{x})] \right\}, \tag{13}$$

which store the features whose attribution has a consistent sign across all good models. After identifying the sign of the local feature attributions, it makes sense to order them according to their magnitude.

Definition 4 (Local Relative Importance) *We say that feature i is locally less important than j (or equivalently j is locally more important than i) according to h if $|\phi_i(h, \boldsymbol{x})| \leq |\phi_j(h, \boldsymbol{x})|$. Formally, the statements take the form $s(h, \boldsymbol{x}) := \mathbb{1}[|\phi_i(h, \boldsymbol{x})| \leq |\phi_j(h, \boldsymbol{x})|]$.*

Note that model consensus on local relative importance leads to a partial order $\preceq_{\epsilon, \boldsymbol{x}}$ on $\mathrm{SA}(\epsilon, \boldsymbol{x})$ defined by:

$$i \preceq_{\epsilon, \boldsymbol{x}} j \iff \forall\, h \in \mathcal{R}(\mathcal{H}, \epsilon) \quad |\phi_i(h, \boldsymbol{x})| \leq |\phi_j(h, \boldsymbol{x})|\,, \tag{14}$$

$\forall i, j \in \mathrm{SA}(\epsilon, \boldsymbol{x})$. By requiring a perfect consensus on the Rashomon Set, we guarantee that the order relations will be transitive. Partial orders differ from the common total orders by allowing some pairs of features to be incomparable when there exist two models with conflicting evidence on relative importance.

Recall that asserting the consensus on a statement over the Rashomon Set (*i.e.* verifying that $\forall h_1, h_2 \in \mathcal{R}(\mathcal{H}, \epsilon),\ s(h_1, \boldsymbol{x}) = s(h_2, \boldsymbol{x}) = 1$) can require checking that uncountably many hypotheses h satisfy that statement. Fortunately, for the specific statements that are of interest to us, this can be rephrased as an optimization problem.

Definition 5 (Local Feature Attribution Consensus) *Given a tolerance level $\epsilon > 0$, a Rashomon Set $\mathcal{R}(\mathcal{H}, \epsilon)$, and a local feature attribution $\boldsymbol{\phi} : \mathcal{H} \times \mathcal{X} \to \mathbb{R}^d$, consensus on statements are asserted via the following optimization problems.*

1. ***Positive (Negative) Gap :*** *There is consensus that the gap $G(h, \boldsymbol{x})$ is positive (resp. negative) if $\inf_{h \in \mathcal{R}(\mathcal{H}, \epsilon)} G(h, \boldsymbol{x}) > 0$ (resp. $\sup_{h \in \mathcal{R}(\mathcal{H}, \epsilon)} G(h, \boldsymbol{x}) < 0$).*
2. ***Positive (Negative) Attribution :*** *There is consensus that feature i has a positive (resp. negative) attribution if $\inf_{h \in \mathcal{R}(\mathcal{H}, \epsilon)} \phi_i(h, \boldsymbol{x}) > 0$ (resp. $\sup_{h \in \mathcal{R}(\mathcal{H}, \epsilon)} \phi_i(h, \boldsymbol{x}) < 0$).*
3. ***Local Relative Importance :*** *Let there be a consensus that the attribution of features i and j have signs s_i and s_j. Under this assumption, the local feature importance becomes $|\phi_i(h, \boldsymbol{x})| = s_i \phi_i(h, \boldsymbol{x})$ for any $h \in \mathcal{R}(\mathcal{H}, \epsilon)$, and similarly for feature j. Consequently, there is a consensus that i is locally less important than j if*

$$\sup_{h \in \mathcal{R}(\mathcal{H}, \epsilon)} s_i \phi_i(h, \boldsymbol{x}) - s_j \phi_j(h, \boldsymbol{x}) \leq 0.$$

These optimization problems may potentially be intractable depending on the hypothesis set $\mathcal{H}$ and loss functions ℓ. Nonetheless, we will see that they can be solved exactly and efficiently for Additive Regression, Kernel Ridge Regression, and Random Forests.

4.1.2 Global

We can also consider global model statements $s : \mathcal{H} \to \{0, 1\}$, which are no longer specific to any input $\boldsymbol{x}$, and assert a consensus over them. When interpreting models globally, there is no need to define the notions of Gap or even sign of the attribution. Indeed, since global feature importance are already positive, we only need to study statements of relative importance.

Definition 6 (Global Relative Importance) *We say that feature i is globally less important than j (or equivalently, j is globally more important than i) according to h if $\Phi_i(h) \leq \Phi_j(h)$. Formally, the statements take the form $s(h) := \mathbb{1}[\Phi_i(h) \leq \Phi_j(h)]$.*

Model consensus on global relative importance defines a partial order $\preceq_\epsilon$ on $[d]$:

$$i \preceq_\epsilon j \iff \forall\, h \in \mathcal{R}(\mathcal{H}, \epsilon) \quad \Phi_i(h) \leq \Phi_j(h). \tag{15}$$

As with local feature attributions, consensus assertion over the Rashomon Set can be rephrased as an optimization problem.

Definition 7 (Global Feature Importance Consensus) *Given a tolerance level $\epsilon > 0$, a Rashomon Set $\mathcal{R}(\mathcal{H}, \epsilon)$, and a Global Feature Importance $\boldsymbol{\Phi} : \mathcal{H} \to \mathbb{R}^d$, there is a consensus that i is globally less important than j if and only if*

$$\sup_{h \in \mathcal{R}(\mathcal{H}, \epsilon)} \Phi_i(h) - \Phi_j(h) \leq 0.$$

4.2 Recommendations for Error Tolerance

It remains to address the specification of the error tolerance ϵ. This is a critical choice because the tolerance controls the size of the Rashomon Set and therefore the number of statements on which consensus is attained. Assuming that h_S is unique, when the tolerance error is set to its minimum value we explain a single model h_S and we have total orders of local/global feature importance. As we increase ϵ, contradicting explanations will arise and the total orders will become partial orders. The number of statements present in these partial orders will diminish and eventually become null for a sufficiently high ϵ. Thus, varying the error tolerance influences *how many* statements about the empirical loss minimizer we abstain from making.

But why would we ever want to abstain from making certain statements supported by h_S? Isn't it the model that is going to be deployed anyway? The risk is that some explanations of h_S might be contradicted by another model with "slightly worst empirical loss". When this occurs, we argue that the explanations of h_S are not trustworthy and we advocate for abstention. Determining the right notion of "slightly worst empirical loss" is a difficult problem. Here we suggest two approaches 1) one based on statistical guarantees 2) a heuristic based on relative error increases.

4.2.1 Capture Bounds

Assume we can find $\epsilon_{\max}$ such that any model with a larger empirical loss can be shown to be suboptimal in terms of population loss $\mathcal{L}_{\mathcal{D}}(h)$. More precisely, with probability $1-\delta$, $\widehat{\mathcal{L}}_S(h) > \epsilon_{\max}$ implies that $\mathcal{L}_{\mathcal{D}}(h) > \mathcal{L}_{\mathcal{D}}(h^\star)$. Then it is not relevant to set $\epsilon > \epsilon_{\max}$ since the Rashomon Set would include models that are likely suboptimal. Assuming $h^\star$ is unique, this $\epsilon_{\max}$ is the smallest value that respects

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^N}[\widehat{\mathcal{L}}_S(h^\star) \leq \epsilon_{\max}] = \mathop{\mathbb{P}}_{S\sim\mathcal{D}^N}[h^\star \in \mathcal{R}(\mathcal{H}, \epsilon_{\max})] > 1-\delta. \tag{16}$$

We shall refer to such statistical guarantees as "Capture Bounds" since they guarantee that the Rashomon Set will "capture" the best-in-class model. By setting $\epsilon = \epsilon_{\max}$, with high probability, any statement on which there is a consensus on the Rashomon Set will also hold for the unknown $h^\star$. That is, we explain the best model without knowing which one it is. We now present three capture bounds

First, if $\mathcal{H}$ is finite and small (*e.g.* $|\mathcal{H}| \leq 100$), we recommend using Model Set Selection (Kissel and Mentch, 2021). We define the subset $E \subseteq \mathcal{H}$ of all models that are not significantly worse than the empirical risk minimizer h_S according to a statistical test *e.g.* paired Student-t tests with significance $1-\delta$. Setting $\epsilon_{\max} = \max\{\widehat{\mathcal{L}}_S(h)\}_{h\in E}$ guarantees that Equation 16 holds. This capture bound was previously applied to the ensemble of five MLPs from **Section 3**.

Second, if strong assumptions can be made on how the target was generated, then the following capture bound can be used.

Proposition 8 *Under the assumption that the data were generated by the optimal model $h^\star$ plus iid zero-mean Gaussian noise*

$$y = h^\star(\boldsymbol{x}) + \Delta, \quad \textit{where} \quad \Delta \sim \mathcal{N}(0, \sigma^2), \tag{17}$$

and using the squared loss $\ell(y', y) = (y'-y)^2$, we have that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^N}[\widehat{\mathcal{L}}_S(h^\star) > \epsilon_{max}] = 1 - F_{\chi^2_N}\left(\frac{N}{\sigma^2}\epsilon_{max}\right), \tag{18}$$

where $F_{\chi^2_N}$ is the CDF of a chi-2 random variable with N degrees of freedom. The proof is provided in ***Appendix A.1****.*

Solving $\delta := 1 - F_{\chi^2_N}(\frac{N}{\sigma^2}\epsilon_{\max})$ for $\epsilon_{\max}$ yields the desired tolerance. If the residuals Δ follow another law than Gaussian, one could replace the χ^2_N CDF by the CDF of the distribution of $1/N\sum_{i=1}^N(\Delta^{(i)})^2$. The assumption that the data was generated by $h^\star$ plus symmetric noise is very strong, but it is ubiquitous in Statistics and Linear Regression (See for instance (Hastie et al., 2009, Section 3.2) and (Wasserman, 2004, Section 13.5)). Therefore, we think this capture bound is *at-least* worth investigating in any regression problem.

Third, we suggest this capture bound if a good reference hypothesis can be chosen apriori *i.e.* before seeing the dataset S on which the empirical loss is computed.

Proposition 9 *Let ℓ be the 0-1 loss, $S \sim \mathcal{D}^N$ be a dataset, $h_{ref} \in \mathcal{H}$ be a reference model that is independent of S, and $h^\star$ be a best in-class hypothesis, for any $\epsilon' \in \mathbb{R}^+$, we have*

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^N}[\widehat{\mathcal{L}}_S(h^\star) \geq \epsilon' + \widehat{\mathcal{L}}_S(h_{ref})] \leq \exp\Big\{ - \frac{N\epsilon'^2}{2}\Big\}. \tag{19}$$

The proof is provided in ***Appendix A.1***

Solving $\delta := \exp\big\{ - \frac{N(\epsilon_{\max} - \widehat{\mathcal{L}}_S(h_{\text{ref}}))^2}{2}\big\}$ for $\epsilon_{\max}$ yields the error tolerance.

4.2.2 Relative Increase Heuristic

Capture bounds rely on very strong assumptions and therefore cannot be used out-of-the-box for all problems. When they are inapplicable, we recommend the heuristic

$$\epsilon = (1 + \epsilon_{\text{rel}}) \times \widehat{\mathcal{L}}_S(h_S), \tag{20}$$

for a ϵ_{rel} typically fixed to 5% (Dong and Rudin, 2019; Coker et al., 2021), although smaller values could be used. Setting ϵ based on this heuristic does not provide any statistical guarantee. Consequently, any alternative model $h' \in \mathcal{R}(\mathcal{H}, \epsilon)$ that is highlighted by the practitioner should be compared to h_S using a paired Student-t test on fresh data. For example, if a model in the Rashomon Set is found to contradict h_S on a statement of interest, then one should assert that the test error of this alternative model is not significantly worse than that of the empirical loss minimizer.

4.2.3 Sensitivity Analysis

When any experimental choice is made, it is important to conduct a sensitivity analysis regarding this choice. Otherwise, the results could be disregarded for being arbitrary or manipulable. Setting the tolerance ϵ is one such critical choice and therefore we must provide evidence that our conclusions are not too sensitive to the specific value of ϵ employed. To measure such sensitivity, we propose to compute the normalized cardinality of the local partial orders

$$|\preceq_{\epsilon,\boldsymbol{x}^{(i)}}| := \left(\frac{1}{2}d(d+1)\right)^{-1} \mathbb{1}[\boldsymbol{x}^{(i)} \in \text{SG}(\epsilon)] \times |\{(j,k) \in \text{SA}(\epsilon, \boldsymbol{x}^{(i)})^2 : j \preceq_{\epsilon,\boldsymbol{x}^{(i)}} k\}|. \tag{21}$$

For any example $\boldsymbol{x}^{(i)}$, this measure goes from 0 (when the gap does not have a consistent sign) to 1 (when we have a total order among the d features). This quantity returns the ratio of statements highlighted by the local partial order to the total number of possible statements $\frac{1}{2}d(d+1) = \frac{1}{2}d(d-1)\,$(local relative importance) $+\, d\,$(attribution sign). Given the cardinality measure, a sensitivity analysis on ϵ would involve asserting the stability of the histograms $\{|\preceq_{\epsilon,\boldsymbol{x}^{(i)}}|\}_{i=1}^N$ for small perturbations of ϵ.

4.3 Relation To Prior Work

Prior methods for characterizing the effect of model uncertainty on local feature attributions have mainly focused on explaining an ensemble of models $E = \{h_k\}_{k=1}^{M}$ trained with the same stochastic learning algorithm $h_k \sim \mathcal{A}(S)$ (Shaikhina et al., 2021; Schulz et al., 2021). We go a step further by studying the feature attributions of all models in the Rashomon Set. For this reason, it may not be immediately clear how our method compares to prior work. The following proposition shows that what we propose is a more conservative alternative to both existing methods.

Proposition 10 *Let $\boldsymbol{\phi}(\cdot, \boldsymbol{x})$ be a linear local feature attribution functional, and $E = \{h_k\}_{k=1}^{M}$ be an ensemble of M models from $\mathcal{H}$ trained with the same stochastic learning algorithm $h_k \sim \mathcal{A}(S)$. Said local feature attribution and ensemble will be employed in the methods of (Shaikhina et al., 2021; Schulz et al., 2021). Moreover, let $\epsilon \geq \max\{\widehat{\mathcal{L}}_S(h_k)\}_{k=1}^{M}$ be an error tolerance, and let $\preceq_{\epsilon,\boldsymbol{x}}$ be the consensus order relation on $SA(\epsilon, \boldsymbol{x})$ (cf. Equation 14). If the relation $i \preceq_{\epsilon,\boldsymbol{x}} j$ holds, we have that i is locally less important than j in the two total orders of prior work (Shaikhina et al., 2021; Schulz et al., 2021).*

This proposition is key as it implies that our framework will not provide users with statements that are not supported by existing approaches. In a way, all we do is abstain from making statements whose uncertainty is highest. We think this is an important property to have because, unlike model predictions, there are no ground truths for feature attributions. For example, a practitioner can apply multiple aggregation mechanisms to model predictions (Arithmetic Mean, Geometric Mean, Majority Vote etc.) and compare the resulting test set performances using the target y as ground truth. However, when aggregating feature attributions using different schemes, there is no metric for what feature importance ranking is the best, or closest to ground truth. This is one of the major challenges currently faced by the explainability community. Still, since our framework only highlights statements supported by existing approaches, we avoid the need for quantitative comparisons.

The following **Section 5, 6, & 7** each presents a practical application of our framework using a different hypothesis space and dataset. The code to reproduce these experiments is available online[1].

1. `https://github.com/gablabc/Partial_Order_in_Chaos`

5. Application to Additive Regression

5.1 Rashomon Set

Additive models have the form $h(\boldsymbol{x}) := \omega_0 + \sum_{j=1}^{d} h_j(x_j)$ where each function h_j only depends on the feature x_j. Since the output is the sum of d functions h_j, the attribution of each individual feature is readily available which is why these models are advertised as transparent. To fit an additive model, one must choose a class of hypotheses for each univariate function h_j. A first method is to represent each of the functions non-parametrically via a sum of univariate decision trees. This scheme is what is currently done in the `ExplainableBoostingMachine` of the `InterpretML` Python library (Nori et al., 2019) for instance. The parametric alternative is to define a basis $\{h_{jk}\}_{k=1}^{M_j}$ along each dimension j (for example using Splines) and represent the additive model using linear combinations of these basis functions (Hastie et al., 2009, Chapter 5)

$$h_{\boldsymbol{\omega}}(\boldsymbol{x}) := \omega_0 + \sum_{j=1}^{d} \underbrace{\sum_{k=1}^{M_j} \omega_{jk} h_{jk}(x_j)}_{h_j(x_j)} = \boldsymbol{\omega}^T \boldsymbol{h}(\boldsymbol{x}) \tag{22}$$

where

$$\boldsymbol{\omega} := [\omega_0, \underbrace{\omega_{11}, \omega_{12}, \ldots, \omega_{1,M_1}}_{\text{feature 1}}, \underbrace{\omega_{21}, \omega_{22}, \ldots, \omega_{2,M_2}}_{\text{feature 2}}, \ldots, \underbrace{\omega_{d1}, \omega_{d2}, \ldots, \omega_{d,M_d}}_{\text{feature } d}]^T,$$

and

$$\boldsymbol{h}(\boldsymbol{x}) := [1, \underbrace{h_{11}(\boldsymbol{x}), h_{12}(\boldsymbol{x}), \ldots, h_{1M_1}(\boldsymbol{x})}_{\text{feature 1}}, \ldots, \underbrace{h_{d1}(\boldsymbol{x}), h_{d2}(\boldsymbol{x}), \ldots, h_{dM_d}(\boldsymbol{x})}_{\text{feature d}}]^T.$$

By letting $\boldsymbol{H}$ be the $N \times (1 + \sum_{j=1}^{d} M_j)$ matrix whose ith row is $\boldsymbol{h}(\boldsymbol{x}^{(i)})^T$, the empirical loss minimizer for the squared loss takes the familiar form

$$\boldsymbol{\omega}_S = (\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{y}. \tag{23}$$

Definition 11 (Rashomon Set for Parametric Additive Regression) *Let $\mathcal{H}$ be the set of Parametric Additive Regression models (cf Equation 22), ℓ be the squared loss, S be a dataset of size N, and $\boldsymbol{\omega}_S = \operatorname{argmin}_{h \in \mathcal{H}} \widehat{\mathcal{L}}_S(h)$ be the least-square estimate. If one uses the performance threshold $\epsilon \geq \widehat{\mathcal{L}}_S(\boldsymbol{\omega}_S)$, then the Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$ consists of all parameters $\boldsymbol{\omega}$ s.t.*

$$(\boldsymbol{\omega} - \boldsymbol{\omega}_S)^T \frac{\boldsymbol{H}^T \boldsymbol{H}}{N} (\boldsymbol{\omega} - \boldsymbol{\omega}_S) \leq \epsilon - \widehat{\mathcal{L}}_S(\boldsymbol{\omega}_S). \tag{24}$$

We see that the Rashomon Set is an ellipsoid in parameter space. Moreover, if we let $\epsilon < \widehat{\mathcal{L}}_S(\boldsymbol{\omega}_S)$, then the Rashomon Set is empty.

