Partial order: Finding Consensus among Uncertain Feature Attributions

Gabriel Laberge[1]
Polytechnique Montréal

Yann Pequignot
Université Laval

Alexandre Mathieu
Université Laval

Mario Marchand
Université Laval

Foutse Khomh
Polytechnique Montréal

Abstract

Post-hoc feature importance is progressively being employed to explain decisions of complex machine learning models. Yet in practice, reruns of the training algorithm and/or the explainer can result in contradicting statements of feature importance, henceforth reducing trust in those techniques. A possible avenue to address this issue is to develop strategies to aggregate diverse explanations about feature importance. While the arithmetic mean, which yields a total order, has been advanced, we introduce an alternative: the consensus among multiple models, which results in partial orders. The two aggregation strategies are compared using Integrated Gradients and Shapley values on two regression datasets, and we show that a large portion of the information provided by the mean aggregation is not supported by the consensus of each individual model, raising suspicion on the trustworthiness of this practice.

1 INTRODUCTION

The Machine Learning (ML) framework has proven to be an essential tool in many data intensive domains such as software engineering, medecine, 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 issue that limits potential trust in ML systems. For this reason, various model-agnostic techniques such as LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), and Integrated Gradient (Sundararajan et al., 2017) have recently been developed to provide explanations of model decisions in the form of feature attributions. These attributions are meant to indicate the contribution (positive/negative) of individual features toward the model prediction, and their magnitudes (positive) 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 settings, it has become apparent that they are subject to variability. Indeed, recomputing the explanation and/or retraining the model can induce a high variance in feature attributions (Visani et al., 2020; Shaikhina et al., 2021), and contradictions in feature importance rankings (Fel et al., 2021; Lyu et al., 2021; Zhou et al., 2021). This is a critical issue for ML explainability: if two practitioners who use the exact same data and learning algorithm cannot agree on the feature attributions/importance, how can they trust these explanations?

Information can be extracted from diverse explanations by aggregating them using various mechanisms. A common aggregation strategy is the arithmetic mean of feature attributions, which was previously used to combine explanations obtained from multiple reruns of the explainer (Molnar, 2020, Chapter 14), explanations of points in the neighborhood of the input of interest (Bhatt et al., 2020) and explanations of independently trained models (Shaikhina et al., 2021).

The alternative combination mechanism we introduce is two-fold. One the one hand, to mitigate variability in the explainer, we compute Confidence Intervals (CIs) that capture the ground-truth attributions with high probability. On the other hand, to address variability in explanations of diverse models, we take the **consensus** of feature importance among independently trained models. Both strategies are used in tandem, resulting in **partial** orders of feature importance (instead of the

[1] Main correspondance: gabriel.laberge@polymtl.ca

more common total orders). These partial orders have the ability of abstaining from considering a feature as more important than another, making them useful at ignoring misleading information. Moreover, they are constructed in such a way that the information they do contain is meaningful, with high probability.

On two regression datasets, the mean and consensus strategies are applied to the SHAP and Integrated Gradient explanations of Multi-Layered Perceptrons (MLP) and the following observations are made.

1. The partial orders resulting from the consensus aggregation are informative (structured), even when combining explanations of about 1K independently trained MLPs.
2. The total orders induced by the mean aggregation contain information that is not supported by the consensus, while information in the latter is supported by the former. This raises doubts on the trustworthiness of the mean strategy.

The remainder of this paper is organised as follows: **Section 2** presents the necessary background knowledge, while **Section 3** discusses the sources of variability (uncertainty) in feature attributions as well as aggregation strategies. **Section 4** lays out the experimental details, results and discussions, and finally **Section 5** concludes the paper.

2 BACKGROUND

Feature attribution methods allow to order features based on their importance in the prediction of a single black-box model. However the ML methodology inevitably leads to a diversity of models and our approach to feature importance in this context relies on the notion of partial order.

2.1 Additive Feature Attribution

The rapid growth of the field of eXplainable Artificial Intelligence (XAI) is primarily driven by the desire to *explain* black box models (Arrieta et al., 2020). Formally, given an input space $\mathcal{X} \subseteq \mathbb{R}^d$, an output space $\mathcal{Y}$, a black box model $h : \mathcal{X} \to \mathcal{Y}$, and a specific input of interest $\boldsymbol{x} \in \mathcal{X}$, there is growing interest in capturing the reasons/logic behind the decision $h(\boldsymbol{x})$. One family of techniques that aim at providing such information are feature attributions, which are vector-valued functionals[1] $\boldsymbol{\phi} : \mathcal{H} \to \mathbb{R}^d$ whose vector output represents the contribution of each feature towards the prediction $h(\boldsymbol{x})$.

[1]Vector $\boldsymbol{\phi}$ also depends on $\boldsymbol{x}$, but to simplify notation, we make this dependence implicit from the context.

This paper focuses on *additive feature attributions* which are further required to distribute the difference between the prediction $h(\boldsymbol{x})$ and the average prediction with respect to a chosen background distribution $\mathcal{B}$ over $\mathcal{X}$ among the features, namely feature attributions which satisfy

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

The component $\phi_i(h)$ is interpreted as the increase/decrease of the model w.r.t the background we can attribute to feature i. Additive feature attributions are motivated by a contrastive question of the form: *Why is $h(\boldsymbol{x})$ so high/low compared to the average prediction on $\mathcal{B}$?*

The notion of feature importance refers to the magnitude of attributions $|\phi_i(h)|$, and can be used to define total orders $\leq_{\boldsymbol{\phi}(h)}$ commonly used to rank features (Ribeiro et al., 2016; Bhatt et al., 2020; Molnar, 2020; Jiarpakdee et al., 2020)

$$i <_{\boldsymbol{\phi}(h)} j \iff |\phi_i(h)| < |\phi_j(h)|. \tag{2}$$

We now recall two additive attribution methods: Shapley values and Integrated Gradients, that are both increasingly being used to explain black boxes.

2.1.1 Shapley Values

The Shapley values are a fundamental concept from cooperative game theory (Shapley, 2016). Letting $[d]$ 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

$$r_P(\boldsymbol{z}, \boldsymbol{x})_i = \begin{cases} x_i & \text{if } i \in P \\ z_i & \text{otherwise.} \end{cases} \tag{3}$$

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 contribution of specifying the ith feature from the background distribution across all coalitions

$$\phi_i^{\text{Shap}}(h) = \mathop{\mathbb{E}}_{\substack{\pi\sim\Omega \\ \boldsymbol{z}\sim\mathcal{B}}} \left[h(\, \boldsymbol{r}_{\pi_{:i}\cup\{i\}}(\boldsymbol{z}, \boldsymbol{x})\,) - h(\, \boldsymbol{r}_{\pi_{:i}}(\boldsymbol{z}, \boldsymbol{x})\,) \right], \tag{4}$$

where Ω is the uniform distribution over all $d!$ permutations of the features.

2.1.2 Integrated Gradient

The Integrated Gradient (IG) 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 the IG is

$$\phi_i^{\mathrm{IG}}(h) := \mathop{\mathbb{E}}_{\substack{\boldsymbol{z}\sim\mathcal{B},\\ \alpha\sim U(0,1)}} \left[(x_i - z_i) \frac{\partial h}{\partial x_i}\Big|_{\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{z}} \right] \quad (5)$$

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.

2.2 Underspecification and Ensembling

Many ML algorithms consist in stochastic methods for learning a model from data and therefore potentially lead to a diversity of models. For simplicity, this paper focuses only on the stochasticity introduced by optimization procedures for MLPs. This includes the random initialization of the parameters, and the random order in which mini-batches are fed to the optimizer. So for given training data S and a learning algorithm $\mathcal{A}_\Phi$ with carefully selected hyperparameters Φ, we can describe the probability distribution $\mathcal{A}_\Phi(S)$ over MLPs via its sampling procedure $h \sim \mathcal{A}_\Phi(S)$: randomly initialise the parameters of a model, perform stochastic gradient descent on the parameters to minimize the loss on S, and retrieve h.

The fact that practitioners, even with fine-tuned hyperparameters, can end up with a distribution over diverse MLPs with equivalent test performance is indicative of the underspecification of the task. The absence of evidence to choose one model over the others can be circumvented by the introduction of an aggregate predictor h_E obtained by averaging predictions

$$h_E(\boldsymbol{x}) := \frac{1}{M}\sum_{k=1}^{M} h^{(k)}(\boldsymbol{x}) \quad \text{with} \quad h^{(k)} \sim \mathcal{A}_\Phi(S), \quad (6)$$

where $E := \{h^{(k)}\}_{k=1}^M$ refers to an ensemble of models. A common justification for Eq. 6 is that averaging high complexity models acts as a regularizer, meaning h_E often performs better on the test set than any individual model in E (Goodfellow et al., 2016, Section 7.11). However, a recent study suggests that in the presence of a mismatch between training/testing and deployment domains, there is a considerable variance of deployment performance between independently trained models, which could reduce the benefits of aggregating predictors (D'Amour et al., 2020). In light of those results, D'Amour et al. suggest that the whole set of good predictors that a learning algorithm can return carries valuable information about underspecification of the ML pipeline.

When feature attribution techniques like SHAP and IG are used to explain the predictions of different models trained on the same data with the same learning algorithm, they can produce different results (Shaikhina et al., 2021), hence suggesting that underspecification also affects feature importance. We argue this underspecification could be induced by a mismatch between the training/testing and explanation domains. Indeed, computing these feature attributions requires evaluating the models on distributions of inputs that differ from the dataset (see **Appendix C**). Note that this domain mismatch has previously been observed and exploited to attack/manipulate SHAP attributions (Slack et al., 2020a). While one could simply decide to consider the feature importance of the single aggregated model h_E to deal with underspecification, the considerations of D'Amour et al. encourage us to map out explanations of all individual models. Combining the explanations of each individual model will naturally lead us to a *partial order* of feature importance, which we now briefly introduce.

2.3 Partial Orders

Definition 2.1. *A partial order $\preceq$ on the set of features $[d]$ is a binary relation that is reflexive, antisymmetric and transitive i.e. for all $i, j, k \in [d]$, we have:*

1. $i \preceq i$ *(reflexivity)*
2. $i \preceq j$ *and* $j \preceq i \implies i = j$ *(antisymmetry)*
3. $i \preceq j$ *and* $j \preceq k \implies i \preceq k$ *(transitivity)*

The subtlety about partial orders is that it is not required that any two features i and j be comparable. Indeed, if neither $i \preceq j$ nor $j \preceq i$ hold, i and j are **incomparable**, denoted by $i \perp j$. The associated strict relation, symbolized with $i \prec j$, is defined by $i \preceq j$ and $i \neq j$. Total orders $\leq$ (typically used for feature importance) can simply be viewed as special cases of of partial orders, those in which every pair of features is comparable.