This result is a simple generalization of the Rashomon Set of Ridge Regression derived in Semenova et al. (2022) to Parametric Additive Regression.

5.2 Asserting Model Consensus

5.2.1 Local Feature Attribution

In addition to having an analytical expression of their Rashomon Set, additive models also have a clear notion of local feature attribution. For instance, running SHAP and EG on an additive model while taking the whole dataset S as the background yields the same result

$$\begin{aligned}\phi_j^{\text{SHAP}}(h, \boldsymbol{x}) = \phi_j^{\text{EG}}(h, \boldsymbol{x}) &= h_j(x_j) - \frac{1}{N}\sum_{i=1}^{N} h_j(x_j^{(i)}) \\ &= \sum_{k=1}^{M_j} \omega_{jk}\left(h_{jk}(x_j) - \frac{1}{N}\sum_{i=1}^{N} h_{jk}(x_j^{(i)})\right) = \sum_{k=1}^{M_j} \omega_{jk}\overline{h}_{jk}(x_j) = \boldsymbol{\omega}_j^T\,\overline{\boldsymbol{h}}_j(\boldsymbol{x}),\end{aligned} \tag{25}$$

which is a linear function of the weights $\boldsymbol{\omega}$. We have seen previously in **Definition 5** that asserting the consensus on local feature attribution statements amounts to optimization problems that are linear with respect to the attributions $\boldsymbol{\phi}$. Therefore, asserting a consensus on the Rashomon Set of Parametric Additive models requires maximizing/minimizing a linear function on an ellipsoid

$$\begin{aligned}\min_{\boldsymbol{\omega}}/\max_{\boldsymbol{\omega}} \quad & \boldsymbol{a}^T\boldsymbol{\omega} \\ \text{with} \quad & (\boldsymbol{\omega} - \boldsymbol{\omega}_S)^T\boldsymbol{A}(\boldsymbol{\omega} - \boldsymbol{\omega}_S) \leq \epsilon - \widehat{\mathcal{L}}_S(\boldsymbol{\omega}_S),\end{aligned} \tag{26}$$

with $\boldsymbol{A} := \frac{\boldsymbol{H}^T\boldsymbol{H}}{N}$ and assuming $\epsilon \geq \widehat{\mathcal{L}}_S(\boldsymbol{\omega}_S)$. The value of $\boldsymbol{a}$ depends on the type of statement and the instance $\boldsymbol{x}^{(i)}$ being explained:

- **Positive (Negative) Gap**
  $\boldsymbol{a} = [0, \overline{h}_{11}(\boldsymbol{x}^{(i)}), \ldots, \overline{h}_{1M_1}(\boldsymbol{x}^{(i)}), \ldots, \overline{h}_{d1}(\boldsymbol{x}^{(i)}), \ldots, \overline{h}_{dM_d}(\boldsymbol{x}^{(i)})]$
- **Positive (Negative) Attribution of Feature j**
  $\boldsymbol{a} = [0, \ldots, \overline{h}_{j1}(\boldsymbol{x}^{(i)}), \overline{h}_{j2}(\boldsymbol{x}^{(i)}), \ldots, \overline{h}_{jM_j}(\boldsymbol{x}^{(i)}), \ldots, 0]$
- **Local Relative Importance of Features i and j**
  $\boldsymbol{a} = [0, \ldots, s_i\overline{h}_{i1}(\boldsymbol{x}^{(i)}), \ldots, s_i\overline{h}_{iM_i}(\boldsymbol{x}^{(i)}), 0, \ldots, 0, -s_j\overline{h}_{j1}(\boldsymbol{x}^{(i)}), \ldots, -s_j\overline{h}_{jM_j}(\boldsymbol{x}^{(i)}), \ldots, 0]$

These optimization problems have an analytical solution that can be computed rapidly using the Cholesky decomposition $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{C}^T$. The optimal values of Equation 26 are

$$\pm\sqrt{\epsilon - \widehat{\mathcal{L}}_S(\boldsymbol{\omega}_S)}\;\|\boldsymbol{a}'\| + \boldsymbol{a}^T\boldsymbol{\omega}_S, \tag{27}$$

where $\boldsymbol{a}' = \boldsymbol{C}^{-1}\boldsymbol{a}$ see **Appendix B.1.1** for more details. This result is a generalization of Theorem 4 from Coker et al. (2021) to Additive models and arbitrary linear functionals of the weights $\boldsymbol{a}^T\boldsymbol{\omega}$. We deduce from Equation 27 that the minimum and maximum values of any linear functional evaluated on the Rashomon Set are a deviation of $\sqrt{\epsilon - \widehat{\mathcal{L}}_S(\boldsymbol{\omega}_S)}\|\boldsymbol{a}'\|$ from $\boldsymbol{a}^T\boldsymbol{\omega}_S$ the value of the functional evaluated on the least-square. Since the deviation is an explicit function of the tolerance ϵ, a consensus on local feature attribution statements can be efficiently asserted at any tolerance level.

5.2.2 Global Feature Importance

Now investigating global feature importance, we observe that the functional $\Phi_j^{[2]}$ is a quadratic form of the weights

$$\begin{aligned}\Phi_j^{[2]}(h) &:= \frac{1}{N}\sum_{i=1}^{N}\phi_j(h,\boldsymbol{x}^{(i)})^2 = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\omega}_j^T\,\overline{\boldsymbol{h}}_j(\boldsymbol{x}^{(i)})\,\overline{\boldsymbol{h}}_j(\boldsymbol{x}^{(i)})^T\,\boldsymbol{\omega}_j \\ &= \boldsymbol{\omega}_j^T\left(\frac{1}{N}\sum_{i=1}^{N}\overline{\boldsymbol{h}}_j(\boldsymbol{x}^{(i)})\,\overline{\boldsymbol{h}}_j(\boldsymbol{x}^{(i)})^T\right)\boldsymbol{\omega}_j = \boldsymbol{\omega}_j^T\boldsymbol{B}_j\boldsymbol{\omega}_j.\end{aligned} \tag{28}$$

Therefore, asserting a consensus on global relative importance statements in the Rashomon Set of Additive Regression (solving **Definition 7**) requires optimizing a quadratic form over an ellipsoid

$$\begin{aligned}\min_{\boldsymbol{\omega}}/\max_{\boldsymbol{\omega}} \quad & \boldsymbol{\omega}_i^T\boldsymbol{B}_i\boldsymbol{\omega}_i - \boldsymbol{\omega}_j^T\boldsymbol{B}_j\boldsymbol{\omega}_j \\ \text{with} \quad & (\boldsymbol{\omega}-\boldsymbol{\omega}_S)^T\boldsymbol{A}(\boldsymbol{\omega}-\boldsymbol{\omega}_S) \leq \epsilon - \widehat{\mathcal{L}}_S(\boldsymbol{\omega}_S),\end{aligned} \tag{29}$$

which is known as the Trust-Region-Subproblem (TRS). Impressively, by Corollary 7.2.2 of (Conn et al., 2000, Section 7.2) this problem has necessary optimality conditions for the global optimum, even when the quadratic form is non-convex. We describe our TRS solver in **Appendix B.1.2**. We end by noting that Fisher et al. (2019) previously defined the Model Class Reliance as the interval $\left[\min_{h\in\mathcal{R}(\mathcal{H},\epsilon)}\Phi_i(h), \max_{h\in\mathcal{R}(\mathcal{H},\epsilon)}\Phi_i(h)\right]$ which they computed for Ridge Regression by solving a TRS. However, our framework is more general because we can also assert a consensus on relative importance relations *i.e.* all good models agree that i is globally less important than j.

5.3 House Price Prediction

The Kaggle-Houses[2] dataset consists of predicting the logarithm of the selling price of 2919 houses based on 79 numerical and categorical features. The training set S contains the first 1460 houses which are labeled, while the test set regroups the remaining 1459 houses whose selling prices are hidden by Kaggle. The only way to measure test performance is to submit predictions on the Kaggle Website.

For simplicity, we only selected numerical features and removed time-related features since we are only interested in the physical properties of the houses. Moreover, features that were perfectly collinear with others were ignored since they would render the matrix $\boldsymbol{H}^T\boldsymbol{H}$ singular. We were left with 19 numerical features which were non-redundant, although some had a very high Spearman correlation : `GarageArea/GarageCars`, `BsmtPercFin/BsmtFullBath`, and `BedroomAbvGrd/TotRmsAbvGrd`. We decided to keep correlated features to see how they impact model underspecification.

Additive Regression requires deciding which h_j to parametrize with spline bases and which to parametrize as linear functions of the input $h_j(x_j) = \omega_j x_j$. For each feature, we fitted the target with a depth-3 decision tree using only that feature as input and selected the k features with the lowest RMSE for spline parametrization. We tuned the hyperparameter

2. `https://www.kaggle.com/c/house-prices-advanced-regression-techniques`

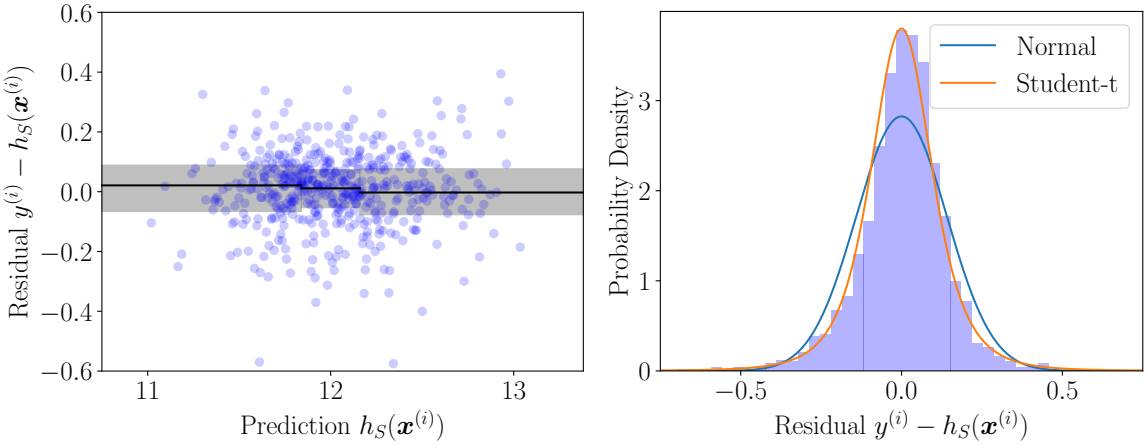


Figure 2: Residuals Analysis of h_S. (Left) Residual as a function of the prediction to assess homogeneity. The horizontal lines represent the 25$^{\text{th}}$, 50$^{\text{th}}$, and 75$^{\text{th}}$ percentiles for three different prediction bins. (Right) Histogram of the residuals and fitted densities.

k, the polynomial degree of the splines, the number of knots, and their positions via five-fold cross-validation. The resulting least-square models had a train RMSE of 0.141. As references for test performance, predicting the average training set target yields a RMSE of 0.426 on Kaggle while Gradient Boosting[3] leads to an error of 0.127. In the case of Additive Regression, we got a test error of 0.150.

To quantify the under-specification of our hypothesis class, we computed the Rashomon Set of all good models on the training set. We could not use the test set since labels are not available. To fix a reasonable value of tolerance ϵ, we investigated whether the assumptions behind the capture bound of **Proposition 8** were reasonable on this dataset. That is, could the labels have been provided by the best-in-class $h^\star$ plus iid noise Δ? We first assumed that h_S and $h^\star$ make similar enough predictions on training data to view the residuals $\{y^{(i)} - h_S(\boldsymbol{x}^{(i)})\}_{i=1}^N$ as noise samples $\{\Delta^{(i)}\}_{i=1}^N$. Figure 2 (Left) supports that the residuals are homogeneous but Figure 2 (Right) reveals they are not Gaussian and are better modeled with a Student-t. Supported by these observations, we modeled the noise Δ with a Student-t distribution fitted on the residuals. Afterward, we approximated the distribution of $\widehat{\mathcal{L}}_S(h^\star) = \frac{1}{N}\sum_{i=1}(\Delta^{(i)})^2$ with the empirical distribution resulting from sampling $\{\Delta^{(i)}\}_{i=1}^N \sim t_\nu^N$ a total of $2{\times}10^5$ times. Taking the 95$^{\text{th}}$ percentile of this empirical distribution yielded the tolerance $\epsilon_{\text{max}} = 0.1444$. Under our assumptions, by fixing $\epsilon = \epsilon_{\text{max}} = 0.1444$ we have an approximate 95% chance that the Rashomon Set will include the best-in-class model.

5.3.1 Local Feature Attribution

Local feature attributions were computed on all houses in the training set using Equation 25. To conduct a sensitivity analysis of our local explanations with respect to the choice of ϵ, we computed the partial order cardinalities (cf. Equation 21) at several tolerance values,

3. `https://www.kaggle.com/code/eesuck/xgboost-regressor`

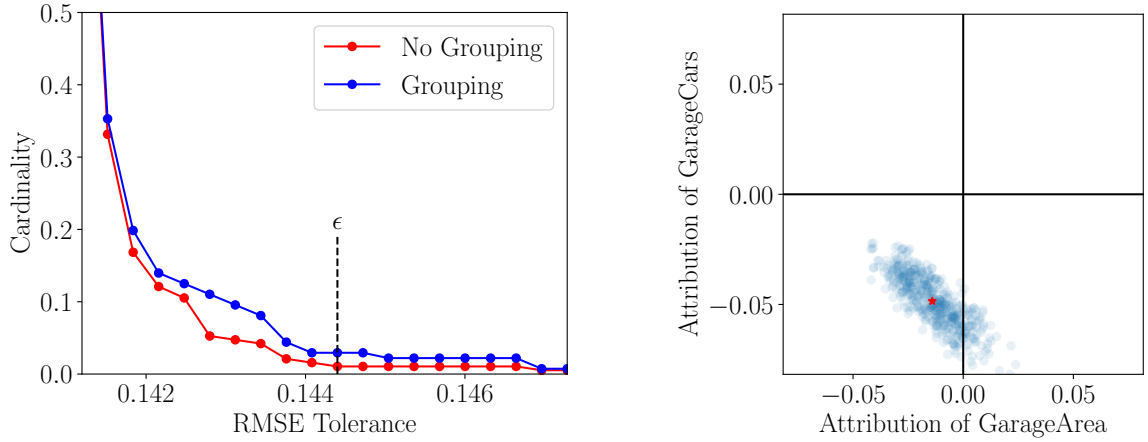


Figure 3: (Left) Sensitivity Analysis regarding the choice of ϵ. The median partial-order cardinalities are shown as a function of the tolerance on training RMSE. The two curves represent whether or not we group correlated features together. (Right) Local Feature Attributions of models sampled from the Rashomon Set boundary. We observe a trade-off between local attributions of correlated features.

see the red curve in Figure 3(Left). We observe that the cardinalities are stable with respect to small perturbations of ϵ. However, the cardinalities are rather small, which we suspect is partly due to feature correlations. To test this hypothesis, we sampled models from the Rashomon Set boundary and compared their local attributions for correlated features, see Figure 3(Right). We observe a trade-off: the more models rely on one feature, the less they rely on the other. To deal with this under-specification, we propose to *group* correlated features i and j and consider their *joint* local attribution

$$\phi_{ij}(h, \boldsymbol{x}) := \phi_i(h, \boldsymbol{x}) + \phi_j(h, \boldsymbol{x}). \tag{30}$$

instead of their separate local attribution. Therefore, we group `GarageArea/GarageCars` into `Garage`, `BsmtPercFin/BsmtFullBath` into `Bsmt` and `BedroomAbvGrd/TotRmsAbvGrd` into `AbvGrd`. Doing so, one obtains partial orders with higher cardinalities as evidenced by the red curve in Figure 3(Left), suggesting that grouping correlated features can reduce explanation under-specification. In the sequel, we will present local/global feature attributions with and without grouping.