The total number of ordered pairs of features that are strictly comparable is called the cardinality of the order $|\preceq| := |\{(i,j) \in [d]^2 : i \prec j\}|$. This quantity takes values between 0 (all features are incomparable), and $d(d-1)/2$ (total order).

With partial orders, we are interested in the two distinct concepts of maximum element and maximal element, which coincide for total orders. Feature i is called maximum if $j \prec i$ for all other features j, while feature i is called maximal if there exists no feature considered more important, i.e. $\nexists j$ s.t. $i \prec j$. Note that a maximum element must be unique and is always maximal,

but when there is no maximum there can be several maximal elements.

A common way to picture a partial order on a finite set is to draw its Hasse diagram.

Definition 2.2. *Given a set of features $[d]$ and a partial order $\preceq$, the associated Hasse diagram $\mathcal{G}(\preceq)$ is a directed graph with vertices $\mathcal{V} := [d]$, and directed edges*

$$\mathcal{E} := \big\{(i,j) \in [d]^2 : \; i \prec j \text{ and } \nexists\, k \in [d] \text{ such that } i \prec k \prec j\big\}. \tag{7}$$

Finally, one of the fundamental properties of partial orders is that they can be intersected to make new ones.

Definition 2.3. *Given a set of partial orders $\{\preceq_k\}_{k=1}^M$, the intersection of these relations $\bigcap_k \preceq_k$ is the binary relation*

$$i \bigcap_{k=1}^{M} \preceq_k \; j \iff i \preceq_k j \quad \forall k = 1, 2, \dots, M\,, \tag{8}$$

which is also a partial order.

The intersection operation formalizes the concept of **consensus** between various orders.

3 ATTRIBUTION UNCERTAINTY

We now discuss two sources of variability in feature attributions: the explainer and model uncertainties, whose existence motivates the development of mechanisms to aggregate diverse explanations

3.1 Explainer Uncertainty

In the majority of practical settings, the attributions presented in Equations 4 and 5 are intractable and must be approximated via the Monte Carlo (MC) method (details in **Appendix B**). Therefore, in experiments, an estimator $\widehat{\boldsymbol{\phi}}$ of $\boldsymbol{\phi}$ will be computed and used to order features as in Eq. 2. Because MC is random in essence, the feature attributions $\widehat{\boldsymbol{\phi}}$ will fluctuate and result in instability of the total order on features.

Fortunately, MC allows for the construction of confidence intervals (CIs) I_i around $\widehat{\phi}_i$ to capture the true value ϕ_i with high probability, i.e., $\mathbb{P}\big(\phi_i(h) \notin I_i\big) \leq \delta$. When CIs are considered, ordering features by relative importance naturally leads us to a first partial order.

Definition 3.1. *Given the approximated feature attribution $\widehat{\phi}_i(h)$, and their CIs I_i, the partial ordering $\preceq_{\widehat{\boldsymbol{\phi}}(h)}$ is defined as*

$$i \preceq_{\widehat{\boldsymbol{\phi}}(h)} j \iff \max |\, I_i \,| \leq \min |\, I_j \,|, \tag{9}$$

where $|I_i|$ refers to the image of the CI under the absolute value map.

The following proposition shows that, with high probability, this partial order is unlikely to contradict the true total order $\leq_{\boldsymbol{\phi}(h)}$ from Eq. 2.

Proposition 3.1. *Suppose that $j <_{\boldsymbol{\phi}(h)} i$ and we have estimates $\widehat{\phi}_i(h), \widehat{\phi}_j(h)$ of the true attributions $\phi_i(h), \phi_j(h)$ with CIs I_i, I_j each of significance δ, then $\mathbb{P}\big(i \prec_{\widehat{\boldsymbol{\phi}}(h)} j\big) \leq 2\delta$.*

3.2 Model Uncertainty

The second source of variability is the randomness inherent to the learning algorithm $\mathcal{A}_\Phi(S)$, which is hard to characterize because it depends on the inner workings of the optimizer and the training data. For this reason, we discuss aggregation strategies to combine explanations of various models.

3.2.1 Mean Aggregation

It has recently been proposed to aggregate explanations of independently trained models by taking arithmetic mean of their feature attributions (Shaikhina et al., 2021), which is equivalent to explaining the aggregated model, as long as the feature attributions are linear w.r.t the model, which is the case for SHAP/IG attributions[2].

Definition 3.2. *Given a linear feature attribution $\widehat{\boldsymbol{\phi}}$, the mean aggregation on E is defined as the feature attribution of the aggregated model h_E. It yields the total order denoted by $\leq_{\widehat{\boldsymbol{\phi}}(h_E)}$.*

This aggregation strategy faces two main challenges. First of all, there can be high variability between attributions of independently trained models (Shaikhina et al., 2021), a useful information that is completely lost when averaging.

Secondly, the order $\leq_{\widehat{\boldsymbol{\phi}}(h_E)}$ lacks theoretical guarantees that, as new models are added to the ensemble, no order relations will suddenly switch. Formally, if $E \subseteq E'$ it can happen that $i <_{\widehat{\boldsymbol{\phi}}(h_{E'})} j$ while $j <_{\widehat{\boldsymbol{\phi}}(h_E)} i$.

3.2.2 Consensus Aggregation

Given the identified shortcomings of the mean aggregation, we introduce another method, the **consensus strategy**, which takes into account the feature importance of each individual model $\bigcap_{k=1}^M \preceq_{\widehat{\boldsymbol{\phi}}(h^{(k)})}$, leading to the following key definition.

[2]This is true for $\boldsymbol{\phi}$ and only true for $\widehat{\boldsymbol{\phi}}$ if the same random seed is used when explaining all models.

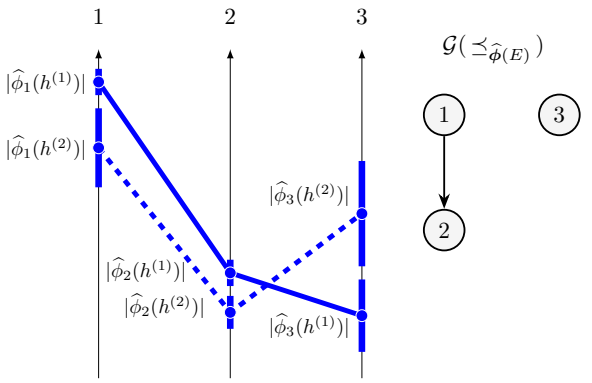


Figure 1: Partial order computed via consensus among two models and by considering CIs.

Definition 3.3. *Given an ensemble E, we define the consensus aggregation as $\preceq_{\widehat{\boldsymbol{\phi}}(E)} := \bigcap_{h\in E} \preceq_{\widehat{\boldsymbol{\phi}}(h)}$. Similarly we let $\leq_{\boldsymbol{\phi}(E)} := \bigcap_{h\in E} \leq_{\boldsymbol{\phi}(h)}$.*

An example of partial order computed via Definition 3.3 is illustrated in Figure 1. One of the crucial properties of the consensus aggregation is that relations cannot flip when new models are added to the ensemble.

Proposition 3.2. *If $E \subseteq E'$, the only possible disagreements between $\preceq_{\widehat{\boldsymbol{\phi}}(E)}$ and $\preceq_{\widehat{\boldsymbol{\phi}}(E')}$ are relations $i \prec_{\widehat{\boldsymbol{\phi}}(E)} j$ for which $i \perp_{\widehat{\boldsymbol{\phi}}(E')} j$.*

In other words, when adding new models to an ensemble, an existing relation between two features cannot be inverted, it can only cease to hold, therefore making them incomparable. Moreover if a relation between two features does not hold, adding new models to the ensemble will not change this fact.

One parameter left to specify for the consensus strategy is the level of confidence δ of the CIs. Considering an ensemble of M models, if the true attributions were known, one would have a ground-truth order $\leq_{\boldsymbol{\phi}(E)}$. However, in practice, one obtains approximated orders $\preceq_{\widehat{\boldsymbol{\phi}}(E)}$, which are unlikely to be identical to the ground-truth. Since relations in the partial order are meant to be informative, we at least want to ensure that every edge in the approximation is present in the ground-truth.

Proposition 3.3. *Given an ensemble $E = \{h^{(k)}\}_{k=1}^{M}$ and estimates $\widehat{\phi}_i(h^{(k)}), \widehat{\phi}_j(h^{(k)})$ of the true attributions $\phi_i(h^{(k)}), \phi_j(h^{(k)})$ with CIs $I_i^{(k)}, I_j^{(k)}$ each of significance δ, if $i \not\prec_{\boldsymbol{\phi}(E)} j$, then $\mathbb{P}\big(i \prec_{\widehat{\boldsymbol{\phi}}(E)} j\big) \leq 2\delta$.*

3.3 Comparing Aggregation Strategies

Comparing the mean and consensus aggregations is not an easy task because they result in partial and total orders. This difficulty is exacerbated by the lack of an oracle feature importance to which both methods can be contrasted. In face of this challenge, we decided to build a pseudo-oracle by labeling order relations supported by both methods as trustworthy. For an ensemble E of models we consider the number of trustworthy relations $T_E = |\leq_{\widehat{\boldsymbol{\phi}}(h_E)} \cap \preceq_{\widehat{\boldsymbol{\phi}}(E)}|$ and use the ratios $R_M = T_E/|\leq_{\widehat{\boldsymbol{\phi}}(h_E)}|$ and $R_C = T_E/|\preceq_{\widehat{\boldsymbol{\phi}}(E)}|$ as measures of trustworthiness of the mean and consensus orders.

Proposition 3.4. *Given an ensemble $E = \{h^{(k)}\}_{k=1}^{M}$ and estimate attributions $\widehat{\phi}_i(h^{(k)}), \widehat{\phi}_j(h^{(k)})$ with CIs $I_i^{(k)}, I_j^{(k)}$, if the estimates $\widehat{\phi}_i(h^{(k)})$ have the same sign for all models (and similarly for feature j), then $i \prec_{\widehat{\boldsymbol{\phi}}(E)} j \implies i <_{\widehat{\boldsymbol{\phi}}(h_E)} j$.*

Although the proposition does not guarantee that $R_C = 1$, it still suggests that R_C should be large in practice.

3.4 Convergence of Partial Orders

Determining the ultimate partial order of feature importance would virtually require all possible models that can be sampled, even though a finite number of them (which is bounded by the number of features) would then already define it. The question therefore arises as to how many models are required to stabilize the partial orders. Considering our inability to determine these ultimate partial orders, we decide to characterize convergence in a simplified setting. The ultimate orders are chosen to be the partial orders computed using an extremely large ensemble E. One can then collect smaller ensembles $E_M \subset E$ consisting of M models randomly picked without replacement from E, and count the number of disagreements between the two orders $\preceq_{\widehat{\boldsymbol{\phi}}(E_M)}$ and $\preceq_{\widehat{\boldsymbol{\phi}}(E)}$ (cf. Proposition 3.2). Repeating this process for increasing values of M will give an idea of the number of models required to stabilize the partial order.

4 EXPERIMENTS

The consensus mechanism proposed in this paper is compared to the mean aggregation using Shapley values and IG on two regression datasets, and it is shown that the total order induced by the mean contains untrustworthy information.

4.1 Experimental Setup

Given a dataset, we start by splitting the instances into train/tests sets with 90% and 10% ratios respectively. Afterward, a MLP architecture is chosen: for datasets with fewer than 10K instances, a single hidden layer with 100 units was used, while for larger datasets, four hidden layers of sizes $[100, 50, 20, 10]$ were considered. For simplicity, the Adam optimizer with default hyperparameters was used, and a grid-search over learning rates, number of epochs, and choice of learning rate scheduler was conducted. With the chosen hyperparameters, around 1000 models were trained and a pool of models E was formed with all MLPs that had a higher test-set performance than a Random Forest baseline. This choice was made because the learning algorithm would sometimes, but rarely, yield models with terrible performances (even worse than linear models).

After obtaining the pool E, we compute the additive feature attributions of the predictions $h(\boldsymbol{x})$ $\forall h \in E$ on some observations of interest $\boldsymbol{x}$ taken from the test set. Following the *Formulate-Estimate-Explain* (FAE) framework introduced for SHAP/IG explanations (Merrick and Taly, 2020), we apply the following steps:

Formulate Given an instance $\boldsymbol{x}$ to explain, formulate a *contrastive* question, which directly specifies the background distribution $\mathcal{B}$ for the additive attribution.

Estimate Feature attributions $\boldsymbol{\phi}$ are estimated with simple MC yielding $\widehat{\boldsymbol{\phi}}$ and CIs I_i. To assess whether or not enough MC samples are used, the relative attribution error $(\Delta - \sum_{i=1}^{d} \widehat{\phi}_i(h))\Delta^{-1}$, where $\Delta = h(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{B}}[h(\boldsymbol{z})]$, is reported for each model h and used as a proxy of MC error. During our experiments, when the attribution error for a model was above 2%, this attribution was rejected and not considered in the aggregation.

Explain Given MC estimates and CIs (with significance level $\delta = 0.1\% \times [d(d-1)]^{-1}$, cf. union bound on Proposition 3.3), feature importance is visualised using a Hasse diagram $\mathcal{G}(\preceq_{\widehat{\boldsymbol{\phi}}(E)})$. For a visual comparisons with $\preceq_{\widehat{\boldsymbol{\phi}}(h_E)}$, the mean attribution of individual features $\widehat{\phi}_i(h_E)$ are reported in the graph nodes, and used as color in a hot/cold heatmap. For quantitative comparisons between the orders, the trustworthiness ratios R_M and R_C are reported.

Finally, to observe the convergence of partial orders, the process discussed in **Section 3.4** was repeated 10,000 times for every value of $M \in \{1, 2, 4, 8, \ldots, 256, 512\}$, and each time recording the number of disagreements between $\preceq_{\widehat{\boldsymbol{\phi}}(E_M)}$ and $\preceq_{\widehat{\boldsymbol{\phi}}(E)}$.

We now present results from the two regression datasets separately, followed by a discussion of the merits and limitation of our proposed approach. Additional results are presented in **Appendix E**.

4.2 House Price Prediction

The Houses regression dataset, available on Kaggle[3], consists of predicting the selling price of 1446 houses based on 79 numerical and categorical features. For simplicity, only numerical features were selected and those with high Spearman correlation (> 0.8) were removed, leading to a total of 20 features. The IG attribution was considered for this dataset. After some data exploration, we became interested in explaining the price of the costliest house from the test set.

Formulate The instance $\boldsymbol{x}$ to explain was the test set house with the largest selling price (430,000 USD predicted as 371,953 USD by h_E), and was compared to the training set $\mathcal{B} := \{\text{Training Set}\}$.

Approximate For MC approximation of IG, a total of 50,000 samples where used leading to relative attribution errors of at most 0.5%. Results for SHAP are reported in **Appendix E.1.1**.

Explain Figure 3 illustrates the feature importance of each model in the pool of size 1115, as well as for the aggregate predictor h_E, while Figure 2 presents the Hasse diagram of the corresponding consensus partial order $\mathcal{G}(\preceq_{\widehat{\boldsymbol{\phi}}(E)})$ whose nodes are colored w.r.t the mean aggregation. Given the high trustworthiness ratio of the partial order ($R_C = 1$), we can analyse the structure of the Hasse diagram with confidence.

First of all, while `1stFlrS=very large` has the largest attribution variance (as evidenced by the spread of blue lines at the top of Figure 3), our approach brings to light that there is consensus among models that the feature has maximum importance. This provides strong evidence that the surface of the first floor explains most of the high price of the house. Secondly, feature `OverallQual=8` (quality of materials and finish of the house from a scale of 1 to 10) is more important than 13 other features. Although this feature is of lesser importance compared to first floor surface, this observation still suggests that material/finish quality is primordial to increase this house selling price. The third most important feature according to mean aggregation is the surface of the second floor `2ndFlrSF=0`, which is given an average negative attribution of $-17,324$ USD. This can appear intuitive at first because one can imagine the house's price would increase if it had a second floor.

[3] https://www.kaggle.com/c/house-prices-advanced-regression-techniques

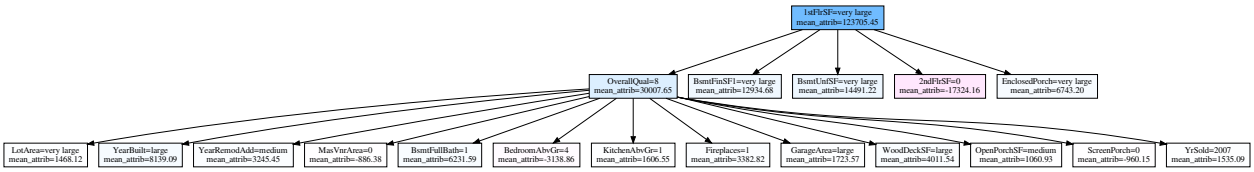


Figure 2: Hasse diagram for an instance in Houses, trustworthiness ($R_M = 34/190, R_C = 34/34$).

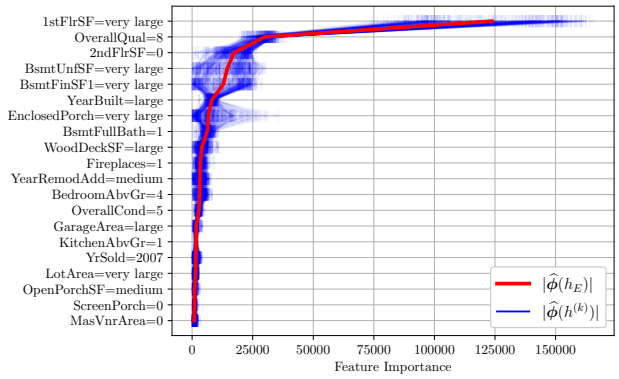


Figure 3: IG feature importance for a Houses instance.

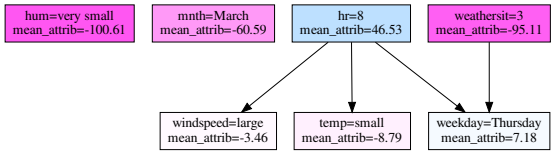


Figure 4: Hasse diagram for a Bikes instance, trustworthiness ($R_M = 4/21, R_C = 4/4$).

However, our approach reveals that the relative importance of this feature varies among models and as result this feature is not considered more important than any other feature. Therefore, to stay cautious, we would abstain from stating that the absence of a second floor explains a substantial drop in the house's price.

Figure 6 (left) shows the convergence of the partial order with up to 512 models considered out of 1115 models. We observe that it takes up to 128 models to guarantee less than 30/190 ($\sim 16\%$) disagreements between the resulting diagram and its limit. Most strikingly, considering a single model leads to around 145 order relations that would disappear when adding more models.

4.3 Bikes Rental Forecast

The Bikes (aka Bike Sharing) dataset available on the UCI repository[4] aims at predicting the hourly count of bike rentals between years 2011 and 2012 in the

[4] https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset

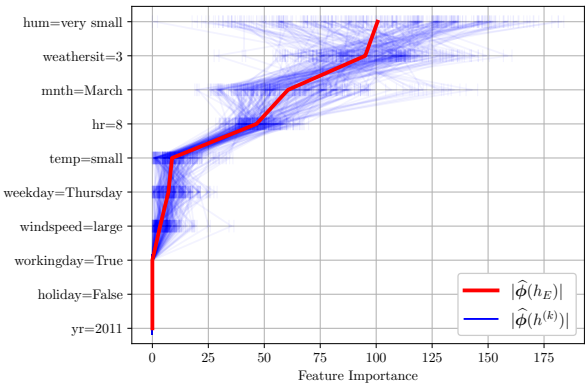


Figure 5: SHAP feature importance for Bikes instance.

Washington state, based on time and weather features. Over all available features, only the felt-temperature was removed because of its high correlation with temperature, leading to 10 features total. During data exploration, it was reported that a lot of bikes were rented at 7h-8h on `WorkingDay=True`. However, there exists an instance in this time period with very low amounts of bike rentals, which we are going to explain.

Formulate Explain the test set instance $\boldsymbol{x}$ with the least amount of bike rentals (44 bikes predicted as 63 by h_E) between 7-8h on workingdays in 2011. The background $\mathcal{B}$ is chosen as the whole training set conditioned on `yr=2011`, `workingday=True` and $7 \leq$ `hr` ≤ 8. It was decided to fix the year to 2011 in the background in order to find attributions that are more relevant to the contrastive question.

Approximate MC was conducted by independently running the `Permutation` explainer of the SHAP library[5] 400 times, with each call using a subset of 100 random instances from the background. As explained in **Appendix B.3**, this procedure is equivalent to simple MC with 400 samples. The highest relative attribution error reported was 2%. Results for IG are discussed in **Appendix E.2.1**.