We explained the predictions on the house with the fifth-smallest selling price: 40K USD. Said predictions ranged from 70K to 100K in the Rashomon Sets of both Scenarios and there was a consensus that the gap was negative. Figure 4 shows the local feature attribution on this instance and the partial orders that summarize all the statements good models agree on. We observe that features `OverallQual=4` (quality of materials and finish of the house from a scale of 1 to 10) and `1stFlrSF=very small` have maximal importance when explaining the drop in price relative to the mean. These statements are robust to the choice of model within the Rashomon Set. `OverallQual=4` also has maximal importance but, because it is incomparable to any other feature, we find it safer to simply ignore it. Moreover, we note that there are no garage-related and basement-related features in the Hasse diagram with-

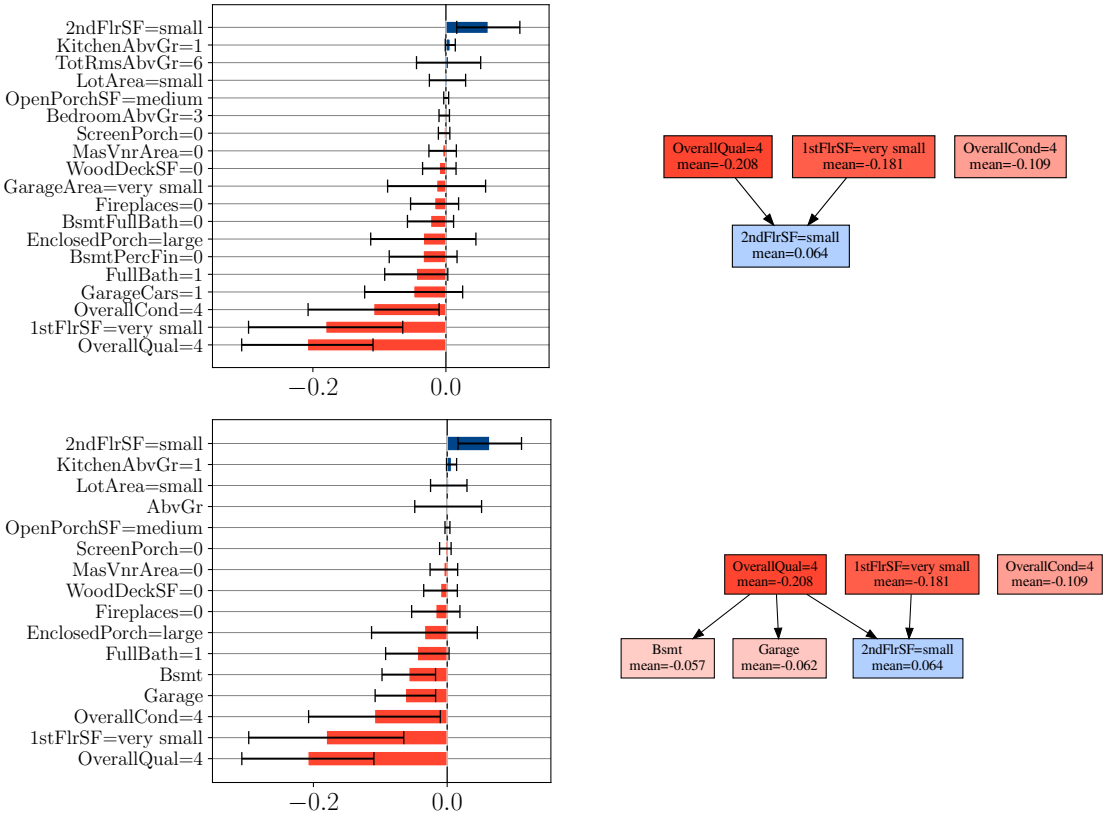


Figure 4: Local feature attributions of a house with a below-average price. (Top) Without grouping. (Bottom) With grouping.

out Grouping. As illustrated in Figure 4 (Top-Left), the attributions of highly correlated features such as `GarageArea/GarageCars` and `BsmtPercFin/BsmtFullBath` do not have a consistent sign. This is because competing models can rely on one feature or the other, which prohibits a consensus on which feature leads to a decrease in selling price. By considering the joint local attribution of correlated features, the attributions of the groups `Garage` and `Bsmt` become consistently negative, see Figure 4 (Bottom-Left). Hence, our framework allows us to get consistent model interpretations in spite of the presence of strong feature correlations.

Finally, we note that, at tolerance $\epsilon = 0.1444$, the sign of the gaps is well-defined for 68% of the houses. For about one-third of houses, it does not make sense to ask the contrastive question: *Why is this house price higher/lower than average?*. We discuss how to deal with those houses in **Appendix C**.

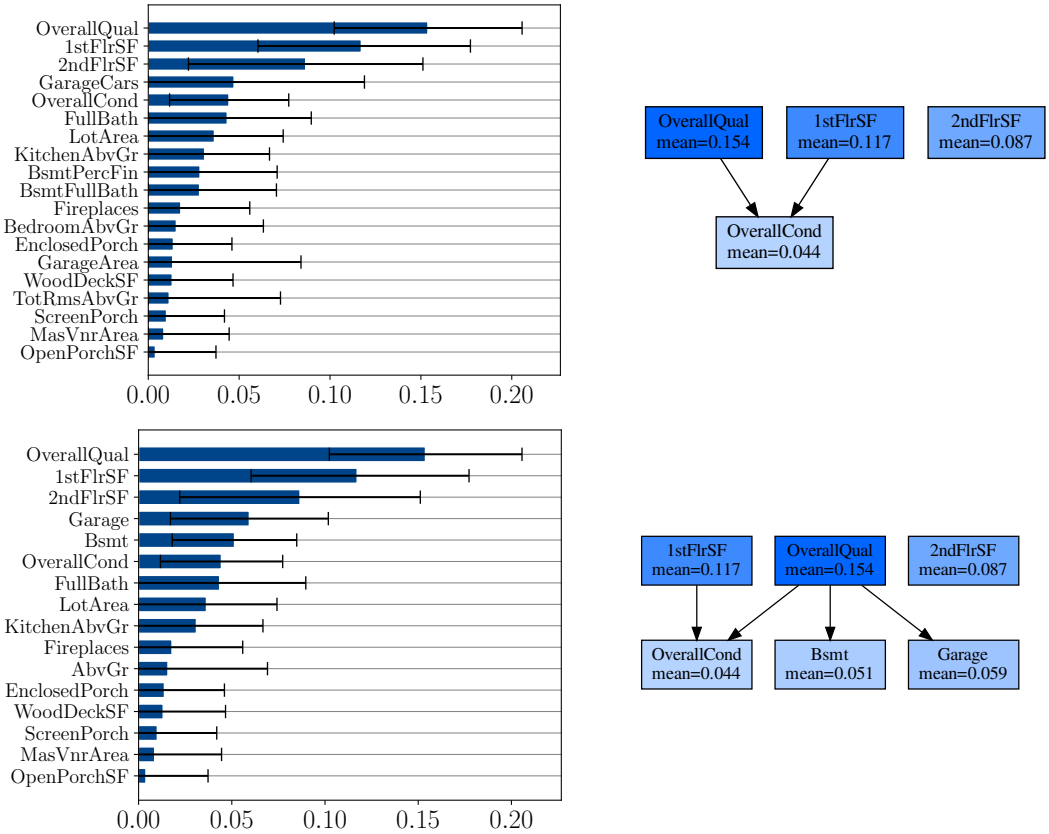


Figure 5: Global Feature Importance of the Kaggle-Houses dataset. (Top) Without grouping. (Bottom) With grouping.

5.3.2 Global Feature Importance

We end this section by presenting global feature importance in Figure 5. For simplicity, we only include in the Hasse diagrams the features whose global importance is non-null across the whole Rashomon set. Such features appear to be necessary in the sense that every model in the Rashomon Set relies on them. As seen previously, the partial order without Grouping does not contain features related to the basement and the garage. We believe that this can be again attributed to strong feature correlations. By grouping correlated features, we discover that the joint effects of `Garage` and `Bsmt` are important for all good models.

As a final observation, all models agree that `1stFlrSF` is more important than `OverallCond` despite the fact that their min-max intervals of global importance intersect. This means that looking at min-max intervals of global feature importance (*i.e.* the Model Class Reliance (Fisher et al., 2019)) does not provide the full picture of the Rashomon Set.

6. Application to Kernel Ridge Regression

6.1 Rashomon Set

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a symmetric positive definite kernel. Then such a kernel induces a functional space called a Reproducing-Kernel-Hilbert-Space (RKHS), which is actually the completion of the Pre-Hilbert space (Mohri et al., 2018)

$$\mathcal{H}_k := \left\{ h_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{R} \alpha_j k(\boldsymbol{x}, \boldsymbol{r}^{(j)}) \text{ for } R \in \mathbb{N}, \boldsymbol{\alpha} \in \mathbb{R}^R, \boldsymbol{r}^{(j)} \in \mathcal{X} \right\} \tag{31}$$

endowed with the scalar product

$$\langle\, k(\cdot, \boldsymbol{r}^{(i)})\,,\, k(\cdot, \boldsymbol{r}^{(j)})\,\rangle_{\mathcal{H}_k} := k(\boldsymbol{r}^{(i)}, \boldsymbol{r}^{(j)}), \tag{32}$$

from which the terminology "Reproducing-Kernel" arises. The space $\mathcal{H}_k$ is infinite-dimensional since it requires specifying any integer R and any R reference inputs $\boldsymbol{r}^{(j)}$. For simplicity, we shall fix the R reference inputs in advance and store them in a dictionary $D := \{\boldsymbol{r}^{(j)}\}_{j=1}^{R}$. We will then use the finite-dimensional approximation

$$\mathcal{H}_k^D := \left\{ h_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{j=1}^{R} \alpha_j k(\boldsymbol{x}, \boldsymbol{r}^{(j)}) \text{ for } \boldsymbol{\alpha} \in \mathbb{R}^R \right\} \tag{33}$$

s.t. $\mathcal{H}_k^D \subset \mathcal{H}_k$ as was done in (Fisher et al., 2019). Since these spaces are still considerably expressive, it is common to apply regularization when learning models from them. From the Rashomon perspective, this implies studying the Rashomon Set

$$\mathcal{R}(\mathcal{H}_k^D, \epsilon) := \left\{ h_{\boldsymbol{\alpha}} \in \mathcal{H}_k^D : \widehat{\mathcal{L}}_D(h_{\boldsymbol{\alpha}}) + \lambda \|h_{\boldsymbol{\alpha}}\|^2 \leq \epsilon \right\}, \tag{34}$$

where $\lambda > 0$ is a regularization hyper-parameter that is fine-tuned by cross-validation and $\|h_{\boldsymbol{\alpha}}\|^2 := \langle h_{\boldsymbol{\alpha}}, h_{\boldsymbol{\alpha}} \rangle_{\mathcal{H}_k} = \sum_{i,j=1}^{R} \alpha_i \alpha_j k(\boldsymbol{r}^{(i)}, \boldsymbol{r}^{(j)})$ is the functional norm induced by the scalar product on $\mathcal{H}_k$. We let $\boldsymbol{K} \in \mathbb{R}^{R \times R}$ be the symmetric positive semi-definite matrix of kernel evaluations on the dictionary $\boldsymbol{K}[i,j] = k(\boldsymbol{r}^{(i)}, \boldsymbol{r}^{(j)})$. The regularized least-square solution is

$$\boldsymbol{\alpha}_D = (\boldsymbol{K} + \lambda R \boldsymbol{I})^{-1} \boldsymbol{y}. \tag{35}$$

Given this notation, we can present the Rashomon Set of Kernel Ridge Regression.

Definition 12 (Rashomon Set for Kernel Ridge Regression) *Let $\mathcal{H}_k^D$ be the space induced by the kernel k and dictionary D, ℓ be the squared loss, $\lambda > 0$ be a regularization hyper-parameter, and $\boldsymbol{\alpha}_D$ be the solution of the regularized least-square. If one uses the performance threshold $\epsilon \geq \widehat{\mathcal{L}}_D(h_{\boldsymbol{\alpha}_D}) + \lambda \|h_{\boldsymbol{\alpha}_D}\|^2$, then the Rashomon set $\mathcal{R}(\mathcal{H}_k^D, \epsilon)$ consists of all models $h_{\boldsymbol{\alpha}}$ s.t.*

$$(\boldsymbol{\alpha} - \boldsymbol{\alpha}_D)^T (\boldsymbol{K}/R + \lambda \boldsymbol{I}) \boldsymbol{K} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_D) \leq \epsilon - \widehat{\mathcal{L}}_D(h_{\boldsymbol{\alpha}_D}) - \lambda \|h_{\boldsymbol{\alpha}_D}\|^2. \tag{36}$$

We see that the Rashomon Set is an ellipsoid in $\mathbb{R}^R$.

The proof is mutatis mutandis like the proof for Ridge Regression in Semenova et al. (2022) but with Kernel Ridge instead.

6.2 Asserting Model Consensus

Unlike the previous section, the model $h_{\boldsymbol{\alpha}}$ is no longer additive, and hence there is no universal way to assign a score ϕ_i to each input feature when explaining a gap in model predictions. Hence, we must rely on either SHAP or Integrated Gradient, which are two principled approaches for computing said scores. Because the exponential burden of Shapley values has not yet been solved for kernel methods, SHAP was not used and we instead employed the Integrated Gradient with a single baseline input $\boldsymbol{z}$. Henceforth, assuming the kernel is continuous and has continuous partial derivatives ($k \in \mathbb{C}^1(\mathcal{X} \times \mathcal{X})$), we compute the IG as follows.

$$\begin{aligned}\phi_i^{\text{IG}}(h_{\boldsymbol{\alpha}}, \boldsymbol{x}, \boldsymbol{z}) &:= (x_i - z_i) \int_0^1 \left.\frac{\partial h_{\boldsymbol{\alpha}}}{\partial x_i}\right|_{t\boldsymbol{x}+(1-t)\boldsymbol{z}} dt \\ &= \sum_{j=1}^{R} \alpha_j \underbrace{\left[(x_i - z_i) \int_0^1 \left.\frac{\partial k(\cdot, \boldsymbol{r}^{(j)})}{\partial x_i}\right|_{t\boldsymbol{x}+(1-t)\boldsymbol{z}} dt\right]}_{\phi_{ij}} = \sum_{j=1}^{R} \alpha_i \phi_{ij},\end{aligned} \tag{37}$$

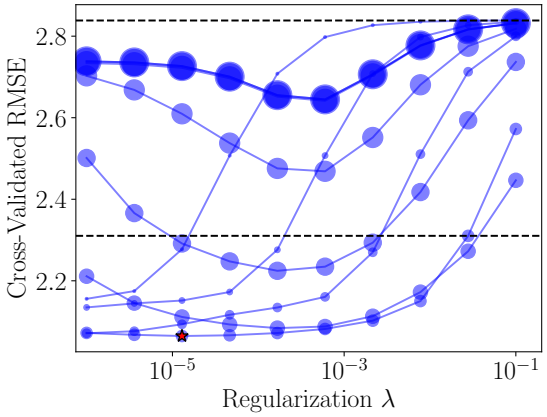
which is a linear function of the coefficients $\boldsymbol{\alpha}$. Consequently, asserting a consensus on IG feature attributions will again amount to optimizing a linear function over an ellipsoid so we can leverage results from the previous section. The only additional step required for Kernel Ridge is to pre-compute the path integrals ϕ_{ij} with common quadrature methods.

Similarly to Additive Models in **Section 5.2.2**, one can combine local feature attributions into global feature importance $\boldsymbol{\Phi}^{[2]}$ which are a quadratic form of the $\boldsymbol{\alpha}$ coefficients. Again, asserting a consensus over the Rashomon Set would require solving a TRS.

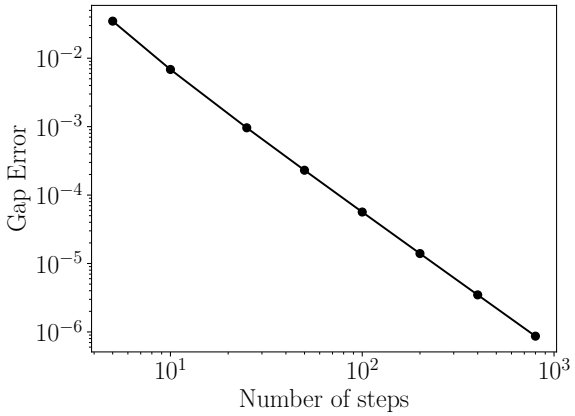
6.3 Criminal Recidivism Prediction

COMPAS is a proprietary model currently employed in the United States to predict the risk of recidivism from individuals that were recently arrested. These risks are encoded as integers going from 1 (low-risk) to 10 (high-risk). The use of this automated tool in the justice system is driven by the promise of providing objective information to judges based on empirical data, thus circumventing human biases. Still, the strong reliance of models on historical data means they can reproduce/perpetuate past injustices. To test such claims, ProPublica has collected several thousands of COMPAS scores from 2013-2014 in the Florida Broward County (Larson et al., 2016). In the resulting article, several pairs of Caucasian and African-American defendants are presented along with their COMPAS scores, the former often being lower than the latter despite the Caucasian defendant having a longer criminal history. These examples of pairs along with the subsequent analysis from the article seem to imply that the proprietary model depends on race. However, the methodology of ProPublica has been heavily criticized alongside the claim that COMPAS depends explicitly on race (Rudin et al., 2018). Hence, there may exist alternative explanations besides race for the discrepancy between scores, so it is pertinent to study the local feature attributions of the whole Rashomon Set of reasonable models when predicting COMPAS scores.

To analyze the dependencies of risk scores on the various features, we repeated the experiments of Fisher et al. (2019) where a Kernel Ridge Regression model was fitted directly on the 1-10 scores from the ProPublica dataset. The same features were employed



(a) Tuning of γ (dot sizes) and λ with Gaussian kernels. The top horizontal line shows the error of the predictor returning the mean, while the bottom line shows the error of a Random Forest with default hyperparameters.



(b) Convergence of the Gap Error w.r.t the number of steps in the quadrature. We observe a second-order convergence. Hence, augmenting the number of steps by a factor 10 reduced the error by a factor 100.

Figure 6

Input	Name	Score	Race	Age	Priors	Charge
$\boldsymbol{x}$	Robert Cannon	6	African-American	22	0	Misdemeanor
$\boldsymbol{z}$	James Rivelli	3	Caucasian	54	3	Felony

Table 2: Comparison of the COMPAS scores of two individuals.

while adding two additional ones related to juvenile misdemeanors and felonies. The dataset was split in train and test sets with ratios of 0.8 and 0.2. The training samples were used to define the dictionary of reference inputs D. We utilized the polynomial kernel $k(\boldsymbol{x}, \boldsymbol{x'}) = (\gamma\langle \boldsymbol{x}, \boldsymbol{x'}\rangle + 1)^p$ with degree $p = 3$ and the Gaussian kernel $k(\boldsymbol{x}, \boldsymbol{x'}) = \exp(-\gamma\|\boldsymbol{x} - \boldsymbol{x'}\|^2)$. The kernel scale hyper-parameter γ and the regularization factor λ were fine-tuned with 5-fold cross-validation on the training set, see the results for Gaussian Kernels in Figure 6 (a). Similar results were obtained with Polynomial Kernels. The test set RMSE of the final model was 2.11 for Gaussian Kernels and 2.12 for Polynomial Kernels. We note that the performances are worse than Fisher et al. (2019) because, unlike them, we predict the recidivism risk scores and not the risk scores for *violent* recidivism, which could be easier to predict. In this paper, we decided to study the recidivism risk scores instead since these are the ones that were actually discussed in the ProPublica article.

After fitting the models, we identified a pair of Caucasian/African-American individuals who were highlighted in the ProPublica piece and applied our explainability framework on them. More specifically, we compared Robert Cannon and James Rivelli, see Table 2. James Rivelli is a 54-year-old Caucasian man who was arrested for shoplifting. Despite having a criminal record with three priors, he was assigned a low COMPAS score. In contrast, Robert Cannon, a 22-year-old African-American charged with petit theft, was assigned a high risk

of recidivism. Letting Robert be the input of $\boldsymbol{x}$ and James be the input $\boldsymbol{z}$, we observe that the differences in scores are also present for the Kernel Ridge models: $h_{\boldsymbol{\alpha}_D}(\boldsymbol{x}) = 4.9$ and $h_{\boldsymbol{\alpha}_D}(\boldsymbol{z}) = 2.5$ for Gaussian Kernels, and $h_{\boldsymbol{\alpha}_D}(\boldsymbol{x}) = 4.9$ and $h_{\boldsymbol{\alpha}_D}(\boldsymbol{z}) = 2.4$ for Polynomial Kernels. Therefore, we have a prediction gap $G(h_{\boldsymbol{\alpha}_D}, \boldsymbol{x}) = h_{\boldsymbol{\alpha}_D}(\boldsymbol{x}) - h_{\boldsymbol{\alpha}_D}(\boldsymbol{z})$ that is positive.

Given the historical racism in the United States, it is very tempting to look at these two individuals and say that Robert Cannon is predicted to have a higher risk "because of his race". Still, there may exist a diversity of alternative explanations for this discrepancy, which we can study by exploring the Rashomon Set of our Kernel Ridge models. The Integrated Gradient was employed using Robert as the input of interest $\boldsymbol{x}$ and James as the reference input $\boldsymbol{z}$ to obtain feature attributions. Since computing the IG feature attributions requires estimating the integrals of Equation 37 with quadratures, we ended up with estimates $\widehat{\boldsymbol{\phi}}^{\text{IG}}(h_{\boldsymbol{\alpha}_D}, \boldsymbol{x})$ of the real attributions $\boldsymbol{\phi}^{\text{IG}}(h_{\boldsymbol{\alpha}_D}, \boldsymbol{x})$. We characterized the estimation error of this discretization by reporting the Gap Error

$$\left| \sum_{i=1}^{d} \widehat{\phi}_i(h_{\boldsymbol{\alpha}_D}, \boldsymbol{x}) - G(h_{\boldsymbol{\alpha}_D}, \boldsymbol{x}) \right|, \tag{38}$$

and used it as a proxy of how well $\widehat{\boldsymbol{\phi}}(h_{\boldsymbol{\alpha}_D}, \boldsymbol{x})$ approximates $\boldsymbol{\phi}(h_{\boldsymbol{\alpha}_D}, \boldsymbol{x})$. By simplicity, the Trapezoid quadrature was implemented, see Figure 6 (b) for the convergence of the Gap Error as the number of steps in the quadrature increases. We note that, as expected, the quadrature converges to the second order. For the remainder of the analysis, we have employed quadratures with 1000 steps.

Now, can we use a capture bound to set the tolerance ϵ? Unfortunately, the empirical loss $\widehat{\mathcal{L}}_D(h_{\boldsymbol{\alpha}}) + \lambda \|h_{\boldsymbol{\alpha}}\|^2$ involves regularization so we cannot guarantee that the Rashomon Set (cf. Equation 34) contains $h^\star$ unless we make a strong (unverifiable) smoothness assumption $\|h^\star\|^2 \leq B$. Without knowledge of B, we instead resort to a relative increase heuristic $\epsilon = 1.01 \times \left[\widehat{\mathcal{L}}_D(h_{\boldsymbol{\alpha}_D}) + \lambda \|h_{\boldsymbol{\alpha}_D}\|^2\right] \approx 4.23$ (an increase of $\epsilon_{\text{rel}} = 1\%$ of the minimum objective value 4.19). Unlike **Sections 5.3 & 7.3**, we did not compute a sensitivity analysis w.r.t. changes in ϵ. Rather, by setting it to a reasonably small value, we only wish to prove the existence of competing models that disagree on their explanation for the discrepancy between James and Robert.

Figure 7 presents the local feature attributions across the Rashomon Sets $\mathcal{R}(\mathcal{H}_k^D, 4.23)$ of Gaussian and Polynomial Kernels. Since the results are consistent across the two types of Kernels, we will only discuss Gaussian Kernels. Inspecting the top bar plot, we see that, according to the Integrated Gradient of the empirical loss minimizer, plotted as the blue/red bars, the features `Age=22` and `Race=Black` have positive attributions while the features `Charge=Misdemeanor` and `Prior=0` have negative attributions. This suggests that one of the possible explanations for the high risk of Robert relative to James is racial discrimination toward African-Americans. However, when we additionally consider the opinion of models with slightly worst performance on the training data, some of our previous statements on feature attribution cease to hold. Importantly, there exists a competing model $h' \in \mathcal{R}(\mathcal{H}_k^D, 4.23)$ that yield a null attribution to the feature `Race=Black`, and whose test error is not significantly worse than $h_{\boldsymbol{\alpha}_D}$ according to a paired Student-t test with $\delta = 0.05$. Therefore, there are reasonable explanations for the disparity between Robert and James, that do not rely on Robert being African-American. Even when considering the whole

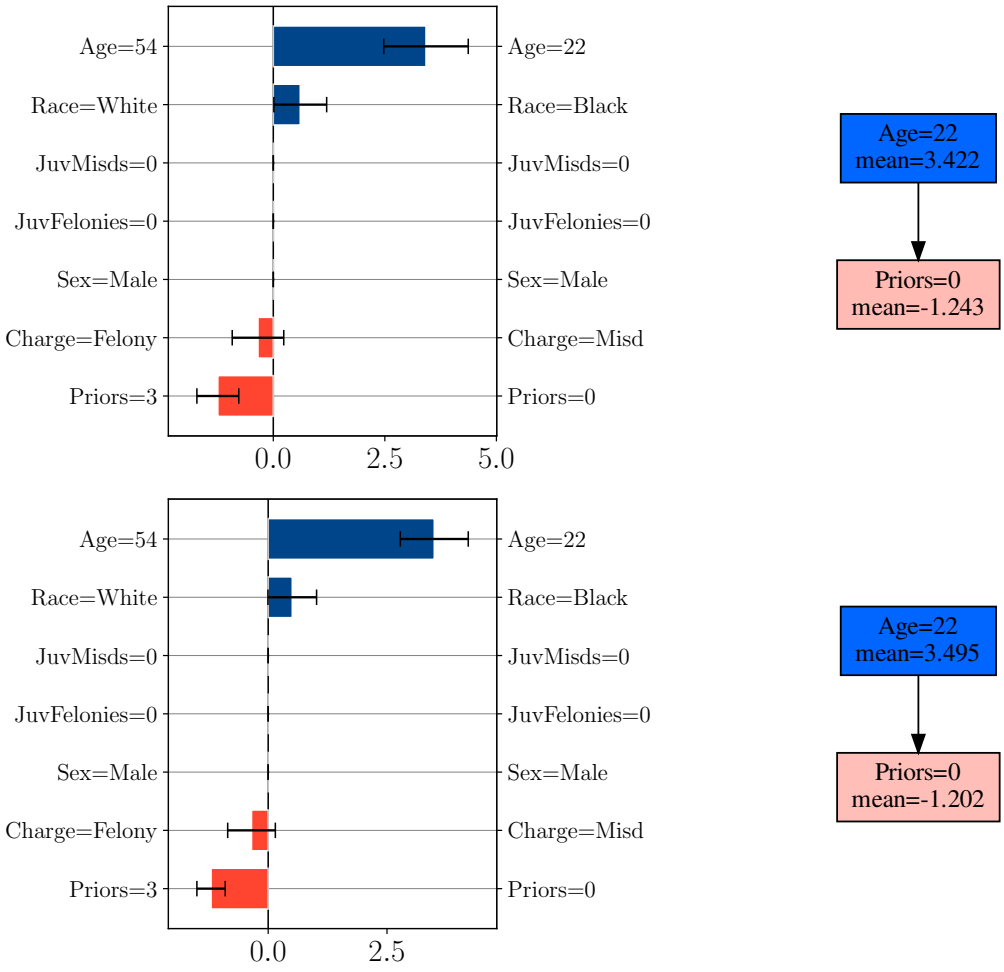


Figure 7: Local feature attribution comparing Robert Cannon to James Rivelli. (Top) Gaussian Kernels. (Bottom) Polynomial Kernels. The features on the left of the bar charts represent James while the values on the right represent Robert.

Rashomon Set, there remain statements on which models reach consensus. Notably, the attribution of the feature `Age=22` remains positive and has maximum importance.

These observations are concordant with previous work of Rudin et al. (2018) which hypothesizes that COMPAS depends strongly on age and (at most) weakly on race. Nonetheless, our local analysis on James and Robert must not be taken as absolute facts about the proprietary model COMPAS. This is because we do not have access to the model and we are surrogating it with Kernel Ridge models fitted on 7 features. The original COMPAS model, on the contrary, takes 137 different factors into consideration to produce a score (Rudin et al., 2018). Our analysis is more of a proof of concept that our explainability framework can make sense of the local feature attributions of competing models and that it can highlight the diversity of explanations for the discrepancies between two individuals.

7. Application to Random Forests

7.1 Rashomon Set

A Random Forest (RF) is an ensemble of independently trained decision trees whose predictions are averaged to yield the final predictions (Breiman, 2001a). To increase diversity, each tree is trained on a different bootstrap sample of the original dataset and each inner split is done among a random subset of features. We let s represent the seed encoding all pseudo-random processes in the training of a single tree t_s. If $\mathcal{S}$ is a distribution over all possible seeds on a computer, the theoretical definition of a RF is

$$h(\boldsymbol{x}) = \mathop{\mathbb{E}}_{s\sim\mathcal{S}}[t_s(\boldsymbol{x})]. \tag{39}$$

Given the finite representation of numbers on a computer, we can assume that the set of possible seeds is finite and of size M. Then, a reasonable choice of distribution over seeds is the uniform over M seeds *i.e.* $\mathcal{S} = U(\{1, 2 \ldots, M\})$. In practice, the expectation $\mathbb{E}_{s\sim\mathcal{S}}$ has to be approximated using Monte-Carlo sampling. Given $m < M$, we subsample m seeds uniformly at random $S \sim \mathcal{S}^m$, and return the sample average as our estimate of the RF

$$h_S(\boldsymbol{x}) = \frac{1}{m}\sum_{s\in S} t_s(\boldsymbol{x}). \tag{40}$$

By the weak law of large numbers, the estimated RF should converge to the true RF (cf. Equation 39) as m increases. Since sampling m seeds out of M with/without replacement assigns a non-zero probability to any subset of m seeds, we conceptualize the space of all **possible** RFs as the collection of all subsets of trees.

Definition 13 *Given a large set $\mathcal{T} = \{t_s\}_{s=1}^M$ of M trees trained with M seeds, the set of all possible RFs of m trees is*

$$\mathcal{H}_m := \left\{ \frac{1}{m}\sum_{t\in T} t \ : \ T \subseteq \mathcal{T} \ \text{ and } \ |T| = m \right\}, \tag{41}$$

i.e. all averages of subsets of m trees from $\mathcal{T}$. Moreover, we define $\mathcal{H}_{m:} := \cup_{k=m}^M \mathcal{H}_k$ as all RFs with least m trees. We interpret $\mathcal{H}_{1:}$ as the set of all possible RF that can ever appear in practice on a given dataset, regardless of the choice of m.

Figure 8 illustrates an example of space $\mathcal{H}_m$ which accentuates their combinatoric nature. We also note the monotonic relation $m < m' \Rightarrow \mathcal{H}_{m:} \supset \mathcal{H}_{m':}$. Since we interpret $\mathcal{H}_{1:}$ as the set of all possible RFs that can ever appear in practice on a dataset, we aim to characterize its Rashomon Set $\mathcal{R}(\mathcal{H}_{1:}, \epsilon)$. Such a Rashomon Set cannot be explicitly represented because if its exponential size ($|\mathcal{H}_{1:}| = 2^M - 1$). Still, we will see that studying the space $\mathcal{H}_{m:}$ for a carefully chosen m can help us characterize a large subset of the Rashomon Set. The reason we want to work with hypotheses $\mathcal{H}_{m:}$ is that they have a desirable property: optimizing a linear functional over them is tractable, as highlighted by the following proposition.

Proposition 14 *Let $\mathcal{T} := \{t_s\}_{s=1}^M$ be a set of M trees, $\mathcal{H}_{m:}$ be the set of all RFs with at least m trees from $\mathcal{T}$, and $\phi : \mathcal{H}_{m:} \to \mathbb{R}$ be a linear functional, then $\min_{h\in\mathcal{H}_{m:}} \phi(h)$ amounts to averaging the m smallest values of $\phi(t_s)$ for $s = 1, 2, \ldots M$.*

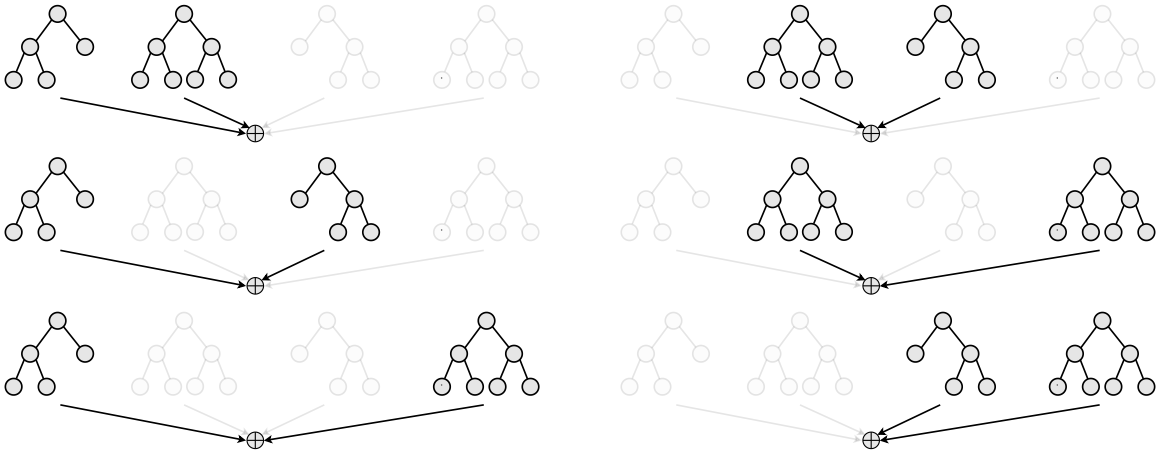


Figure 8: Example of the space $\mathcal{H}_2$ representing all the possible groupings of 2 decision trees out of $M = 4$.

The proof of this proposition is presented in **Appendix A.3**. Examples of linear functionals $\phi : \mathcal{H}_{m:} \to \mathbb{R}$ include the model prediction at fixed input $h(\boldsymbol{x})$ and the SHAP feature attribution which we can compute efficiently with TreeSHAP (Lundberg et al., 2020).