Explain The feature importance of all models in the pool of size 988 is shown in Figure 5 and the Hasse dia-

[5] https://shap.readthedocs.io/en/latest/generated/shap.explainers.Permutation.html

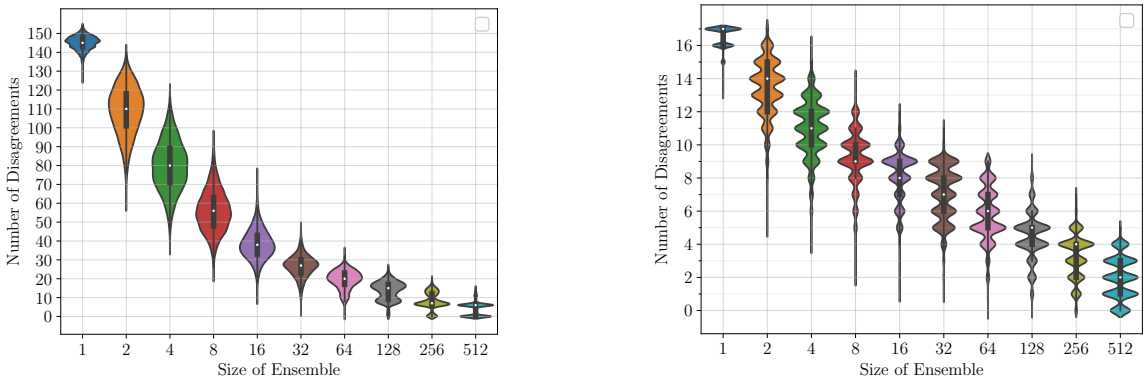


Figure 6: Number of disagreements between the ultimate order $\preceq_{\widehat{\phi}(E)}$ and the orders $\preceq_{\widehat{\phi}(E_M)}$ of ensemble increasing in size, for Houses (left) and Bikes (right).

gram is presented in Figure 4. Once again, we observe that all information contained in the partial order is supported by the total one ($R_C = 1$), while only one fifth of the information in the total order could be considered trustworthy. More specifically, we note that feature `humidity=very small` is found to be given the most importance by h_E to explain the drop in bike rentals. By itself this observation is already quite suspicious because prior knowledge suggests that people would rather be more inclined to rent bikes when humidity is low. Looking at Figure 5, we observe an important variability in the importance that different models attribute to humidity, and looking even closer, we can see that models further disagree on the relative importance of humidity compared to other features, which is why the feature is incomparable to any other in the partial order. It appears that presenting `humidity` as the most important feature, as done by the arithmetic mean, may be problematic and risky. Similar observations can be made for feature `mnth=March`. This example shows that using the arithmetic mean can be misleading as the relation of two of the top-three most important features to other features are not supported by the consensus strategy.

Figure 6 (right) illustrates the convergence of the Hasse diagram. As previously, hundreds of models are needed to obtain stability, e.g. it takes 256 models to ensure that less than 7/45 ($\sim$ 15%) of the order relations disagree with the limit. When only considering a single model, we see that about 17 order relations would disappear as the ensemble size is increased, only leaving 4 relations in the limit.

4.4 Discussion

The main takeaway from the case studies is that combining attributions via the arithmetic mean (or equivalently explaining the aggregate model h_E) often yields untrustworthy information about feature importance that is not supported by the consensus strategy. Moreover, we find that partial orders contain information (i.e. there are features which are deemed more important by every single model), even when considering up to $\sim$1K MLPs, which suggests that SHAP and IG feature attributions can still drive decision-making despite being subject to variability.

As presented in Figure 6, hundreds of models are required for the partial orders to stabilize, which is quite high compared to the size of ensembles commonly used by practitioners. However, we argue that reporting the number of relations present in $\preceq_{\widehat{\phi}(E_M)}$ that are missing in the limit is a pessimistic measure of diagram stability. Indeed, some differences between $\preceq_{\widehat{\phi}(E_M)}$ and $\preceq_{\widehat{\phi}(E)}$ may be more critical than others for end-users. For example, disagreements of order relations between features deemed irrelevant by domain knowledge may have a lesser impact on interpretability than disagreements between features relevant to the user. It is therefore necessary to study more task-specific measures of stability for partial orders.

The performance of individual models is of high concern when combining explanations via models consensus because it suffices that a **single** model disagrees with all others to make two features incomparable. For this reason, there is a strong assumption that all models are equally as trustworthy. In this study we used the performance of a Random Forest baseline to fix an upper bound and select acceptable models, but the exact role of performance should be investigated further in the future.

4.5 Related Work

Confidence intervals for feature attributions have recently been applied for LIME (Visani et al., 2020) and SHAP (Merrick and Taly, 2020). To the best of our knowledge, we are the first to apply said confidence

intervals to define partial orders of feature importance, via Definition 3.1. Moreover, confidence intervals with $\delta = 5\%$ are systematically encouraged in the literature, although applying the union bound to Proposition 3.3 suggests that δ should decrease by a factor $d(d-1)$. Reducing δ so drastically will increase the width of the intervals, resulting in less relations in the partial order. Nonetheless, we see this as a natural price to pay in order to ensure remaining edges convey trustworthy information.

The explanation uncertainty induced by the diversity of good models has previously been investigated (Shaikhina et al., 2021; Fisher et al., 2019). While Shaikhina et al. characterize said uncertainty by computing the variance of feature attributions among independently trained models, Fisher et al. search for the maximum and minimum feature importance among all models with acceptable performance. Our approach is fundamentally different in that we are not measuring the variability of attributions of each single feature feature i, but focus instead on whether or not every single model agrees that feature i is more important than feature j. Moreover, as presented in the case study of **Section 4.2**, the feature with highest variance in attribution, can still be the most informative according to the partial order, suggesting attribution variance offers an incomplete view of explanation uncertainty.

An alternative to the frequentist approach to attribution uncertainty described in **Section 3.1** is offered by BayesSHAP (Slack et al., 2020b), which leverages principles from Locally Weighted Bayesian Regression to compute posterior distributions over plausible Shapley values. However, while BayesSHAP can provide confidence intervals for attributions, the method does not consider model uncertainty and focuses on explaining a single model.

5 CONCLUSION

In this work, sources of uncertainty for feature attributions were described. A novel aggregation mechanism based on consensus among models was introduced and compared to the arithmetic mean. It was observed that the total orders resulting from the mean attributions can contain possibly misleading information that is absent from the more conservative partial orders provided by the consensus aggregation. Moreover, the partial orders were shown to be informative (structured), even when considering $\sim$1K MLPs, suggesting that there exists partial order in the seeming chaos of models produced by a learning algorithm and that IG/SHAP attributions are useful tools despite suffering from variability.

As future work, we intend to study other post-hoc explanations (LIME, SmoothGrad, Permutation Importance, Entropic Variable Projection), employ alternative models such as Gradient Boosted Trees, and apply our methodology to more practical settings on which the nuance introduced by partial orders will hopefully prove most beneficial.

References

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.

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.

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.

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.

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.

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.

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.

Thomas Fel, David Vigouroux, Rémi Cadène, and Thomas Serre. How good is your explanation? al-

gorithmic stability measures to assess the quality of explanations for deep neural networks. 2021.

Yinggzhe Lyu, Gopi Krishnan Rajbahadur, Dayi Lin, Boyuan Chen, and Zhen Ming Jack Jiang. Towards a consistent interpretation of aiops models. 2021.

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

Christoph Molnar. *"Limitations of Interpretable Machine Learning*, chapter 14. 2020.

Umang Bhatt, Adrian Weller, and José MF Moura. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

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

Jirayus Jiarpakdee, Chakkrit Tantithamthavorn, Hoa Khanh Dam, and John Grundy. An empirical study of model-agnostic techniques for defect prediction models. *IEEE Transactions on Software Engineering*, 2020.

Lloyd S Shapley. *A value for n-person games*. Princeton University Press, 2016.

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.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

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.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020a.

Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 17–38. Springer, 2020.

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.

Dylan Slack, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. How much should i trust you? modeling uncertainty of black box explanations. *arXiv preprint arXiv:2008.05030*, 2020b.

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

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

Tiago Botari, Frederik Hvilshøj, Rafael Izbicki, and Andre CPLF de Carvalho. Melime: meaningful local explanation for machine learning models. *arXiv preprint arXiv:2009.05818*, 2020.

Supplementary Material

A PROOFS

A.1 Proposition 3.1

Proposition A.1 (Proposition 3.1)**.** *Suppose that $j <_{\boldsymbol{\phi}(h)} i$ and we have estimates $\widehat{\phi}_i(h), \widehat{\phi}_j(h)$ of the true attributions $\phi_i(h), \phi_j(h)$ with CIs I_i, I_j each of significance δ, then $\mathbb{P}\big(i \prec_{\widehat{\boldsymbol{\phi}}(h)} j\big) \leq 2\delta$.*

Proof. Given that $j <_{\boldsymbol{\phi}(h)} i$, the event $i \prec_{\widehat{\boldsymbol{\phi}}(h)} j$, or equivalently $\max|I_i| < \min|I_j|$, is included in the event $|\phi_i(h)| \notin |I_i|$ or $|\phi_j(h)| \notin |I_j|$. From the union bound, the latter event occurs with probability at most 2δ. □

Since the true order $\leq_{\boldsymbol{\phi}(h)}$ is total, for each of the $d(d-1)/2$ unordered pair of features $\{i, j\}$, either $i <_{\boldsymbol{\phi}(h)} j$ or $j <_{\boldsymbol{\phi}(h)} i$ (we assume that $\phi_i(h) = \phi_j(h)$ is impossible considering the attributions are real numbers). Hence, using the union bound, the probability that there exists some unordered pair $\{i, j\}$ such that $j \prec_{\hat{\boldsymbol{\phi}}(h)} i$ while $i <_{\boldsymbol{\phi}(h)} j$, or $i \prec_{\hat{\boldsymbol{\phi}}(h)} j$ while $j <_{\boldsymbol{\phi}(h)} i$ is at most $d(d-1)\delta$.

A.2 Proposition 3.2

Before proving the proposition, we wish to specify what are disagreements between orders. For two distinct features i and j and a partial order $\preceq$, there are exactly three mutually exclusive situations about relative feature importance: either $i \prec j$, $j \prec i$, or $i \perp j$.

We say that two partial orders $\preceq_1$ and $\preceq_2$ disagree on an unordered pair of features $\{i, j\}$ if they fail to place $\{i, j\}$ in the same configuration (among the three possible ones stated above). Thereof, six type of disagreements can occur between arbitrary orders.

Proposition A.2 (Proposition 3.2)**.** *Given two ensembles E and E' such that $E \subseteq E'$, the only possible disagreements between $\preceq_{\widehat{\boldsymbol{\phi}}(E)}$ and $\preceq_{\widehat{\boldsymbol{\phi}}(E')}$ are relations $i \prec_{\widehat{\boldsymbol{\phi}}(E)} j$ for which $i \perp_{\widehat{\boldsymbol{\phi}}(E')} j$.*

Proof. It suffices to prove that four out of six possible disagreements between orders are impossible.