At this point, we assume that the desired tolerance on error ϵ has been fixed and so we wish to identify a value $m(\epsilon)$ that guarantees that $\mathcal{H}_{m(\epsilon):} \subseteq \mathcal{R}(\mathcal{H}_{1:}, \epsilon)$, or equivalently, that $\max_{h \in \mathcal{H}_{m(\epsilon):}} \widehat{\mathcal{L}}_S(h) \leq \epsilon$. This value $m(\epsilon)$ should be as small as possible so that the space $\mathcal{H}_{m(\epsilon):}$ is as large as possible. With this goal in mind, we restrict ourselves to losses $\ell(y', y)$ that are monotonically increasing w.r.t $|y' - y|$. This includes the 0 -1 loss and the squared loss for example. Such losses are of interest because $\max_{y' \in \mathcal{Y}'} \ell(y', y) = \max\{\ell(\min_{y' \in \mathcal{Y}'} y', y)\,,\, \ell(\max_{y' \in \mathcal{Y}'} y', y)\}$ for any set $\mathcal{Y}'$, meaning that the worst loss on a point must be attained by either of the two most extreme predictions at that point. Remembering that model predictions are linear functionals of the trees, **Proposition 14** can be used to efficiently identify the min/max predictions at any input. Therefore, it makes sense to define the upper bound

$$\begin{aligned}\max_{h \in \mathcal{H}_{m:}} \widehat{\mathcal{L}}_S(h) \leq & \frac{1}{N} \sum_{i=1}^{N} \max_{h \in \mathcal{H}_{m:}} \ell(h(\boldsymbol{x}^{(i)}), y^{(i)}), \\ = & \frac{1}{N} \sum_{i=1}^{N} \max \left\{ \ell\big(\min_{h \in \mathcal{H}_{m:}} h(\boldsymbol{x}^{(i)}), y^{(i)} \big)\,,\, \ell\big(\max_{h \in \mathcal{H}_{m:}} h(\boldsymbol{x}^{(i)}), y^{(i)} \big) \right\} := \epsilon^+(m),\end{aligned} \tag{42}$$

which can be computed efficiently at any $m \leq M$ in time $\mathcal{O}(NM \log M)$. Because of the scalability of this process w.r.t M, the total number of tree M must be reasonable, but still large enough so that $\mathcal{T} = \{t_s\}_{s=1}^{M}$ is representative of all trees that would be produced with all possible seeds on a computer. We will see in the experiments of **Section 7.3** that setting $M = 1000$ can be representative of all trees fitted on real-world data.

Now, given an absolute tolerance ϵ on the empirical loss, we search for the smallest number of trees m we can keep while ensuring that $\epsilon^+(m) \leq \epsilon$

$$m(\epsilon) := \min\{m \in [M] : \epsilon^+(m) \leq \epsilon\}. \tag{43}$$

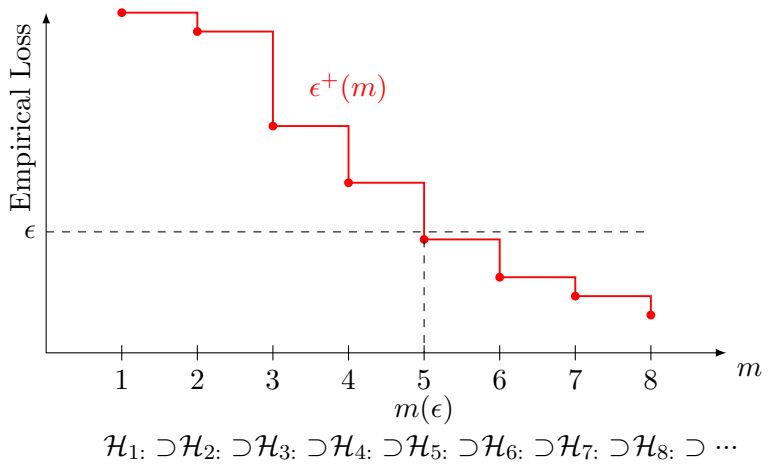


Figure 9: Choosing m based on the error tolerance ϵ.

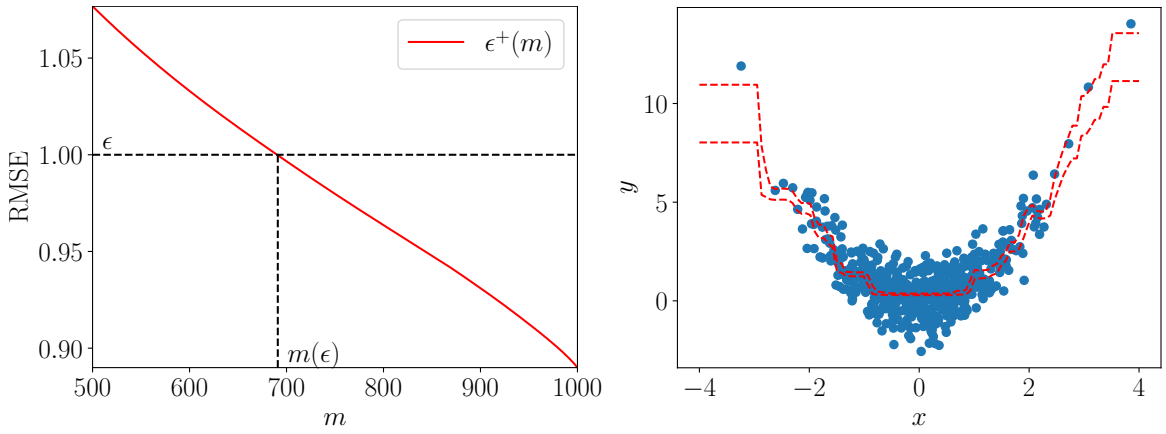
The intuition behind the computation of $m(\epsilon)$ is presented in Figure 9. Since setting $m = m(\epsilon)$ guarantees that $\max_{h \in \mathcal{H}_{m:}} \widehat{\mathcal{L}}_S(h) \leq \epsilon^+(m) \leq \epsilon$, we have $\mathcal{H}_{m(\epsilon):} \subseteq \mathcal{R}(\mathcal{H}_{1:}, \epsilon)$. Hence, we are going to employ $\mathcal{H}_{m(\epsilon):}$ as an under-estimate of the Rashomon Set over which we can efficiently optimize linear functionals such as model predictions or the SHAP local feature attributions.

We end this subsection by presenting in detail the computation of $\epsilon^+(m)$ on a toy example. We designed a regression task where the input follows a $\mathcal{N}(0, 1)$ Gaussian and the output y is a quadratic function x^2 plus some noise of amplitude 0.9. A total of $M = 1000$ different seed values were used to independently generate 1000 decision trees. Figure 10 (a) shows the upper bound $\epsilon^+(m)$ of any RF containing at least m trees. Given a threshold on the RMSE of $\epsilon = 1$, the smallest m we can safely consider is $m(\epsilon) = 691$. Hence, we suggest employing the set $\mathcal{H}_{691:}$ as a subset of $\mathcal{R}(\mathcal{H}_{1:}, 1)$. Figure 10 (b) presents the minimum and maximum predictions $\min_{h \in \mathcal{H}_{691:}} h(x)$ and $\max_{h \in \mathcal{H}_{691:}} h(x)$ at various values of x. We see that the min-max prediction intervals are wider in low-data density regions near the boundaries. This means that there is more disagreement among individual trees on these points. Such an observation makes sense because each tree is fitted on a bootstrap sample of the dataset and therefore some trees have never seen the boundary points.

7.2 Asserting Model Consensus

7.2.1 Local Feature Attribution

We discuss how to assert model consensus on local feature attributions statements at any level of tolerance ϵ. Given an error tolerance ϵ, we set m to $m(\epsilon)$, and assert the consensus on $\mathcal{H}_{m:}$ via optimization problems (cf. **Definition 5**) that we solve efficiently with **Proposition 14**. For example, to compute $\min_{h \in \mathcal{H}_{m:}} \phi_i(h, \boldsymbol{x})$, we calculate the vector of feature attributions of all trees $[\phi_i(t_1, \boldsymbol{x}), \phi_i(t_2, \boldsymbol{x}), \ldots, \phi_i(t_M, \boldsymbol{x})]^T$ with TreeSHAP, then we sort it and average its m smallest values. The overall complexity of this procedure w.r.t M is $\mathcal{O}(M \log M)$.



(a) Performance bound $\epsilon^+(m)$. Given a RMSE tolerance of $\epsilon = 1$, the smallest m we can safely consider is $m(\epsilon) = 691$.

(b) Toy regression data. The min-max predictions over the hypothesis space $\mathcal{H}_{691:}$ are shown as red lines.

Figure 10

7.2.2 Global Feature Importance

Asserting model consensus on global feature importance statements is a lot more complicated since the functionals $\mathbf{\Phi}^{[1]}, \mathbf{\Phi}^{[2]}$ are not linear w.r.t the model. Thus, we cannot leverage directly apply **Proposition 14**. We refer to **Appendix B.2** for the full details of how we deal with global feature importance. In short, we employ the functional $\mathbf{\Phi}^{[1]}$ and create an ensemble E containing

1. Approximates of $\text{argmin/max}_h \, \Phi_i^{[1]}(h)$ for $1 \leq i \leq d$.
2. Approximates of $\text{argmin/max}_h \, \Phi_i^{[1]}(h) - \Phi_j^{[1]}(h)$ for $1 \leq i < j \leq d$.

After, we assert a consensus among all models in $E \subset \mathcal{H}_{m(\epsilon):}$ leading to the partial order

$$i \, \widehat{\preceq_\epsilon} \, j \iff \forall \, h \in E \quad \Phi_i(h) \leq \Phi_j(h). \tag{44}$$

We consequently underestimate the diversity of our models, but the resulting partial order of global importance is guaranteed to be transitive.

7.3 Income Prediction

The Adult-Income dataset available on the UCI repository[4] contains the census data of 48,842 individuals collected in 1994. It consists of a binary classification task with the goal of predicting whether or not a person makes more ($y = 1$) or less ($y = 0$) than 50k USD per year based on 14 attributes. Out of all these features, we removed `fnlwgt` because we do not fully understand what it represents and `native-country` because it is a categorical feature with very high cardinality. We were finally left with five numerical features and seven one-hot-encoded categorical ones. After encoding, we were left with a data matrix of

4. https://archive.ics.uci.edu/ml/datasets/adult

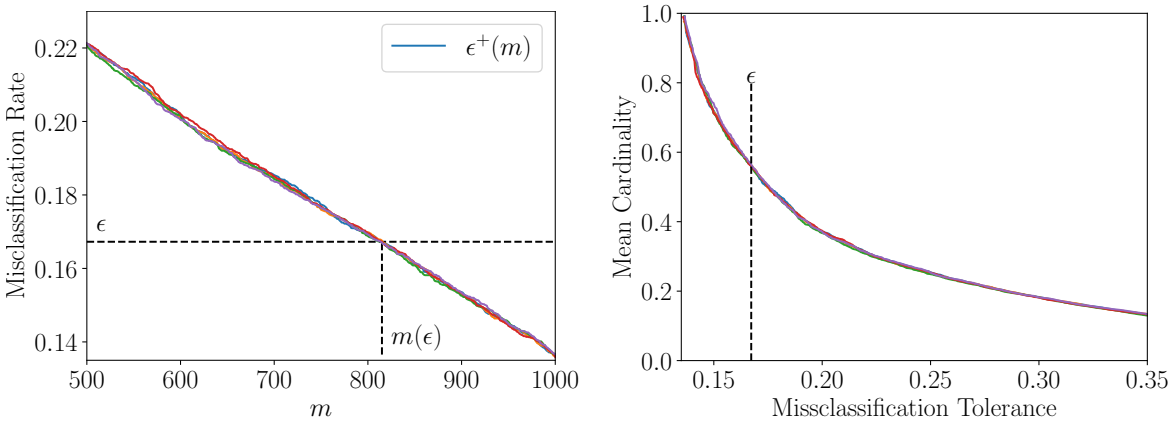


Figure 11: Estimating the Rashomon Set of RFs on Adult-Income. Each curve is associated with a different tree collection $\mathcal{T}_i$.

40 columns. The data was split into train and test sets with ratios 0.8 and 0.2 respectively. The training set was used to obtain the set $\mathcal{T}$ of M iid trees. For the model, we utilized Scikit-Learn's `RandomForestClassifiers` whose hyperparameters were tuned with a 100 steps random search and 5-fold cross-validation. Then, we trained $M = 1000$ trees in order to generate a set $\mathcal{T}$. The training was actually repeated 5 times so that we ended up with 5 distinct sets of 1000 trees $\mathcal{T}_i$ with $i = 1, 2, \ldots, 5$. We do not expect practitioners to fit several sets $\mathcal{T}_i$ when applying our methodology. This was done to verify our assumption that $\mathcal{T}$ is representative of all trees trained with bootstrapped data and random splits.

After obtaining large collections of trees, we estimated the Rashomon Set containing all RFs that perform well on the test set. The loss employed was the 0-1 loss meaning the Rashomon Set contains all models with a Misclassification Rate below some threshold ϵ. The tolerance ϵ was set via the capture bound of **Proposition 9** using $h_{\text{ref}} = 1/M \sum_{s=1}^{M} t_s$ as the reference model. This proposition is applicable since we compute the Rashomon Set on test data that is independent of the hypothesis h_{ref} which was fitted on training data. Using a confidence $\delta = 1\%$, the proposition led to an error tolerance $\epsilon = \sqrt{-2\log(1\%)/N} + \widehat{\mathcal{L}}_S(h_{\text{ref}}) \approx 3\% + \widehat{\mathcal{L}}_S(h_{\text{ref}})$. By computing the upper bound $\epsilon^+(m)$ on test samples, we set the minimum number of trees $m(\epsilon) = 815$, see Figure 11 (a). At this tolerance level, the sign of the gap is consistent for 90.8% of the individuals. Therefore, under-specification prohibits us from explaining one-tenth of the data. We refer to **Appendix C** for how we deal with those unexplainable instances.

7.3.1 Local Feature Attribution

The model outputs $h(\boldsymbol{x}) \in [0, 1]$ must be interpreted as estimates of the conditional probabilities of y given $\boldsymbol{x}$ and not as hard 0/1 predictions. Therefore, local feature attributions should sum up to a difference in conditional probabilities. We computed local feature attributions with the efficient algorithm TreeSHAP. In fact, seeing that categorical features were one-hot-encoded, which is not supported in the TreeSHAP implementation of the `SHAP`

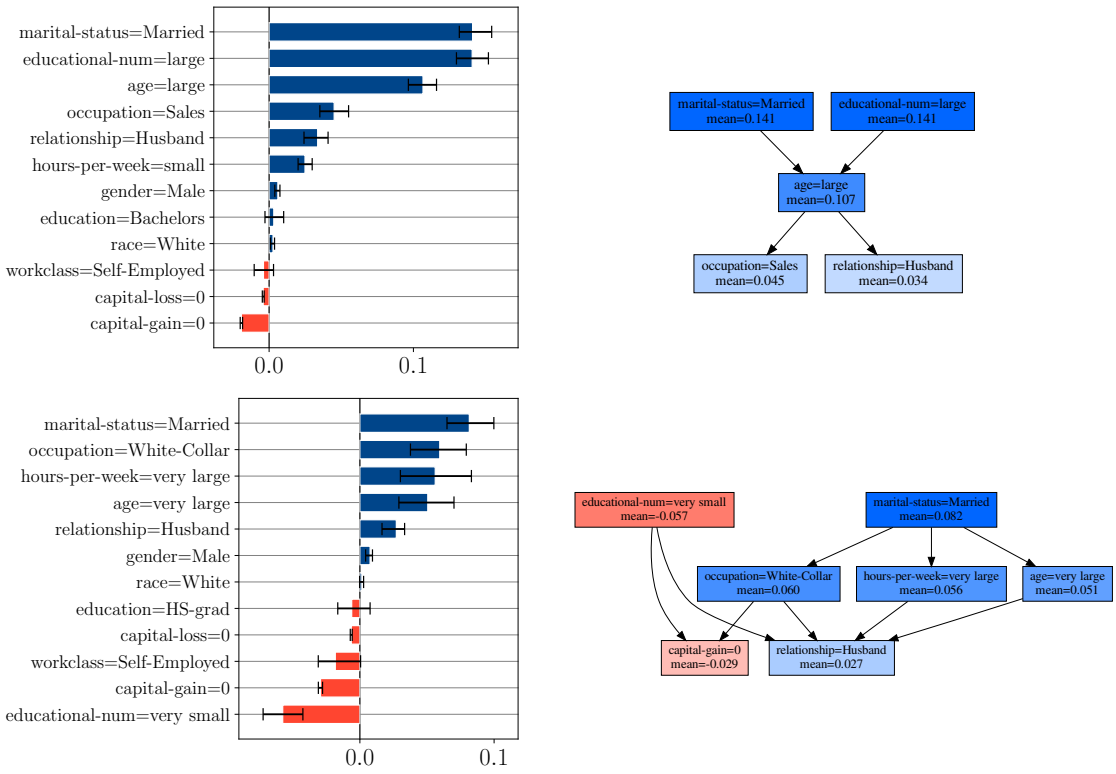


Figure 12: Local feature attributions on two individuals (Top) A person with a high prediction, (Bottom) Individual near the decision boundary. The Hasse Diagrams only show the first three ranks.

library, we used the Partition-TreeSHAP algorithm described in (Laberge and Pequignot, 2022). The feature attribution requires a background distribution $\mathcal{B}$ to serve as a reference and we used the empirical distribution of the whole training set. Still, given the considerable size of the Adult dataset, we had to subsample B instances from the training set and use them to estimate Shapley values. So, we ended up explaining the models with estimates $\widehat{\boldsymbol{\phi}}$ rather than ground-truths $\boldsymbol{\phi}$. A proxy of the error made by subsampling is the Gap Error presented in Equation 38. We found that the Gap Errors would stabilize to around 0.2% at $B = 500$ and so we employed 500 background samples. This led to a ten-minute runtime for explaining $M = 1000$ decision trees on 2000 test instances.

Figure 11 (b) presents the mean partial-order cardinality as a function of tolerance on test error. We observe that the five curves are very similar which suggests that fitting $M = 1000$ trees can be representative of all trees possibly generated for RFs. For error tolerances smaller than the ϵ employed, the mean cardinality decreases very rapidly. This means that our partial orders abstain from making many statements supported by $h_{\text{ref}} = 1/M \sum_{s=1}^{M} t_s$, but which are contradicted by other RFs with slightly worst test performance. We now discuss two instances that were explained with our framework.

The first instance is an individual who makes more than 50k per year and whose predictions range from 0.69 to 0.74 across $\mathcal{H}_{815:}$. The average prediction on the background for all

trees is 0.23 so this individual has a positive gap, which we aim to explain with TreeSHAP. Figure 12 (Top) illustrates this person's local feature attribution and the resulting partial order that encodes the statements on which there is a consensus in $\mathcal{H}_{815:}$. We observe that the features `educational-num=large` and `matiral-status=Married` have maximal positive importance for understanding why this individual has higher-than-average predictions. At the second rank is the feature `age=large`, which is also important but to a lesser extent. Looking at the bar char on the top left, we note that the feature `gender=Male` is given a small yet consistently positive attribution across all models. It appears that all RFs with at least 815 trees exhibit a small gender bias. We will come back to this in our analysis of global feature importance.

The second instance is a person who makes more than 50k and whose predictions range from 0.30 to 0.50. The prediction gap is still positive in that case but it is smaller than the previous example. Figure 12 (Bottom) shows how our framework would explain the positive gaps. We focus on the two features `capital-gain=0` and `workclass=Self-Employed` which both have a negative attribution according to the average model. Looking at the error bars on the bar chart, we observe that the model uncertainty is higher for `workclass` than with `capital-gain`. This means that there is more agreement among RFs that `capital-gain=0` reduced the model output. For `workclass=Self-Employed`, the model uncertainty is so high that the min-max interval crosses the origin, which implies the existence of RFs with satisfactory performance that yield a positive attribution to this feature. Our framework identified this ambiguity and hence removed the feature `workclass` from the partial order despite it having a negative attribution according to the average model.

7.3.2 Global Feature Importance

Figure 13 presents the global feature importance. The associated Hasse diagram is not shown because the feature ordering is a total order. Indeed, the rankings are consistent across all RFs with at least 815 trees. Interestingly, there were more disagreements when looking at local feature attributions. This highlights that combining local attributions $\boldsymbol{\phi}$ into global ones $\boldsymbol{\Phi}$ can result in information loss. Hence, it is primordial to investigate explanations under-specification both globally and locally.

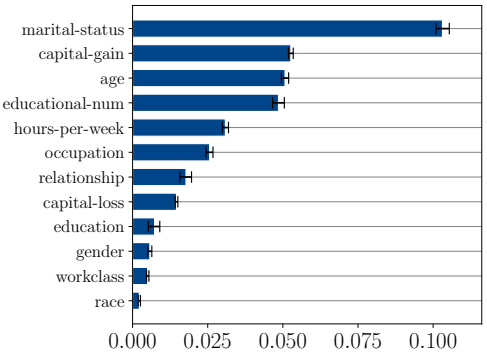


Figure 13: Global Feature Importance on Adult-Income.

Notice that all features have non-null importance across the Rashomon Set. This was not true for the hypothesis class of Additive Models, see Figure 5. We suspect that this is due to the training procedure of RFs. Indeed, when growing trees, a random subset of candidate features is chosen at each internal node. The optimal split is then chosen among these features. Hence, even if a feature is irrelevant for predicting y, there is a non-zero probability it will be used by some of the trees in the forest. This is unfortunate in the context of biases because any of our good RFs uses the `gender` features for prediction.

8. Discussion

As suggested by our experiments, model under-specification has an important impact on feature attributions on real data, and taking into account this uncertainty seems necessary to derive reliable insight from machine learning models. Our conservative approach only retains the information on features attributions on which all models agree and still succeeds in finding partial order in this chaos. This in itself is an important observation because one could have expected the partial orders to be trivial and contain no interesting structure (no arrows).

The principal limitation of our approach is that we are currently restricted to Additive Regression, Kernel Ridge Regression, and Random Forests. It is therefore primordial to extend our work to other models, especially to more Classifiers. We envision using techniques from previous work to sample Logistic Regression models and Decision Trees (Dong and Rudin, 2019; Kissel and Mentch, 2021; Semenova et al., 2022). Once an ensemble of models is available, we could apply Model Set Selection to choose ϵ (cf. **Section 4.2.1**) and assert consensus of the selected models.

Still, there may also exist hypothesis spaces whose Rashomon Set is too large to be realistically estimated, for instance, Neural Networks. Moreover, the cost of training/explaining multiple models may be too high for practitioners to see any benefit. A potential solution to derive careful conclusions from these large models would be to employ only a few models, but train them in a way that ensures they are as diverse as possible. This application of our framework is left for future work as it involves unique and novel challenges regarding the training of Neural Networks.

The main characteristic of our approach is that we require a perfect consensus among all good models. However, when employing our methodology with a finite ensemble of models, one may wonder why not also consider statements on which a majority of models agree (or at least 90% of the models agree). As a more extreme example, a practitioner may have 1000 models and 999 of these models state something while a single one states the opposite. Our approach would abstain from making any statement in that case, which may seem unnecessarily strict. An important argument for requiring a perfect consensus is that it ensures the transitivity of the order relations. This property is crucial for the interpretability of the feature orderings. We note that some prior work has produced partial orders from the consensus of at least $\alpha\%$ of the models via the transitive closure and fine-tuning of α to avoid cycles (Cheng et al., 2010). Nonetheless, in our context of local explainability, this method has two issues. First, it would require fine-tuning α for each instance $\boldsymbol{x}^{(i)}$ and therefore the interpretation of order relations would change on an instance-by-instance basis. Second, because they rely on transitive closure, the resulting Hasse diagrams could be misinterpreted seeing as the existence of a directed path between two features would not imply a consensus among at least $\alpha\%$ of the models that one feature is more important. Our diagrams, on the other hand, remain simple to interpret: for any instance $\boldsymbol{x}^{(i)}$, a directed path between two features means that all models agree on the relative importance statement and the absence of such path means that at least one model disagrees on that statement. Still, we think that imperfect consensus is a pertinent future work direction, especially for extending our framework to Bayesian methods.

On a more philosophical level, a justification for perfect consensus is that, given that the error threshold ϵ was fixed at a value that represents a satisfactory performance, any single model that disagrees with the rest is still a good model, and its mere existence is enough to put into question the claim supported by the others. Going back to the extreme scenario of 999 models disagreeing with a single one, if this solitary model had the worst performance of the whole ensemble, slightly reducing the error tolerance would remove this model from the Rashomon Set and we would reach a consensus.

Speaking of tuning the error tolerance ϵ, similar to prior work (Fisher et al., 2019; Marx et al., 2020; Hsu and Calmon, 2022), we explore a range of tolerance values and inspect the effect of under-specification on conclusions drawn from models. Nonetheless, it is not clear what is the right value for ϵ, however, we argue that this is a limitation shared by multiple studies on the Rashomon Set (D'Amour et al., 2020; Dong and Rudin, 2019; Semenova et al., 2022; Coker et al., 2021). It is well understood that the ϵ parameter should be "small enough" to represent negligible performance differences. But, there is still no agreement on what "small enough" means depending on the ML task and hypothesis space. We think the most promising directions in tackling this limitation are Proposition 7 from Fisher et al. (2019), Profile Likelihoods (Coker et al., 2021, Appendix C.1), Model Set Selection (Kissel and Mentch, 2021), and our **Propositions 8 & 9**. All these statistical guarantees suggest to define the "set of all good models" as a set that contains the best-in-class $h^\star$ with high probability. Future work should investigate these theoretical results jointly.

9. Conclusion

In this work, we propose a new approach to explanations in the context of model uncertainty. Rather than considering the mean attributions or the mean rank, we identify properties and relations of feature attributions that are consistent across a set of models with good performance. These logical statements about local/global feature attribution naturally lead to a partial order of feature importance, which we show can provide more nuanced explanations than the more common total orders based on mean attributions. As such, we believe that our work opens a new perspective on post-hoc explanations in the context of model uncertainty.

In future work, we intend to study more Classifiers (Logistic Regression, Decision Trees, Neural Networks) and other local/global post-hoc explanations (LIME, Permutation Importance, SAGE). Moreover, we shall apply our methodology to more practical settings, especially those where there are clear *actionable* features on which a human subject is able to act upon. We hope that in these scenarios, the nuance introduced by partial orders will prove most beneficial.

Acknowledgements

This work is supported by the DEEL Project CRDPJ 537462-18 funded by the National Science 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.[5]

5. `https://deel.quebec`

Appendix A. Proofs

A.1 Statistical Bounds

Proposition 15 (Proposition 8) *Under the assumption that the data was generated by the optimal model $h^\star$ plus zero-mean Gaussian noise*

$$y = h^\star(\boldsymbol{x}) + \Delta, \quad \textit{where} \quad \Delta \sim \mathcal{N}(0, \sigma^2), \tag{45}$$

and using the squared loss $\ell(y', y) = (y' - y)^2$, we have that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^N}[\widehat{\mathcal{L}}_S(h^\star) > \epsilon_{max}] = 1 - F_{\chi^2_N}\left(\frac{N}{\sigma^2}\epsilon_{max}\right), \tag{46}$$

where $F_{\chi^2_N}$ is the CDF of a chi-2 random variable with N degrees of freedom.

Proof Under the assumption that Equation 45 is valid, we have that

$$\widehat{\mathcal{L}}_S(h^\star) = \frac{1}{N}\sum_{i=1}^{N}(h^\star(\boldsymbol{x}) - y^{(i)})^2 = \frac{1}{N}\sum_{i=1}^{N}(\Delta^{(i)})^2,$$

where each $\Delta^{(i)}$ is sampled iid from a $\mathcal{N}(0, \sigma^2)$ Gaussian. Now we have

$$\begin{aligned}
\mathop{\mathbb{P}}_{S\sim\mathcal{D}^N}[\widehat{\mathcal{L}}_S(h^\star) > \epsilon_{\max}] &= \mathop{\mathbb{P}}_{\boldsymbol{\Delta}\sim\mathcal{N}(0,\sigma^2)^N}\left[\frac{1}{N}\sum_{i=1}^{N}(\Delta^{(i)})^2 > \epsilon_{\max}\right] \\
&= \mathop{\mathbb{P}}_{\boldsymbol{\Delta}\sim\mathcal{N}(0,\sigma^2)^N}\left[\sum_{i=1}^{N}\left(\frac{\Delta^{(i)}}{\sigma}\right)^2 > \frac{N}{\sigma^2}\epsilon_{\max}\right] \\
&= \mathop{\mathbb{P}}_{\boldsymbol{\Delta}\sim\mathcal{N}(0,1)^N}\left[\sum_{i=1}^{N}(\Delta^{(i)})^2 > \frac{N}{\sigma^2}\epsilon_{\max}\right] \\
&= \mathop{\mathbb{P}}_{c\sim\chi^2_N}\left[c > \frac{N}{\sigma^2}\epsilon_{\max}\right] \\
&= 1 - F_{\chi^2_N}\left(\frac{N}{\sigma^2}\epsilon_{\max}\right).
\end{aligned} \tag{47}$$

■

Proposition 16 (Proposition 9) *Let $\ell(\widehat{y}, y) = \mathbb{1}(\widehat{y} \neq y)$ be the 0-1 loss, $S \sim \mathcal{D}^N$ be a dataset, $h_{ref} \in \mathcal{H}$ be a reference model that is independent of S, and $h^\star$ be a best in-class hypothesis, for any $\epsilon' \in \mathbb{R}^+$, we have*

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^N}[\widehat{\mathcal{L}}_S(h^\star) \geq \epsilon' + \widehat{\mathcal{L}}_S(h_{ref})] \leq \exp\left\{ -\frac{N\epsilon'^2}{2} \right\}. \tag{48}$$

Proof We assume that $\widehat{\mathcal{L}}_S(h^\star) \geq \epsilon' + \widehat{\mathcal{L}}_S(h_{\text{ref}})$ and show that this implies the occurrence of an unlikely event. We first have

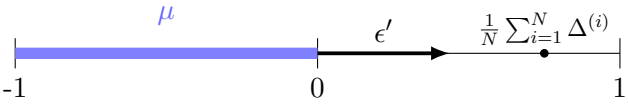
$$\widehat{\mathcal{L}}_S(h^\star) - \widehat{\mathcal{L}}_S(h_{\text{ref}}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[h^\star(\boldsymbol{x}^{(i)}) \neq y^{(i)}] - \mathbb{1}[h_{\text{ref}}(\boldsymbol{x}^{(i)}) \neq y^{(i)}] \tag{49}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \Delta^{(i)}, \tag{50}$$

where the N random variables $\Delta^{(i)} := \mathbb{1}[h^\star(\boldsymbol{x}^{(i)}) \neq y^{(i)}] - \mathbb{1}[h_{\text{ref}}(\boldsymbol{x}^{(i)}) \neq y^{(i)}]$ are iid, take values between -1 and 1, and have the expectancy

$$\mu = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^N}[\Delta^{(i)}] = \mathop{\mathbb{E}}_{(\boldsymbol{x}^{(i)},y^{(i)})\sim\mathcal{D}}[\Delta^{(i)}] = \mathcal{L}_{\mathcal{D}}(h^\star) - \mathcal{L}_{\mathcal{D}}(h_{\text{ref}}). \tag{51}$$

We accentuate that Equation 51 only holds if the reference model h_{ref} is independent on the dataset S used to assess model performance. Now by definition of $h^\star$, we have $\mu \leq 0$. However, under our assumption that $\widehat{\mathcal{L}}_S(h^\star) \geq \epsilon' + \widehat{\mathcal{L}}_S(h_{\text{ref}})$, we have that $\frac{1}{N}\sum_{i=1}^{N} \Delta^{(i)} \geq \epsilon' \geq 0$. Hence we have a bounded random variable Δ whose true mean is negative but whose empirical mean is large and positive. This event becomes highly improbable as ϵ' increases or the sample size N increases, see the following Figure.



Formally, using Hoeffding's inequality yields

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^N}[\widehat{\mathcal{L}}_S(h^\star) \geq \epsilon' + \widehat{\mathcal{L}}_S(h_{\text{ref}})] = \mathop{\mathbb{P}}_{S\sim\mathcal{D}^N}\left[\frac{1}{N}\sum_{i=1}^{N} \Delta^{(i)} \geq \epsilon'\right] \tag{52}$$

$$\leq \mathop{\mathbb{P}}_{S\sim\mathcal{D}^N}\left[\frac{1}{N}\sum_{i=1}^{N} \Delta^{(i)} - \mu \geq \epsilon'\right] \qquad \text{(Since } \mu \leq 0\text{)}$$

$$\leq \exp\left\{ -\frac{N\epsilon'^2}{2} \right\}, \qquad \text{(With Hoeffding's inequality)}$$

concluding the proof. ■

A.2 Relation to Prior Work

Proposition 17 (Proposition 10) *Let $\boldsymbol{\phi}(\cdot, \boldsymbol{x})$ be a linear feature attribution functional, and $E = \{h_k\}_{k=1}^M$ be an ensemble of M models from $\mathcal{H}$ trained with the same stochastic learning algorithm $h_k \sim \mathcal{A}(S)$. Said feature attribution and ensemble will be employed in the methods of (Shaikhina et al., 2021; Schulz et al., 2021). Moreover let $\epsilon \geq \max\{\widehat{\mathcal{L}}_S(h_k)\}_{k=1}^M$ be an error tolerance, and let $\preceq_{\epsilon,\boldsymbol{x}}$ be the consensus order relation on $SA(\epsilon, \boldsymbol{x})$ (cf. Equation 14). If the relation $i \preceq_{\epsilon,\boldsymbol{x}} j$ holds, we have that i is less important than j in the two total orders of prior work (Shaikhina et al., 2021; Schulz et al., 2021).*

Proof We first note that, since $i, j \in \mathrm{SA}(\epsilon, \boldsymbol{x})$, there is a consensus across the Rashomon Set that these features attributions have sign s_i and s_j respectively. As a reminder, this simplifies the expression of the feature importance : $\forall\, h \in \mathcal{R}(\mathcal{H}, \epsilon) \quad |\phi_i(h, \boldsymbol{x})| = s_i \phi_i(h, \boldsymbol{x})$. Additionally, our assumption that $\epsilon \geq \max\{\widehat{\mathcal{L}}_S(h_k)\}_{k=1}^M$, guarantees that $E \subseteq \mathcal{R}(\mathcal{H}, \epsilon)$. We now prove that the order relation $i \preceq_{\epsilon,\boldsymbol{x}} j$ is present in the two rankings from the literature.

Shaikhina et al. (2021) compute the average model $h_E = \frac{1}{M}\sum_{k=1}^M h_k$ and rank features according to their importance for this model $|\boldsymbol{\phi}(h_E, \boldsymbol{x})|$. For any $i, j \in \mathrm{SA}(\epsilon, \boldsymbol{x})$, we deduce

$$
\begin{aligned}
i \preceq_{\epsilon,\boldsymbol{x}} j &\Rightarrow \forall\, h \in \mathcal{R}(\mathcal{H}, \epsilon) \quad |\phi_i(h, \boldsymbol{x})| \leq |\phi_j(h, \boldsymbol{x})| \\
&\Rightarrow \forall\, h \in \mathcal{R}(\mathcal{H}, \epsilon) \quad s_i \phi_i(h, \boldsymbol{x}) \leq s_j \phi_j(h, \boldsymbol{x}) \\
&\Rightarrow \forall\, h \in E \qquad\quad s_i \phi_i(h, \boldsymbol{x}) \leq s_j \phi_j(h, \boldsymbol{x}) \\
&\Rightarrow \frac{1}{M}\sum_{k=1}^M s_i \phi_i(h_k, \boldsymbol{x}) \leq \frac{1}{M}\sum_{k=1}^M s_j \phi_j(h_k, \boldsymbol{x}) \\
&\Rightarrow s_i \phi_i(h_E, \boldsymbol{x}) \leq s_j \phi_j(h_E, \boldsymbol{x}) && \text{(By Linearity of } \boldsymbol{\phi}) \\
&\Rightarrow |\phi_i(h_E, \boldsymbol{x})| \leq |\phi_j(h_E, \boldsymbol{x})|, && \text{(By Linearity of } \boldsymbol{\phi},\ s_i = \mathtt{sign}[\,\phi_i(h_E, \boldsymbol{x})\,])
\end{aligned}
$$

thus proving that the order relation is also present when explaining the average model.