Now let us show that $i \prec_{\widehat{\boldsymbol{\phi}}(E)} j$ and $j \prec_{\widehat{\boldsymbol{\phi}}(E')} i$ is impossible. Indeed, suppose that $j \prec_{\widehat{\boldsymbol{\phi}}(E')} i$ and so by the above remark $j \prec_{\widehat{\boldsymbol{\phi}}(E)} i$. This means that $j \preceq_{\widehat{\boldsymbol{\phi}}(E)} i$ and $j \neq i$. Now if $i \prec_{\widehat{\boldsymbol{\phi}}(E)} j$ also holds, i.e. $i \preceq_{\widehat{\boldsymbol{\phi}}(E)} j$ and $i \neq j$, we get by antisymmetry that $i = j$, a contradiction. Similarly, $j \prec_{\widehat{\boldsymbol{\phi}}(E)} i$ and $i \prec_{\widehat{\boldsymbol{\phi}}(E')} j$ is impossible.

Next, we show that $i \perp_{\widehat{\boldsymbol{\phi}}(E)} j$ and $j \prec_{\widehat{\boldsymbol{\phi}}(E')} i$ is also impossible. Indeed, if $j \prec_{\widehat{\boldsymbol{\phi}}(E')} i$, then $j \prec_{\widehat{\boldsymbol{\phi}}(E)} i$ by our remark. But then $i \perp_{\widehat{\boldsymbol{\phi}}(E)} j$, which stands for $i \not\preceq_{\widehat{\boldsymbol{\phi}}(E)} j$ and $j \not\preceq_{\widehat{\boldsymbol{\phi}}(E)} i$ does not hold. Finally, $i \perp_{\widehat{\boldsymbol{\phi}}(E)} j$ and $i \prec_{\widehat{\boldsymbol{\phi}}(E')} j$ is also impossible for similar reasons.

□

A.3 Proposition 3.3

Proposition A.3 (Proposition 3.3)**.** *Given an ensemble $E = \{h^{(k)}\}_{k=1}^{M}$ and estimates $\widehat{\phi}_i(h^{(k)}), \widehat{\phi}_j(h^{(k)})$ of the true attributions $\phi_i(h^{(k)}), \phi_j(h^{(k)})$ with CIs $I_i^{(k)}, I_j^{(k)}$ each of significance δ, if $i \not\prec_{\boldsymbol{\phi}(E)} j$, then $\mathbb{P}\big(i \prec_{\widehat{\boldsymbol{\phi}}(E)} j\big) \leq 2\delta$.*

Proof. Since $i \not\prec_{\boldsymbol{\phi}(E)} j$, there exists at least one model (lets call it $h^{(m)}$) for which $|\phi_i(h^{(m)})| \geq |\phi_j(h^{(m)})|$. Now note that the event $i \prec_{\widehat{\boldsymbol{\phi}}(E)} j$ is included in the event $i \prec_{\widehat{\boldsymbol{\phi}}(h^{(m)})} j$ whose probability is bounded by 2δ (from Proposition 3.1). □

A.4 Proposition 3.4

Proposition A.4 (Proposition 3.4)**.** *Given an ensemble $E = \{h^{(k)}\}_{k=1}^{M}$ and estimate attributions $\widehat{\phi}_i(h^{(k)}), \widehat{\phi}_j(h^{(k)})$ with CIs $I_i^{(k)}, I_j^{(k)}$, if the estimate attributions $\widehat{\phi}_i(h^{(k)})$ have the same sign for all models (and similarily for feature j), then $i \prec_{\widehat{\boldsymbol{\phi}}(E)} j \implies i <_{\widehat{\boldsymbol{\phi}}(h_E)} j$.*

Proof. Because $\widehat{\phi}_i(h^{(k)})$ has the same sign for all models (and similarly for feature j), the absolute value function commutes with the arithmetic mean. Therefore assuming $i \prec_{\widehat{\boldsymbol{\phi}}(E)} j$, we have

$$
\begin{aligned}
|\widehat{\phi}_i(h_E)| = \left|\frac{1}{M}\sum_{i=1}^{M}\widehat{\phi}_i(h^{(k)})\right| &= \frac{1}{M}\sum_{i=1}^{M}|\widehat{\phi}_i(h^{(k)})| \\
&\leq \frac{1}{M}\sum_{k=1}^{M}\max|I_i^{(k)}| && \text{since } \widehat{\phi}_i(h^{(k)}) \in I_i^{(k)} \quad \forall k \\
&< \frac{1}{M}\sum_{k=1}^{M}\min|I_j^{(k)}| && \text{because } i \prec_{\widehat{\boldsymbol{\phi}}(E)} j \\
&\leq \frac{1}{M}\sum_{i=1}^{M}|\widehat{\phi}_j(h^{(k)})| && \text{since } \widehat{\phi}_j(h^{(k)}) \in I_j^{(k)} \quad \forall k \\
&= \left|\frac{1}{M}\sum_{i=1}^{M}\widehat{\phi}_j(h^{(k)})\right| = |\widehat{\phi}_j(h_E)|,
\end{aligned}
$$

and so $i <_{\widehat{\boldsymbol{\phi}}(h_E)} j$ as desired. □

B MONTE-CARLO

B.1 Introduction

Monte-Carlo is fundamentally about approximating expectations of random variables. Letting $\boldsymbol{\psi} \in \mathbb{R}^n$ be a random vector, and $(\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \dots, \boldsymbol{\psi}^{(N)})$ be a random sequence of N iid observations of $\boldsymbol{\psi}$, we wish to approximate the expectation $\mu = \mathbb{E}[f(\boldsymbol{\psi})]$ for some function $f : \mathbb{R}^n \to \mathbb{R}$ given our N iid observations. The frequentist interpretation of probabilities would suggest that the following

$$\widehat{\mu}_N := \frac{1}{N} \sum_{\ell=1}^{N} f(\boldsymbol{\psi}^{(\ell)}) \tag{10}$$

is a good estimator of μ, the quality of which is assessed using the following theorem.

Theorem B.1 (Central Limit Theorem (Owen, 2013))**.** *Let $(\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \dots, \boldsymbol{\psi}^{(N)})$ be a sequence of N iid observations of $\boldsymbol{\psi}$, moreover let $\mu = \mathbb{E}[f(\boldsymbol{\psi})]$ and $\sigma^2 = \mathbb{E}[\,(f(\boldsymbol{\psi}) - \mu)^2\,]$ be finite, then the following holds for any $\delta \in [0, 1]$:*

$$\lim_{N \to \infty} \mathbb{P}\left(|\widehat{\mu}_N - \mu| \geq \Phi^{-1}(\delta/2) \frac{s_N}{\sqrt{N}} \right) = \delta,$$

where $\Phi^{-1} : [0, 1] \to \mathbb{R}$ is the normal quantile function and $s_N^2 = 1/(N-1) \sum_{\ell=1}^{N} (f(\boldsymbol{\psi}^{(\ell)}) - \widehat{\mu}_N)^2$ is the sample variance.

Roughly speaking, the theorem states that one can build *approximate* Confidence Intervals (CIs) around the estimates $\widehat{\mu}_N$ i.e. $I := [\widehat{\mu}_N - \Delta, \widehat{\mu}_N + \Delta]$ with $\Delta = \Phi^{-1}(\delta/2)\, s_N/\sqrt{N}$. As N increases, CIs are guaranteed to eventually capture the true expectation with probability $1 - \delta$.

We now discuss how to apply MC to estimate the Integrated Gradient and Shapley feature attributions. To improve the clarity of the analysis, we shall make the dependence of the attribution on the input $\boldsymbol{x}$ of interest and the background distribution $\mathcal{B}$ explicit by writting $\boldsymbol{\phi}(h, \boldsymbol{x}, \mathcal{B})$, unlike in the main paper.

B.2 Integrated Gradient

As a reminder the IG attribution of feature i is the following

$$\phi_i^{\mathrm{IG}}(h, \boldsymbol{x}, \mathcal{B}) := \mathop{\mathbb{E}}_{\substack{\boldsymbol{z} \sim \mathcal{B}, \\ \alpha \sim U(0,1)}} \left[(x_i - z_i) \frac{\partial h}{\partial x_i}\bigg|_{\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{z}} \right]. \tag{11}$$

By defining the random vector $\boldsymbol{\psi} = (\boldsymbol{z}, \alpha) \sim \mathcal{B} \times U(0, 1)$, we can rewrite IG in the general form

$$\phi_i^{\mathrm{IG}}(h, \boldsymbol{x}, \mathcal{B}) = \mathbb{E}\left[f(\boldsymbol{\psi}) \right]. \tag{12}$$

whose MC estimate is immediately apparent

$$\widehat{\phi}_i^{\mathrm{IG}}(h, \boldsymbol{x}, \mathcal{B}) = \frac{1}{N} \sum_{\ell=1}^{N} f(\boldsymbol{\psi}^{(\ell)}) = \frac{1}{N} \sum_{\ell=1}^{N} (x_i - z_i^{(\ell)}) \frac{\partial h}{\partial x_i}\bigg|_{\alpha^{(\ell)} \boldsymbol{x} + (1-\alpha^{(\ell)}) \boldsymbol{z}^{(\ell)}} \tag{13}$$

B.3 Shapley Values

We once again show the definition of the Shapley attributions

$$\phi_i^{\mathrm{Shap}}(h, \boldsymbol{x}, \mathcal{B}) = \mathop{\mathbb{E}}_{\substack{\pi \sim \Omega \\ \boldsymbol{z} \sim \mathcal{B}}} \left[h(\, \boldsymbol{r}_{\pi_{:i} \cup \{i\}}(\boldsymbol{z}, \boldsymbol{x})\,) - h(\, \boldsymbol{r}_{\pi_{:i}}(\boldsymbol{z}, \boldsymbol{x})\,) \right]. \tag{14}$$

By letting $\boldsymbol{\psi} = (\boldsymbol{z}, \pi) \sim \mathcal{B} \times \Omega$ be a random vector, sampling it N times iid, and taking the empirical average of the expression inside the brackets of Equation 14, we recover the IME algorithm (Štrumbelj and Kononenko, 2014).

However, this approach has a significant drawback: it evaluates the model h on redundant inputs. Indeed, several samples of $\boldsymbol{r}_{\pi_{:i}}(\boldsymbol{z}, \boldsymbol{x})$ can yield the exact same point in $\mathbb{R}^d$, because some background samples may have similar feature values as $\boldsymbol{x}$. Considering the cost of evaluating large ensembles at any given input, it is preferable to ignore duplicate inputs.

To this end, the SHAP library provides efficient estimates of Equation 14 via the `PermutationExplainer`. This algorithm is optimized for speed and therefore the SHAP source code discourages feeding the whole background distribution $\mathcal{B}$ to the explainer, but rather use a subset of 100 iid samples from $\mathcal{B}$, which we call a background minibatch.

Before introducing the algorithm in detail, it is necessary to clarify and introduce notation. As stated previously, π is a permutation over $[d]$, $\pi(i)$ represents the position of feature i in π, and $\pi^{-1}(m)$ is the feature at position m. Moreover $\pi_{:i} = \{j \in [d] : \pi(j) < \pi(i)\}$ and $\pi_{i:} = \{j \in [d] : \pi(i) < \pi(j)\}$ are the subsets of features that appear before and after i in π.

Algorithm 1 presents the pseudo-code of the procedures within the `PermutationExplainer` (vectors start at index 1).

Algorithm 1 `PermutationExplainer` given a budget of p permutation samples.

Input: $h, \boldsymbol{x}$
Parameters: Background minibatch $(\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(100)}) \sim \mathcal{B}^{100}$ and Permutations $(\pi^{(1)}, \ldots, \pi^{(p)}) \sim \Omega^p$
Output: Estimator of $\boldsymbol{\phi}^{\text{Shap}}(h, \boldsymbol{x}, \mathcal{B})$

Initialize $\widehat{\boldsymbol{\phi}} = \mathbf{0} \in \mathbb{R}^d$;
Store averages $\boldsymbol{c} = \mathbf{0} \in \mathbb{R}^{2d+1}$;
Set background $c_1 = c_{2d+1} = \frac{1}{100} \sum_{\ell=1}^{100} h(\boldsymbol{z}^{(\ell)})$;
Set prediction $c_{d+1} = h(\boldsymbol{x})$;
for $j = 1, 2, \ldots, p$ **do**
  Let $\pi = \pi^{(j)}$;
  for $m = 2, \ldots, d$ **do**
    Let $i = \pi^{-1}(m)$
    $c_m = \texttt{Efficiently\_evaluate}\big[\frac{1}{100} \sum_{\ell=1}^{100} h(\, \boldsymbol{r}_{\pi_{:i}}(\boldsymbol{z}^{(\ell)}, \boldsymbol{x})\,)\big]$;
  end for
  for $m = 1, \ldots, d-1$ **do**
    Let $i = \pi^{-1}(m)$;
    $c_{m+d+1} = \texttt{Efficiently\_evaluate}\big[\frac{1}{100} \sum_{\ell=1}^{100} h(\, \boldsymbol{r}_{\pi_{i:}}(\boldsymbol{z}^{(\ell)}, \boldsymbol{x})\,)\big]$;
  end for
  for $m = 1, \ldots, d$ **do**
    $i = \pi^{-1}(m)$;
    $\widehat{\phi}_i \; \mathrel{+}= c_{m+1} - c_m + c_{m+d} - c_{m+d+1}$;
  end for
end for
return $\widehat{\boldsymbol{\phi}}/(2p)$

Note that the complex optimizations made by SHAP are hidden in the repeated calls of the function `Efficiently_evaluate`, which is able to reuse model evaluations from its previous call.

The efficiency of this estimator comes at a cost, the loss of the iid assumption necessary to the application of the Central Limit Theorem. Indeed, once again thinking of $\boldsymbol{\psi} = (\boldsymbol{z}, \pi)$ as a random vector, we observe that Algorithm 1 does not provide iid samples of $\boldsymbol{\psi}$, but rather a $100 \times 2p$ grid. This is because, the algorithm considers all pairings between permutations and background minibatch samples.

Our conservative solution to allow the application of the Central Limit Theorem is a direct consequence of the following observation: considering the random vector $\boldsymbol{\psi} = (\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(100)}, \pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(p)}) \sim \mathcal{B}^{100} \times \Omega^p$ being fed to the function $f(\boldsymbol{\psi}) = \texttt{PermutationExplainer}(h, \boldsymbol{x}; \boldsymbol{\psi})_i$, the following holds

$$\phi_i^{\text{Shap}}(h, \boldsymbol{x}, \mathcal{B}) = \mathbb{E}\left[f(\boldsymbol{\psi})\right] = \mathbb{E}\left[\texttt{PermutationExplainer}(h, \boldsymbol{x}; \boldsymbol{\psi})_i\right], \tag{15}$$

which states that $\texttt{PermutationExplainer}(h, \boldsymbol{x}; \boldsymbol{\psi})$, provided by SHAP, is an unbiased estimator.

Proof. We let

$$S_i(h, \boldsymbol{x}; \boldsymbol{z}, \pi) = h(\, \boldsymbol{r}_{\pi_{:i}}(\boldsymbol{z}, \boldsymbol{x})\,) - h(\, \boldsymbol{r}_{\pi_{:i}\cup\{i\}}(\boldsymbol{z}, \boldsymbol{x})\,), \qquad \bar{S}_i(h, \boldsymbol{x}; \boldsymbol{z}, \pi) = h(\, \boldsymbol{r}_{\pi_{i:}\cup\{i\}}(\boldsymbol{z}, \boldsymbol{x})\,) - h(\, \boldsymbol{r}_{\pi_{i:}}(\boldsymbol{z}, \boldsymbol{x})\,).$$

We observe that Algorithm 1 boils down to computing

$$\frac{1}{200p}\sum_{j=1}^{p}\sum_{\ell=1}^{100}\left[S_i(h, \boldsymbol{x}; \boldsymbol{z}^{(\ell)}, \pi^{(j)}) + \bar{S}_i(h, \boldsymbol{x}; \boldsymbol{z}^{(\ell)}, \pi^{(j)})\right]. \tag{16}$$

Taking the expectation of this expression w.r.t the random vector $\boldsymbol{\psi} = (\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(100)}, \pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(p)}) \sim \mathcal{B}^{100} \times \Omega^p$, we obtain

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\psi}}\left[\texttt{PermutationExplainer}(h, \boldsymbol{x}; \boldsymbol{\psi})_i\right] &= \frac{1}{200p}\sum_{j=1}^{p}\sum_{\ell=1}^{100}\mathop{\mathbb{E}}_{\substack{\pi^{(j)}\sim\Omega\\ \boldsymbol{z}^{(\ell)}\sim\mathcal{B}}}\left[S_i(h, \boldsymbol{x}; \boldsymbol{z}^{(\ell)}, \pi^{(j)}) + \bar{S}_i(h, \boldsymbol{x}; \boldsymbol{z}^{(\ell)}, \pi^{(j)})\right] \\
&= \frac{1}{100p}\sum_{j=1}^{p}\sum_{\ell=1}^{100}\mathop{\mathbb{E}}_{\substack{\pi^{(j)}\sim\Omega\\ \boldsymbol{z}^{(\ell)}\sim\mathcal{B}}} S_i(h, \boldsymbol{x}; \boldsymbol{z}^{(\ell)}, \pi^{(j)}) \\
&= \frac{1}{100p}\sum_{j=1}^{p}\sum_{\ell=1}^{100}\phi_i^{\text{Shap}}(h, \boldsymbol{x}, \mathcal{B}) \\
&= \phi_i^{\text{Shap}}(h, \boldsymbol{x}, \mathcal{B})
\end{aligned} \tag{17}$$

where to go from the first line to the second we used the following property

$$\mathop{\mathbb{E}}_{\substack{\pi^{(j)}\sim\Omega\\ \boldsymbol{z}^{(\ell)}\sim\mathcal{B}}}\left[S_i(h, \boldsymbol{x}; \boldsymbol{z}^{(\ell)}, \pi^{(j)})\right] = \mathop{\mathbb{E}}_{\substack{\pi^{(j)}\sim\Omega\\ \boldsymbol{z}^{(\ell)}\sim\mathcal{B}}}\left[\bar{S}_i(h, \boldsymbol{x}; \boldsymbol{z}^{(\ell)}, \pi^{(j)})\right] \tag{18}$$

based on the symmetry of the distribution Ω over permutations. □

Since `PermutationExplainer` is an unbiased estimator, running it N times with iid permutations and background minibatches, and averaging results is a MC estimate of Shapley values

$$\widehat{\phi}_i^{\text{Shap}}(h, \boldsymbol{x}, \mathcal{B}) = \frac{1}{N}\sum_{\ell=1}^{N} f(\boldsymbol{\psi}^{(\ell)}) = \frac{1}{N}\sum_{\ell=1}^{N}\texttt{PermutationExplainer}(h, \boldsymbol{x}; \boldsymbol{\psi}^{(\ell)}). \tag{19}$$

C DISTRIBUTIONAL SHIFT IN EXPLANABILITY

This section presents a preliminary investigation of why independently trained models can yield diverse (and even contradicting) feature attributions/importance. We provide empirical evidence of a distributional shift between the train/test domains and the explanation domain.

We denote probability distributions as $\mathcal{D}$ and their empirical counterparts by $\widehat{\mathcal{D}} := \frac{1}{N}\sum_{\ell=1}^{N} \delta_{z^{(\ell)}}$ which attribute uniform weights to each instance of a finite set $\{z^{(1)}, \ldots, z^{(N)}\}$ sampled iid from $\mathcal{D}$. Moreover, letting f be a (measurable) function, we define the push-forward distribution $f \circ \mathcal{D}$ of the distribution $\mathcal{D}$ mapped though$f$ i.e.

$$z \sim f \circ \mathcal{D} \iff z = f(w) \quad \text{with} \quad w \sim \mathcal{D}. \tag{20}$$

C.1 Investigating Explanation Distributions

All models in the ensemble are minimizing the empirical risk on the training data and the selection of hyperparameters and models is guided by the same performance metric. While this indirectly ensures that all models will give similar predictions on the train/test instances, they can disagree on other inputs. As Monte-Carlo (MC) estimates of the SHAP and IG attributions rely on the evaluation of the models on points which may not belong to the train/test datasets, we hypothesize that our models disagree more on MC points leading to variability in feature attribution. It can be argued the underspecification of our models (materialized by a stronger disagreement on MC points) translates to the underspecification of the feature importance, which our partial order based consensus approach aims to address.

To investigate this hypothesis, we now formalize the distributions of input generated during MC estimations of IG and SHAP, which we refer to as *Explanation Distributions*. We then study the distributional shift they represent from the empirical data distribution $\widehat{\mathcal{D}}^{\text{Data}}$, and more specifically, how the shift affects the models confidence as measured by the pointwise predictive variance

$$\Delta(\boldsymbol{x}) := \frac{1}{M}\sum_{k=1}^{M} \left(h^{(k)}(\boldsymbol{x}) - h_E(\boldsymbol{x})\right)^2. \tag{21}$$

C.1.1 SHAP Distribution

When running Algorithm 1 to estimate Shapley values, the models will be evaluated on inputs $\boldsymbol{r}_{\pi,i}(\boldsymbol{z}, \boldsymbol{x})$ sampled by replacing subsets $\pi_{:i}$ of features of $\boldsymbol{x}$ with features of a background instance $\boldsymbol{z} \sim \mathcal{B}$. Unless features are statistically independent (which they rarely are) the action of replacing certain features will generate unrealistic inputs, e.g. a very humid and hot day of January, or a house that was renovated before it was build. This observation has already been made and exploited to attack SHAP explanations by modifying the model on those unrealistic inputs, resulting in arbitrary explanations without affecting the test performance (Slack et al., 2020a). However, connections with distributional shift in ML remains to be fully explored.