Schulz et al. (2021) order features using the mean rank $\frac{1}{M}\sum_{k=1}^M \mathbf{r}[\,|\boldsymbol{\phi}(h_k, \boldsymbol{x})|\,]$, where $\mathbf{r} : \mathbb{R}_+^d \to [d]$ is the rank function. By the definition, for any model h, we have $|\phi_i(h, \boldsymbol{x})| \leq |\phi_j(h, \boldsymbol{x})| \iff r_i[\,|\boldsymbol{\phi}(h, \boldsymbol{x})|\,] \leq r_j[\,|\boldsymbol{\phi}(h, \boldsymbol{x})|\,]$. Therefore,

$$
\begin{aligned}
i \preceq_{\epsilon,\boldsymbol{x}} j &\Rightarrow \forall\, h \in \mathcal{R}(\mathcal{H}, \epsilon) \quad |\phi_i(h, \boldsymbol{x})| \leq |\phi_j(h, \boldsymbol{x})| \\
&\Rightarrow \forall\, h \in E \qquad\quad |\phi_i(h, \boldsymbol{x})| \leq |\phi_j(h, \boldsymbol{x})| \\
&\Rightarrow \forall\, h \in E \qquad r_i[\,|\boldsymbol{\phi}(h, \boldsymbol{x})|\,] \leq r_j[\,|\boldsymbol{\phi}(h, \boldsymbol{x})|\,] \\
&\Rightarrow \frac{1}{M}\sum_{k=1}^M r_i[\,|\boldsymbol{\phi}(h_k, \boldsymbol{x})|\,] \leq \frac{1}{M}\sum_{k=1}^M r_j[\,|\boldsymbol{\phi}(h_k, \boldsymbol{x})|\,],
\end{aligned}
$$

which implies that the order relation is also supported by the mean ranks. ■

A.3 Random Forests

Proposition 18 (Proposition 14) *Let $\mathcal{T} := \{t_s\}_{s=1}^M$ be a set of M trees, $\mathcal{H}_{m:}$ be the set of all subsets of at least m trees from $\mathcal{T}$, and $\phi : \mathcal{H}_{m:} \to \mathbb{R}$ be a linear functional, then $\min_{h\in\mathcal{H}_{m:}} \phi(h)$ amounts to averaging the m smallest values of $\phi(t_s)$ for $s = 1, 2, \dots M$.*

Proof We can compute the linear functional on every tree $\{\phi(t_s)\}_{s=1}^M$ and store the indices of the m smallest ones in a set C_m s.t. $|C_m| = m$ and

$$s \in C_m \text{ and } s' \notin C_m \Rightarrow \phi(t_s) \leq \phi(t_{s'}). \tag{53}$$

Now, to prove to proposition, we must show that $\phi(\frac{1}{m}\sum_{s\in C_m} t_s) \leq \phi(h)\ \forall h \in \mathcal{H}_{m:}$. Since $\min_{h\in\mathcal{H}_{m:}} \phi(h) = \min_{k=m,\dots,M} \min_{h\in\mathcal{H}_k} \phi(h)$, the proof can be done in two parts: first for a fixed k we prove that $\phi(\frac{1}{k}\sum_{s\in C_k} t_s) \leq \phi(h)\ \forall h \in \mathcal{H}_k$ and secondly prove that $\operatorname{argmin}_{k=m,\dots,M} \phi(\frac{1}{k}\sum_{s\in C_k} t_s) = m$.

Part 1 By linearity $\phi(\frac{1}{k}\sum_{s\in C_k} t_r) = \frac{1}{k}\sum_{s\in C_k} \phi(t_r)$. Also, remember that any model $h \in \mathcal{H}_k$ is associated to a subset C'_k of k seeds *i.e.* $h = \frac{1}{k}\sum_{s\in C'_k} t_r$. Importantly, since C_k and C'_k have the same size, the two sets $C_k \backslash C'_k$ and $C'_k \backslash C_k$ have a one-to-one correspondence. We get

$$\begin{aligned}
\frac{1}{k}\sum_{s\in C_k} \phi(t_s) &= \frac{1}{k}\Bigg(\sum_{s\in C_k\cap C'_k} \phi(t_s) + \sum_{s\in C_k\backslash C'_k} \phi(t_s)\Bigg) \\
&\leq \frac{1}{k}\Bigg(\sum_{s\in C_k\cap C'_k} \phi(t_s) + \sum_{s'\in C'_k\backslash C_k} \phi(t_{s'})\Bigg) && \text{(cf. Equation 53)} \\
&= \frac{1}{k}\sum_{s\in C'_k} \phi(t_s) = \phi\Bigg(\frac{1}{k}\sum_{s\in C'_k} t_s\Bigg) = \phi(h).
\end{aligned}$$

Part 2 We now prove that $\operatorname{argmin}_{k=m,\dots,M} \phi(\frac{1}{k}\sum_{s\in C_k} t_s) = m$. The key insight is that given $m' > m$, the set C_m contains the m smallest elements of $C_{m'}$. We get

$$\begin{aligned}
\frac{1}{m'}\sum_{s\in C_{m'}} \phi(t_s) &= \frac{1}{m'}\Bigg(\sum_{s\in C_m} \phi(t_s) + \sum_{s'\in C_{m'}\backslash C_m} \phi(t_{s'})\Bigg) \\
&\geq \frac{1}{m'}\Bigg(\sum_{s\in C_m} \phi(t_s) + \sum_{s'\in C_{m'}\backslash C_m} \Bigg[\frac{1}{m}\sum_{s\in C_m} \phi(t_s)\Bigg]\Bigg) \\
&= \frac{1}{m'}\Bigg(\sum_{s\in C_m} \phi(t_s) + \frac{m'-m}{m}\sum_{s\in C_m} \phi(t_s)\Bigg) \\
&= \frac{1}{m'}\frac{m'}{m}\sum_{s\in C_m} \phi(t_s) = \frac{1}{m}\sum_{s\in C_m} \phi(t_s),
\end{aligned}$$

which ends the proof. ■

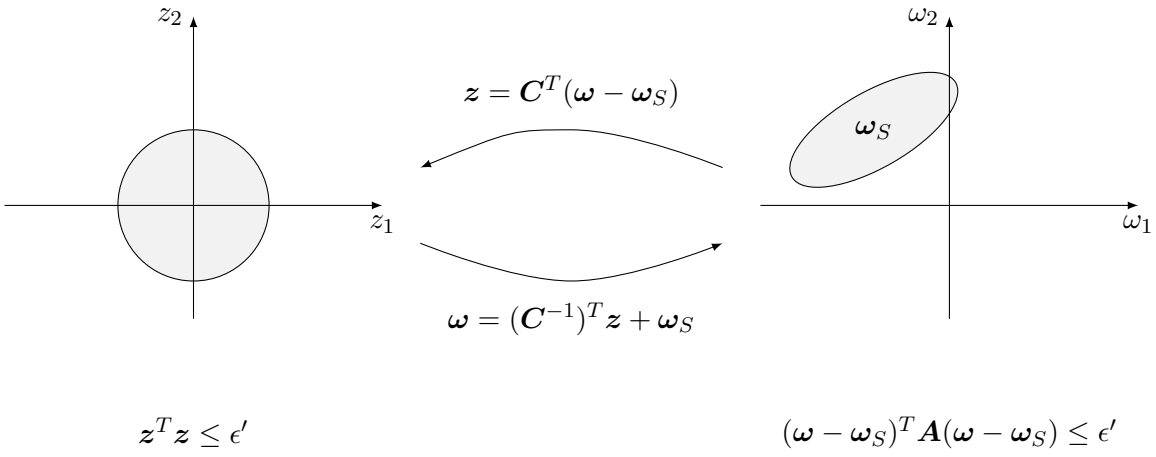


Figure 14: Mapping an ellipsoid to the unit sphere.

Appendix B. Optimization

B.1 Optimization over a Ellipsoid

B.1.1 Linear Objective

We study the optimization of a linear function over an ellipsoid

$$\begin{aligned} \max_{\boldsymbol{\omega}} \quad & \boldsymbol{a}^T\boldsymbol{\omega} \\ \text{s.t.} \quad & (\boldsymbol{\omega} - \boldsymbol{\omega}_S)^T\boldsymbol{A}(\boldsymbol{\omega} - \boldsymbol{\omega}_S) \leq \epsilon - \widehat{\mathcal{L}}_S(\boldsymbol{\omega}_S), \end{aligned} \tag{54}$$

which is necessary to compute the local feature attribution consensus on the Rashomon Set of Additive Regression and Kernel Ridge Regression. To lighten the notation, we will introduce the variable $\epsilon' := \epsilon - \widehat{\mathcal{L}}_S(\boldsymbol{\omega}_S)$. Solving Equation 54 can be done efficiently with a Cholesky decomposition of $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{C}^T$, which we know exists since $\boldsymbol{A}$ is symmetric positive definite. We also have $\boldsymbol{A}^{-1} = (\boldsymbol{C}^{-1})^T\boldsymbol{C}^{-1}$. Now, it is always possible to map an ellipsoid back to a sphere by defining a new variable

$$\boldsymbol{z} := \boldsymbol{C}^T(\boldsymbol{\omega} - \boldsymbol{\omega}_S), \tag{55}$$

see Figure 14. Applying the inverse change of variable to $\boldsymbol{\omega}$ in Equation 54, we get

$$\begin{aligned} \boldsymbol{a}^T\boldsymbol{\omega} &= \boldsymbol{a}^T\big(\,(\boldsymbol{C}^{-1})^T\boldsymbol{z} + \boldsymbol{\omega}_S\,\big) \\ &= \underbrace{\boldsymbol{a}^T(\boldsymbol{C}^{-1})^T}_{\boldsymbol{a}'^T}\boldsymbol{z} + \boldsymbol{a}^T\boldsymbol{\omega}_S, \end{aligned} \tag{56}$$

leading to the optimization problem

$$\begin{aligned} \max_{\boldsymbol{z}} \quad & \boldsymbol{a}'^T\boldsymbol{z} + \boldsymbol{a}^T\boldsymbol{\omega}_S \\ \text{s.t.} \quad & \boldsymbol{z}^T\boldsymbol{z} \leq \epsilon. \end{aligned} \tag{57}$$

Importantly, the optimization problems of Equations 54 and 57 both reach the same optimal values. Since the objective $\boldsymbol{a}'^T\boldsymbol{z}$ is a scalar product, it reaches its maximum objective value $\sqrt{\epsilon'}\|\boldsymbol{a}'\|$ when the vector $\boldsymbol{z}$ points in the same direction as $\boldsymbol{a}'$. The minimum and maximum values of the objective are therefore $\pm\sqrt{\epsilon - \widehat{\mathcal{L}}_S(\boldsymbol{\omega}_S)}\|\boldsymbol{a}'\| + \boldsymbol{a}^T\boldsymbol{\omega}_S$.

B.1.2 Quadratic Objective

We now investigate the optimization of a quadratic form over an ellipsoid

$$\begin{aligned} \min_{\boldsymbol{\omega}} \quad & \boldsymbol{\omega}_i^T \boldsymbol{B}_i \boldsymbol{\omega}_i - \boldsymbol{\omega}_j^T \boldsymbol{B}_j \boldsymbol{\omega}_j \\ \text{s.t.} \quad & (\boldsymbol{\omega} - \boldsymbol{\omega}_S)^T \boldsymbol{A} (\boldsymbol{\omega} - \boldsymbol{\omega}_S) \leq \epsilon'. \end{aligned} \tag{58}$$

Letting $\boldsymbol{\omega}_{ij} \in \mathbb{R}^{M_i+M_j}$ be the concatenation of $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_j$, and relabelling the least-square $\widehat{\boldsymbol{\omega}} := \boldsymbol{\omega}_S$, we express the optimization problem as

$$\begin{aligned} \min_{\boldsymbol{\omega}_{ij}} \quad & \boldsymbol{\omega}_{ij}^T \boldsymbol{B}_{ij} \boldsymbol{\omega}_{ij} \\ \text{s.t.} \quad & (\boldsymbol{\omega}_{ij} - \widehat{\boldsymbol{\omega}}_{ij})^T \boldsymbol{A}_{ij} (\boldsymbol{\omega}_{ij} - \widehat{\boldsymbol{\omega}}_{ij}) \leq \epsilon', \end{aligned} \tag{59}$$

where $\boldsymbol{B}_{ij}$ is a block-diagonal matrix containing $\boldsymbol{B}_i$ and $-\boldsymbol{B}_j$, and $\boldsymbol{A}_{ij}$ is the Schur complement of $\boldsymbol{A}$. The Schur complement is computed because we must project the Rashomon Set (which is an ellipsoid in $\mathbb{R}^{1+\sum_j M_j}$) onto the subspace $\mathbb{R}^{M_i+M_j}$ in which $\boldsymbol{\omega}_{ij}$ resides. Importantly, the projection of an ellipsoid on a subspace is still an ellipsoid whose covariance matrix is the Schur complement. Taking the Cholesky decomposition $\boldsymbol{A}_{ij} = \boldsymbol{C}\boldsymbol{C}^T$ and using the change of variable in Equation 55, we get

$$\boldsymbol{\omega}_{ij}^T \boldsymbol{B}_{ij} \boldsymbol{\omega}_{ij} = (\boldsymbol{z}_{ij} - \widehat{\boldsymbol{z}}_{ij})^T \boldsymbol{B}'_{ij} (\boldsymbol{z}_{ij} - \widehat{\boldsymbol{z}}_{ij}), \tag{60}$$

with $\boldsymbol{B}'_{ij} = \boldsymbol{C}^{-1}\boldsymbol{B}_{ij}(\boldsymbol{C}^{-1})^T$ and $\widehat{\boldsymbol{z}}_{ij} := -\boldsymbol{C}^T\widehat{\boldsymbol{\omega}}_{ij}$. Thus, we can express the optimization in standard TRS form

$$\begin{aligned} \min_{\boldsymbol{z}_{ij}} \quad & (\boldsymbol{z}_{ij} - \widehat{\boldsymbol{z}}_{ij})^T \boldsymbol{B}'_{ij} (\boldsymbol{z}_{ij} - \widehat{\boldsymbol{z}}_{ij}) \\ \text{s.t.} \quad & \boldsymbol{z}_{ij}^T \boldsymbol{z}_{ij} \leq \epsilon' \end{aligned} \tag{61}$$

and solve the following necessary optimality conditions adapted from Corollary 7.2.2 in (Conn et al., 2000, Section 7.2).

Corollary 19 (TRS Necessary Optimality Condition) *Letting $\{\sigma_k\}_k$ be the eigenvalues of the matrix $\boldsymbol{B}'_{ij}$, any global minimizer $\boldsymbol{z}_{ij}$ of the TRS (Equation 61) must satisfy*

$$\boldsymbol{B}'_{ij}(\boldsymbol{z}_{ij} - \widehat{\boldsymbol{z}}_{ij}) = \lambda \boldsymbol{z}_{ij} \tag{62}$$

$$\lambda(\boldsymbol{z}_{ij}^T \boldsymbol{z}_{ij} - \epsilon') = 0, \tag{63}$$

*for some $\lambda \geq \max\{0\} \cup \{\text{-}\sigma_k\}_k$. If $\lambda > \max\{\text{-}\sigma_k\}_k$ then $\boldsymbol{z}_{ij}$ is the **unique** global minimizer.*

To solve these conditions, we diagonalize $\boldsymbol{B}'_{ij} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$, define $\boldsymbol{\alpha} = \boldsymbol{V}^T\boldsymbol{z}_{ij}$ and $\widehat{\boldsymbol{\alpha}} = \boldsymbol{V}^T\widehat{\boldsymbol{z}}_{ij}$. Then, assuming $\lambda > \max\{-\sigma_k\}_k$, we rewrite Equation 62 as

$$\boldsymbol{\alpha} = (\boldsymbol{D} + \lambda \boldsymbol{I})^{-1} \boldsymbol{D} \widehat{\boldsymbol{\alpha}}. \tag{64}$$

Also assuming $\lambda > 0$, Equation 63 becomes $\boldsymbol{\alpha}^T\boldsymbol{\alpha} = \epsilon'$, which combined with Equation 64 yields

$$q(\lambda) := \sum_k \frac{\sigma_k^2}{(\sigma_k + \lambda)^2}\widehat{\alpha}_k = \epsilon'. \tag{65}$$

We finally solve the non-linear Equation $q(\lambda) = \epsilon'$ for $\lambda > \max\{0\}\cup\{\text{-}\sigma_k\}_k$ with the bissection algorithm. From the resulting λ we can determine the TRS solution $\boldsymbol{z}_{ij}$.

If we do not assume $\lambda > \max\{0\}\cup\{\text{-}\sigma_k\}_k$, there are two additional cases to consider:

1. The solution is inside the ball ($\lambda = 0$).
2. The so-called "Hard Case" where $\lambda = \max\{\text{-}\sigma_k\}_k$ and $(\boldsymbol{D} + \lambda\boldsymbol{I})$ becomes singular.