By assuming that only training instances as are ever used in background samples while test set points are systematically being explained, we can formalize the distribution $\mathcal{D}^{\text{Shap}}$ of all points that will ever be generated when computing SHAP attributions in practice, see Algorithm 2.

Algorithm 2 SHAP distribution: $\mathcal{D}^{\text{Shap}}$

Parameters: Training Distr. $\widehat{\mathcal{D}}^{\text{Train}}$ and Test Distr. $\widehat{\mathcal{D}}^{\text{Test}}$
Output: A single sample from $\mathcal{D}^{\text{Shap}}$

1: Sample feature $i \sim U([d])$;
2: Sample permutation $\pi \sim \Omega$;
3: Sample input to explain $\boldsymbol{x} \sim \widehat{\mathcal{D}}^{\text{Test}}$;
4: Sample background instance $\boldsymbol{z} \sim \widehat{\mathcal{D}}^{\text{Train}}$;
5: **return** $\boldsymbol{r}_{\pi:i}(\boldsymbol{z}, \boldsymbol{x})$

C.1.2 IG Distribution

When estimating IG with MC, the models will be fed random convex combinations of the input to explain $\boldsymbol{x}$ and background samples $\boldsymbol{z} \sim \mathcal{B}$. Formally, all points that will ever be generated when using IG in a practical settings, assuming only train points are ever used for background and only test points are explained, are sampled from the distribution $\mathcal{D}^{\text{IG}}$ described in Algorithm 3.

Algorithm 3 IG distribution: $\mathcal{D}^{\text{IG}}$

Parameters: Training Distr. $\widehat{\mathcal{D}}^{\text{Train}}$ and Test Distr. $\widehat{\mathcal{D}}^{\text{Test}}$
Output: A single sample from $\mathcal{D}^{\text{IG}}$

1: Sample constant $\alpha \sim U([0, 1])$;
2: Sample input to explain $\boldsymbol{x} \sim \widehat{\mathcal{D}}^{\text{Test}}$;
3: Sample background instance $\boldsymbol{z} \sim \widehat{\mathcal{D}}^{\text{Train}}$;
4: **return** $\alpha\boldsymbol{x} + (1 - \alpha)\boldsymbol{z}$

C.2 Toy Example

This toy regression dataset previously used by Botari et al. (2020) generates input points along a 2D spiral and labels then with the cumulative length of the spiral up to that point. Due to the manifold structure of the dataset, it is easy to visualize which inputs are close or far from the data, an information that is supported by the variance (disagreement) among model predictions according to Figure 7. We indeed observe that the pointwise variance between models is smaller on inputs close to the manifold compared to elsewhere by orders of magnitude.

Nonetheless, since the explanation distributions differ from the data-generating one, some inputs generated by SHAP and IG fall outside the manifold, and even into regions where model disagreements are ~ 100 times larger than on the data points.

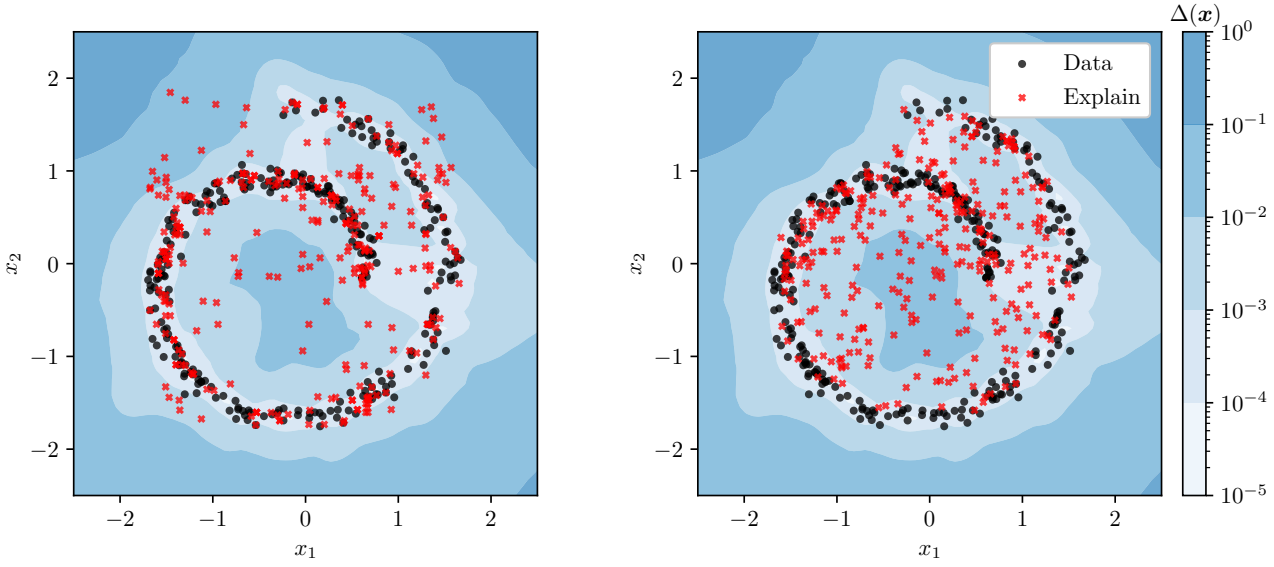


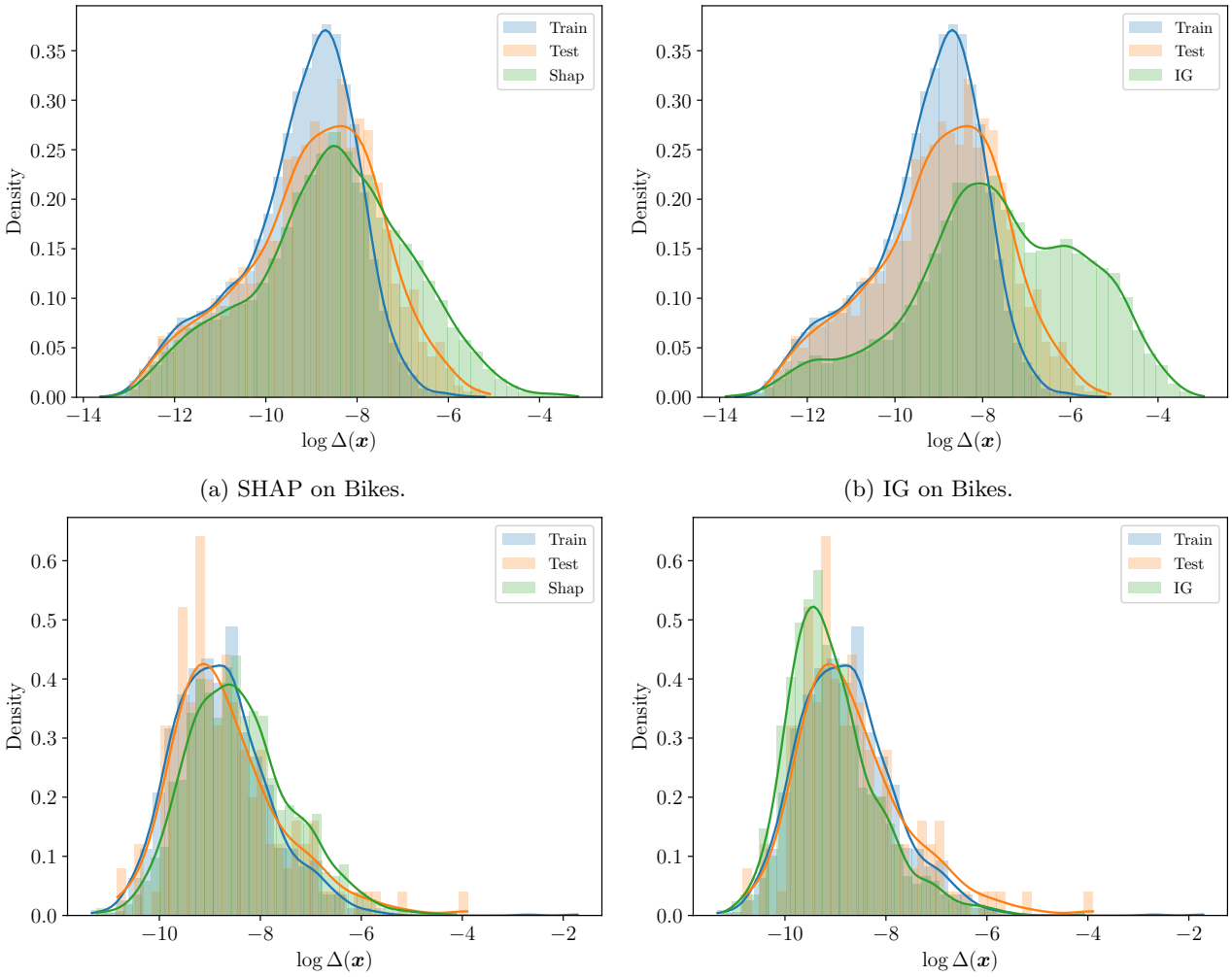
Figure 7: Scatter plot of training data and samples from the Explainer Distributions $\mathcal{D}^{\text{Shap}}$ (left) and $\mathcal{D}^{\text{IG}}$ (right) for the Spiral toy dataset. We see that the 10 models disagree more, as measured by $\Delta(\boldsymbol{x})$, on some samples from the Explainer Distributions than on most training points.

.

C.3 Real Datasets

When studying higher-dimensional datasets, we cannot as easily visualise the distributional shift. An alternative is visualize histograms of the empirical distributions $\log \circ \Delta \circ \widehat{\mathcal{D}}$ with $\widehat{\mathcal{D}} = \widehat{\mathcal{D}}^{\text{Train}}$, $\widehat{\mathcal{D}}^{\text{Test}}$, and $\widehat{\mathcal{D}}^{\text{Explain}}$, see Figure

8. In total, 20 models were used to compute the pointwise variance and the empirical explanation distributions $\widehat{\mathcal{D}}^{\text{Explain}}$ were constructed by running Algorithms 2 and 3 as many times as there are instances in the datasets.



(a) SHAP on Bikes.

(b) IG on Bikes.

(c) SHAP on Houses.

(d) IG on Houses.

Figure 8: Model disagreements on inputs sampled by the explanation distributions on two real datasets.