For simplicity, we do not address them in this Appendix. We instead refer to (Conn et al., 2000, Section 7.3) for discussion on these technicalities.

B.2 Combinatorial Optimization and Relaxations

B.2.1 Min/Max of Global Importance

In this section we discuss the combinatorial optimization problems that occur when computing the global feature importance over the Rashomon Set of Random Forests. As a reminder, we have defined

$$\mathcal{H}_m := \left\{ \frac{1}{m}\sum_{t\in T} t \;:\; T \subseteq \mathcal{T} \;\text{ and }\; |T| = m \right\}, \tag{66}$$

as the set of RFs containing m trees. An alternative way to represent such a set is to introduce binary variables $\boldsymbol{z} \in \{0,1\}^M$ with $\sum_{s=1}^M z_s = m$ and view all RFs from $\mathcal{H}_m$ as $\frac{1}{m}\sum_{s=1}^M z_s t_s$ for some $\boldsymbol{z}$.

Now letting ϕ_j be the SHAP local feature attribution of feature j, we wish to find the minimum and maximum values of the global feature importance $\Phi_j^{[1]}(h) := \frac{1}{N}\sum_{i=1}^N |\,\phi_j(h, \boldsymbol{x}^{(i)})\,|$ across all RFs with m trees

$$\begin{aligned} \min_{\boldsymbol{z}}/\max_{\boldsymbol{z}} \quad & \frac{1}{Nm}\sum_{i=1}^N \left| \sum_{s=1}^M z_s\, \phi_j(t_s, \boldsymbol{x}^{(i)}) \right| \\ \text{s.t.} \quad & \boldsymbol{z} \in \{0,1\}^M \text{ and } \sum_{s=1}^M z_s = m. \end{aligned} \tag{67}$$

These are non-linear combinatorial problems that are extremely hard to solve. For that reason, we will provide quick approximate solutions based on a Linear relaxation of Equation 67. The first step of the relaxation is to enlarge the domain of $\boldsymbol{z}$ to allow fractional values.

$$\begin{aligned} \min_{\boldsymbol{z}}/\max_{\boldsymbol{z}} \quad & \sum_{i=1}^N \left| \sum_{s=1}^M z_s\, \phi_j(t_s, \boldsymbol{x}^{(i)}) \right| \\ \text{s.t.} \quad & \boldsymbol{z} \in [0,1]^M \text{ and } \sum_{s=1}^M z_s = m. \end{aligned} \tag{68}$$

The corresponding domain is a polytope and so it is compatible with Linear Programs. The second step of the Linear relaxation is to rephrase the absolute value function $|\cdot|$ as a Linear Program

$$
\begin{aligned}
|\alpha| = \min_{\beta} \quad & \beta \\
\text{s.t.} \quad & \alpha \leq \beta \\
& \text{-}\alpha \leq \beta
\end{aligned}
\tag{69}
$$

$$
\begin{aligned}
|\alpha| = \max_{\beta} \quad & \beta\alpha \\
\text{s.t.} \quad & \text{-}1 \leq \beta \\
& \beta \leq 1
\end{aligned}
\tag{70}
$$

After we get a solution to the relaxation of Equation 68, we project $\boldsymbol{z}$ back on $\{0,1\}^M$ using the following heuristic: if there are o components with $z = 1$, we select $M-o$ fractional values in decreasing order and set them to one. The other fractional values are set to zero. For example, if we have $M = 3$ and find a solution $\boldsymbol{z} = [0, 1, 1, 0.75, 0.25]$ to the relaxation, we would discretize the solution to get $\boldsymbol{z} = [0, 1, 1, 1, 0]$. This heuristic may be sub-optimal but our goal is to provide quick approximate solutions.

Maximize By leveraging Equation 70, we can reformulate Equation 68 as

$$
\begin{aligned}
\max_{\boldsymbol{z}} \sum_{i=1}^{N} \left| \sum_{s=1}^{M} z_s \, \phi_j(t_s, \boldsymbol{x}^{(i)}) \right| &= \max_{\boldsymbol{z}} \sum_{i=1}^{N} \max_{\beta_i \in [-1,1]} \beta_i \sum_{s=1}^{M} z_s \phi_j(t_s, \boldsymbol{x}^{(i)}) \\
&= \max_{\boldsymbol{z},\boldsymbol{\beta}} \sum_{i=1}^{N} \sum_{s=1}^{M} z_s \, \beta_i \phi_j(t_s, \boldsymbol{x}^{(i)}) \\
&= \max_{\boldsymbol{z},\boldsymbol{\beta}} \boldsymbol{\beta}^T \boldsymbol{B} \boldsymbol{z},
\end{aligned}
\tag{71}
$$

where $\boldsymbol{z}$ and $\boldsymbol{\beta}$ are each restricted to a separate polytope and $B_{is} \equiv \phi_j(t_s, \boldsymbol{x}^{(i)})$. Equation 71 is known as a Bilinear Program which is a non-convex optimization problem that can be solved to local optima via the coordinate ascent algorithm (Nahapetyan, 2009). In our setting, the output of the coordinate ascent algorithm will already respect $\boldsymbol{z} \in \{0,1\}^M$ since $\max_{\boldsymbol{z}} \boldsymbol{\beta}^T \boldsymbol{B} \boldsymbol{z}$ under the constraints on $\boldsymbol{z}$ yields $z_s = 1$ for the m smallest values of $\sum_{i=1}^{N} \beta_i \phi_j(t_s, \boldsymbol{x}^{(i)})$ and $z_s = 0$ for the others.

Minimize By leveraging Equation 69, we can reformulate Equation 68 as

$$
\begin{aligned}
\min_{\boldsymbol{z},\boldsymbol{\beta}} \quad & \sum_{i=1}^{N} \beta_i \\
\text{s.t.} \quad & \boldsymbol{z} \in [0,1]^M \text{ and } \sum_{s=1}^{M} z_s = m \\
& \sum_{s=1}^{M} z_s \, \phi_j(t_s, \boldsymbol{x}^{(i)}) \leq \beta_i \\
& \text{-} \sum_{s=1}^{M} z_s \, \phi_j(t_s, \boldsymbol{x}^{(i)}) \leq \beta_i,
\end{aligned}
\tag{72}
$$

which is a Linear Program with $N + M$ variables and $2(N + M) + 1$ constraints that we can solve efficiently if N and M are not too large. However, the solution of this LP can have fractional values so we must use the discretization heuristic to get the final solution $\boldsymbol{z} \in \{0, 1\}^M$. In our experiments on Adult-Income, about 0.25% of the non-null components of $\boldsymbol{z}$ would be fractional so we suspect our discretization heuristic provided good solutions in that setting.

Now that we have discussed approximate schemes to get the min/max global feature importance across $\mathcal{H}_m$, we are left with addressing *relative* importance relations between features.

B.2.2 Global Relative Importance

To assert a consensus on global relative importance (cf. **Definitions 6 & 7**) we must solve $\min/\max_h \Phi_j^{[1]}(h) - \Phi_k^{[1]}(h)$. However, as previously discussed, we cannot guarantee to minimize/maximize $\Phi_j^{[1]}(h)$ to optimality for Random Forests. Consequently, we cannot guarantee to solve $\min/\max_h \Phi_j^{[1]}(h) - \Phi_k^{[1]}(h)$ to optimality either. This is a critical issue because the resulting partial order may not be transitive. Our solution is to create an ensemble E containing

1. Approximates of $\operatorname{argmin}/\max_h \Phi_j^{[1]}(h)$ for $1 \leq j \leq d$.
2. Approximates of $\operatorname{argmin}/\max_h \Phi_j^{[1]}(h) - \Phi_k^{[1]}(h)$ for $1 \leq j < k \leq d$.

After, we assert a consensus among all models in $E \subset \mathcal{H}_{m(\epsilon)}$: leading to the partial order

$$j \mathrel{\widehat{\preceq_\epsilon}} k \iff \forall\, h \in E \quad \Phi_j(h) \leq \Phi_k(h). \tag{73}$$

We underestimate the diversity of our models but the resulting partial order of global importance is guaranteed to be transitive. To approximate $\operatorname{argmin}/\max_h \Phi_j^{[1]}(h) - \Phi_k^{[1]}(h)$, we propose to define the set

$$S_{jk} := \{i \in [N] : \forall h \in \mathcal{H}_m \ \texttt{sign}[\phi_j(h, \boldsymbol{x}^{(i)})] = s_{ij} \text{ and } \texttt{sign}[\phi_k(h, \boldsymbol{x}^{(i)})] = s_{ik}\} \tag{74}$$

representing all data instances whose local attributions for features j and k has a consistent sign across the $\mathcal{H}_m$. Then we solve

$$\begin{aligned}
&\operatorname*{argmin}/\max_{\boldsymbol{z}} \sum_{i \in S_{jk}} \left| \sum_{s=1}^{M} z_s\, \phi_j(t_s, \boldsymbol{x}^{(i)}) \right| - \left| \sum_{s=1}^{M} z_s\, \phi_k(t_s, \boldsymbol{x}^{(i)}) \right| \\
&= \operatorname*{argmin}/\max_{\boldsymbol{z}} \sum_{i \in S_{jk}} s_{ij} \sum_{s=1}^{M} z_s\, \phi_j(t_s, \boldsymbol{x}^{(i)}) - s_{ik} \sum_{s=1}^{M} z_s\, \phi_k(t_s, \boldsymbol{x}^{(i)}) \\
&= \operatorname*{argmin}/\max_{\boldsymbol{z}} \sum_{s=1}^{M} z_s \Big(\sum_{i \in S_{jk}} s_{ij} \phi_j(t_s, \boldsymbol{x}^{(i)}) - s_{ik}\, \phi_k(t_s, \boldsymbol{x}^{(i)}) \Big) \\
&= \operatorname*{argmin}/\max_{\boldsymbol{z}} \sum_{s=1}^{M} z_s a_{jks}
\end{aligned} \tag{75}$$

which is a linear function of $\boldsymbol{z}$ thus we can leverage **Proposition 14**.

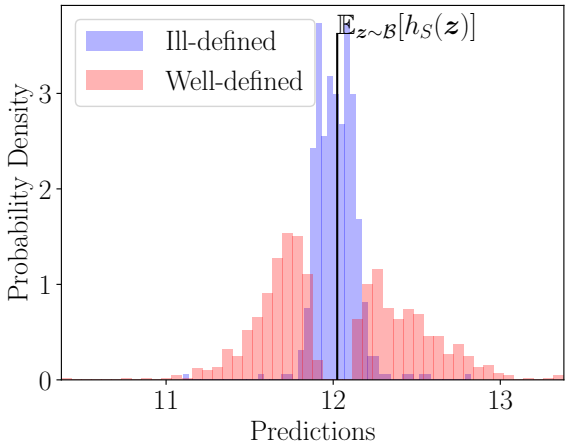


Figure 15: Distributions of predictions for houses with ill-defined and well-defined gaps across the Rashomon Set of Kaggle-Houses. The background $\mathcal{B}$ is the empirical distribution over the whole training data.

Appendix C. Ill-Defined Gaps

In this appendix, we investigate instances $\boldsymbol{x}^{(i)}$ whose gap is ill-defined given the underspecification of the ML task. That is, there exists two models $h_1, h_2 \in \mathcal{R}(\mathcal{H}, \epsilon)$ which assign gaps $G(h_1, \boldsymbol{x}^{(i)}) < 0$ and $G(h_2, \boldsymbol{x}^{(i)}) > 0$. When this occurs, it does not make sense to compute local feature attributions at $\boldsymbol{x}^{(i)}$ since the different models end up answering different contrastive questions. We now present instances with an ill-defined gap and show that redefining the background $\mathcal{B}$ can help make these points explainable.

C.1 Kaggle-Houses

As a reminder, the background $\mathcal{B}$ employed on Kaggle-Houses was the empirical distribution over the training data. Figure 15 shows the distributions of predictions for instance whose gap is well-defined or ill-defined across the Rashomon Set. We note that instances whose gap does not have a consistent sign tend to have predictions $h_S(\boldsymbol{x}^{(i)})$ near the baseline $\mathbb{E}_{\boldsymbol{z}\sim\mathcal{B}}[h_S(\boldsymbol{z})]$ so that the Gap $G(h_S, \boldsymbol{x}^{(i)})$ is very small. This could explain why models with similar performance can assign different signs to the gap. Importantly, model underspecification warns us that the contrastive question is not well-posed on these houses and it would be better to use another background $\mathcal{B}'$ when explaining them. We redefined $\mathcal{B}'$ to be the empirical distribution over all houses with a predicted price below the first quartile. Consequently, the prediction gaps increased and 97% of the houses that were previously unexplainable suddenly became explainable.

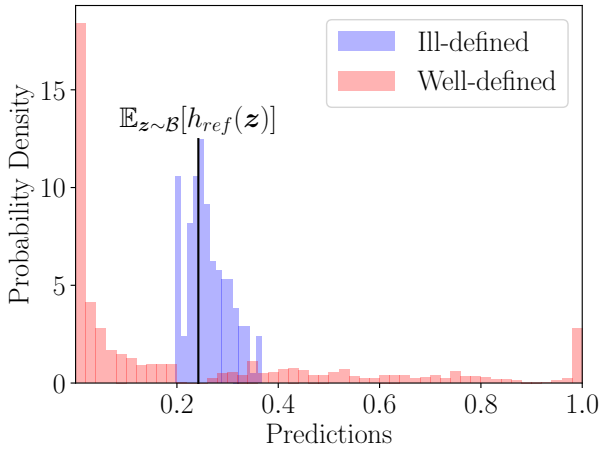


Figure 16: Distributions of predictions for instance with ill-defined and well-defined gaps across the Rashomon Set for Adult-Income. The background $\mathcal{B}$ is the empirical distribution over 500 uniform samples from the training data.

C.2 Adult-Income

As a reminder, the background $\mathcal{B}$ employed on Adult-Income was 500 instances sampled uniformly at random from the training data. Figure 16 shows the distributions of predictions for individuals whose gap is well-defined or ill-defined across the Rashomon Set. Again, individuals whose gap is ill-defined tend to have predictions $h_{\text{ref}}(\boldsymbol{x}^{(i)})$ near the baseline $\mathbb{E}_{\boldsymbol{z}\sim\mathcal{B}}[h_{\text{ref}}(\boldsymbol{z})]$ so that the Gap $G(h_{\text{ref}}, \boldsymbol{x}^{(i)})$ is small. Once more, we replace the background and re-explain those inputs. Letting $\mathcal{B}'$ be 500 uniformly-chosen adults who were predicted to make more than 50K (*i.e.* $h_{\text{ref}}(\boldsymbol{x}^{(i)}) > 0.5$), the gaps became highly negative and 100% of the previously unexplainable individuals were now explainable.

References

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.

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

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001b.

Weiwei Cheng, Michaël Rademaker, Bernard De Baets, and Eyke Hüllermeier. Predicting partial orders: ranking with abstention. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 215–230. Springer, 2010.

Beau Coker, Cynthia Rudin, and Gary King. A theory of statistical inference for ensuring the robustness of scientific results. *Management Science*, 67(10):6174–6197, 2021.

Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223, 2020.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

Jiayun Dong and Cynthia Rudin. Variable importance clouds: A way to explore variable importance for the set of good models. *arXiv preprint arXiv:1901.03209*, 2019.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, pages 1–12, 2021.

Geanderson Esteves, Eduardo Figueiredo, Adriano Veloso, Markos Viggiato, and Nivio Ziviani. Understanding machine learning software defect predictions. *Automated Software Engineering*, 27(3):369–392, 2020.

Thomas Fel, David Vigouroux, Rémi Cadène, and Thomas Serre. How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. 2021.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Hsiang Hsu and Flavio du Pin Calmon. Rashomon capacity: A metric for predictive multiplicity in probabilistic classification. *arXiv preprint arXiv:2206.01295*, 2022.

Naira Kaieski, Cristiano Andre da Costa, Rodrigo da Rosa Righi, Priscila Schmidt Lora, and Bjoern Eskofier. Application of artificial intelligence methods in vital signs analysis of hospitalized patients: A systematic literature review. *Applied Soft Computing*, page 106612, 2020.

Nicholas Kissel and Lucas Mentch. Forward stability and model path selection. *arXiv preprint arXiv:2103.03462*, 2021.

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

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica*, May 2016. URL `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`.

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.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1): 56–67, 2020.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pages 6765–6774. PMLR, 2020.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Artyom G Nahapetyan. Bilinear programming., 2009.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

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.

Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *arXiv preprint arXiv:1811.00731*, 2018.

Azar Salih, Subhi T Zeebaree, Sadeeq Ameen, Ahmed Alkhyyat, and Hnan M Shukur. A survey on the role of artificial intelligence, machine learning and deep learning for cybersecurity attack detection. In *2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic"(IEC)*, pages 61–66. IEEE, 2021.

Jonas Schulz, Rafael Poyiadzi, and Raul Santos-Rodriguez. Uncertainty quantification of surrogate explanations: an ordinal consensus approach. *arXiv preprint arXiv:2111.09121*, 2021.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858, 2022.

Torgyn Shaikhina, Umang Bhatt, Roxanne Zhang, Konstantinos Georgatzis, Alice Xiang, and Adrian Weller. Effects of uncertainty on the quality of feature importance explanations. *AAAI Workshop on Explainable Agency in Artificial Intelligence*, 2021.

Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, pages 307–317, 1953.

Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems*, 34, 2021.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017.

Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. Statistical stability indices for lime: obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, pages 1–11, 2020.

Larry Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

Zhengze Zhou, Giles Hooker, and Fei Wang. S-lime: Stabilized-lime for model explanation. *arXiv preprint arXiv:2106.07875*, 2021.