Looking at the top right plot, it is apparent that there is a considerable portion of the IG samples generated inputs on which the models disagree more than on train/test instances. The same behavior is also present on the top left plot (SHAP), although it is less pronounced. The difference between IG and SHAP could be explained by the fact that Bikes contain many features (hour, day of the week, month, and year) structured as a regular grid. Since linear interpolations between two vertices of a grid are unlikely to land on another vertex, models need not agree on IG points even though they are trained on the same data. On the other hand, replacing some feature values of a vertex with another one (as done in SHAP) necessarily yields a vertex of the grid. Focusing on the bottom row of Figure 8, the difference between the train/test/explain distributions are not as striking, although IG has a slight tendency to generate inputs on which the models agree more than on most train/test instances.

To refine our analysis, we computed the Mann-Whitney U-statistic as a principled mean to compare the explanation distributions and the dataset distribution $\widehat{\mathcal{D}}^{\text{Data}}$ (train and test). Formally, given two empirical distributions $\widehat{\mathcal{D}}_1$, and $\widehat{\mathcal{D}}_2$, the U statistic in common language effect size is defined by

$$\mathcal{U}(\widehat{\mathcal{D}}_1, \widehat{\mathcal{D}}_2) := \mathop{\mathbb{P}}_{\substack{z_1 \sim \widehat{\mathcal{D}}_1 \\ z_2 \sim \widehat{\mathcal{D}}_2}} [z_1 < z_2]. \tag{22}$$

It coincides with the area under the receiver operator characteristic curve (AUROC) and can be interpreted

	BIKES		HOUSES	
Explain. distr.	**U-stat**	**p-val**	**U-stat**	**p-val**
SHAP	0.631 ± 0.003	0	0.610 ± 0.006	$(2 \pm 4) \times 10^{-22}$
IG	0.756 ± 0.002	0	0.418 ± 0.009	1

Table 1: U-statistics $\mathcal{U}(\Delta \circ \widehat{\mathcal{D}}^{\text{Data}}, \Delta \circ \widehat{\mathcal{D}}^{\text{Explain}})$ comparing the distributions of pointwise variance $\Delta(\boldsymbol{x})$ on datasets and explanation distributions

as a dissimilarity measure between distributions: it takes values near 1/2 when distributions are very similar, and values close to 0 or 1 when one distribution is significantly shifted compared to the other. We employed the statistic in the one-sided Mann-Whitney test of SciPy[6], whose null hypothesis $H_0 : \mathcal{D}_1 = \mathcal{D}_2$ is rejected in favor of the alternative $H_1 : \mathcal{D}_1 < \mathcal{D}_2$ when the U statistic is high enough compared to 1/2. Table 1 presents the U-statistics (and their respective p-values) resulting from comparing the datasets and explanations distributions across 10 reruns to account for the randomness of Algorithms 2 and 3. The main observation is that three out of four null hypotheses were rejected with high significance (all p-values way smaller than 1%), meaning that in these scenarios we can clearly measure a shift of model confidence between the dataset and explanation distributions. Taking a closer look at the U-statistics, we note as previously that, on the Bikes dataset, IG tend to generate inputs on which the models have a higher disagreement compared to SHAP, as evidenced by its higher U-statistic (0.756 versus 0.631).

Interestingly, IG yields a U-statistic below 0.5 on Houses, suggesting that models tend to agree more on interpolations between train/test instances than on the dataset as a whole. Since we used a one-sided test, the associated p-value of 1 means that we failed to provide strong evidence that models trained on Houses have larger disagreements on IG points than on the whole dataset. However, despite this fact, we still ended up with a partial order when taking the consensus of IG feature importance in **Section 4.2**, meaning that Table 1 presents an incomplete view of the underspecification of feature attributions. In truth, while we consider here all the inputs that occur in explaining *any* test instance with the whole training set as background, it could happen that the Explainer Distribution corresponding to a specific instance (and possibly different background) shows a different behavior, possibly suggesting that models have diverse explanations for that specific instance. Indeed, this Table is principally meant as a preliminary mean to illustrate the general phenomenon of distributional shift occurring when applying a given explanation method to a specific dataset, and not as a definitive proof of why models have diverse explanations.

[6] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html`

D EXPERIMENTAL DETAILS

All experiments involving MLPs were performed using a GPU (NIVIDIA GeForce RTX 2080 Ti) resulting in around 50 hours of runtime total. To allow for the fast training and explanations of the 1K MLPs, we made sure to systematically store 50 models on the graphics card and evaluate them in parallel. Otherwise, we would not have been able to run SHAP `PermutationExplainer` 400 times for 1000 models in $\sim 1h$.

The online platform Weights & Biases[7] was used to help design experiments and to fine-tune hyperparameters. Notably it performed grid searches over number of epochs, learning rate, and learning rate scheduler[8], following Table 2. Given a specific set of hyperparameters Φ, the procedure shown in Algorithm 4 was employed with $m = 5$ and $M = 10$ to evaluate their fitness. The hyperparameters with highest fitness were then used to train $\sim$1K MLPs on the whole training set.

Table 2: Simple Grid Search

HYPERPARAM	RANGE
`n_epochs`	[50, 100, 150, 200, 250]
`learning_rate`	`logspace`(-4, -1, 7)
`use_lr_scheduler`	[`True`, `False`]

Algorithm 4 Fitness of the learning algorithm $\mathcal{A}_\Phi$, with hyperparameters Φ.

Input: Training set S, Hyperparameters Φ
Output: RMSE

Initialize RMSE $= 0$;
for rep $= 1, 2, \ldots, m$ **do**
  $T, V =$ `ShuffleSplit`(S)
  for $k = 1, 2, \ldots, M$ **do**
    $h^{(k)} \sim \mathcal{A}_\Phi(T)$;
  end for
  $h = \frac{1}{M} \sum_{k=1}^{M} h^{(k)}$
  RMSE $+= \sqrt{\frac{1}{|V|} \sum_{j \in V} \left(h(\boldsymbol{x}^{(j)}) - y^{(j)}\right)^2}$
end for
return RMSE$/m$

[7] `https://wandb.ai/site`

[8] `https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html#torch.optim.lr_scheduler.ReduceLROnPlateau`

E ADDITIONAL RESULTS

In this section, we explain additional instances for the Houses and Bikes datasets and present the results. Moreover, we conduct an additional experiment, which consists of computing the trustworthiness ratios R_C discussed in **Section 3.3** for ensembles E_M of M models sampled without replacement from E. This experiment is repeated 5,000 times for each sizes $M = \{2, 3, 4, 5, \ldots, 13, 14\}$.

E.1 Houses

E.1.1 SHAP on test Instance A

Section 4.2 explained the high price of a given house via the IG attribution. To avoid confusion, we shall refer to this instance as instance A. As a complement to the IG, we present the explanations obtained by running SHAP 200 times with 100 new background samples each time. The maximal relative attribution error was 1.09%. Results are presented in Figure 9. Note that the Hasse diagram is identical to what we had with IG, which is a surprising result considering SHAP and IG are completely different mathematical quantities. However, considering that SHAP and IG attributions coincide on additive models, we hypothesize that feature interactions are very weak in Houses compared to the main effects, an hypothesis that still needs to be empirically confirmed.

E.1.2 IG on Test Instance B

Similarly to instance A, we aim at explaining extreme values of house prices from the test set, but now look at low prices.

Formulate The instance B to explain was the test set house with the lowest selling price (79,000 USD predicted as 95,041 USD by h_E), and was compared to all training instances.

Approximate For MC approximation of IG, a total of 50,000 samples where used leading to relative attribution errors of at most 0.5%.

Explain Figure 10 shows the results on a pool of 1115 selected models.

E.2 Bike-Sharing

E.2.1 IG on Test Instance A

In **Section 4.3**, we used SHAP to explain a test instance from Bike-Sharing with very few rentals, which we refer to as instance A. To complement this study, we attempted to explain the same instance with IG but we were not able to obtain relative attribution errors below 2% in a reasonable time. For this reason, we abstain from trusting IG explanations of instance A and do not present results.

E.2.2 SHAP on Test Instance B

Formulate Explain the test set instance $\boldsymbol{x}$ with the most amount of bike rentals (487 bikes predicted as 443 by h_E) between 7-8h on workingdays in 2011. The background $\mathcal{B}$ is once again the whole training set is conditioned on `yr=2011`, `workingday=True` and $7 \leq$ `hr` ≤ 8.

Approximate We conducted MC by independently running the `PermutationExplainer` 200 times, with each call using a subset of 100 random instances from the background with the default number of permutations from SHAP. The highest relative attribution error reported was 1.875%.

Explain After accounting for performance and relative attribution errors, 990 models were considered for the consensus and the results are shown in Figure 11.

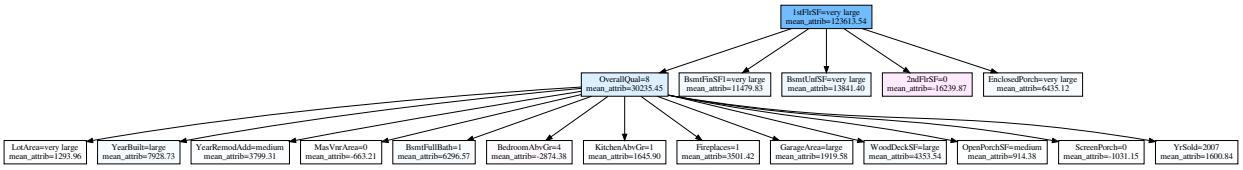
E.2.3 SHAP on Test Instance C

As we were exploring the dataset, it came to our attention that instances with the most bikes rented between 10 and 17h were on off-days, which is intuitive. However, there was still a lot of variance in bike rentals within this subset of the data and we attempt to explain one of the extreme values observed.

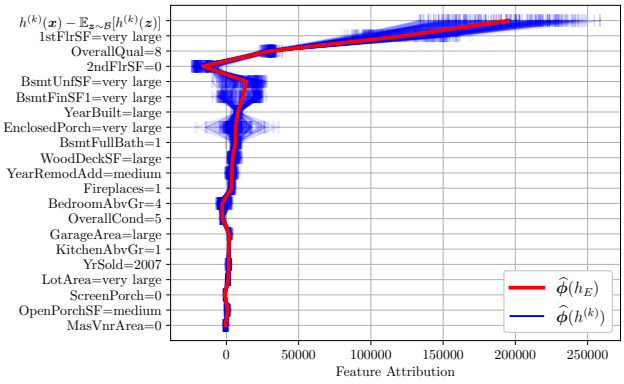
Formulate Explain the test set instance $\boldsymbol{x}$ with the most amount of bike rentals (715 bikes predicted as 722 by the aggregate model) between 10-17h on off-days in 2012. The background $\mathcal{B}$ is chosen as the whole training set conditioned on `yr=2012`, `workingday=False` and $10 \leq$ `hr` ≤ 17.

Approximate MC was done as previously resulting in a maximal relative attribution error of 0.36%.

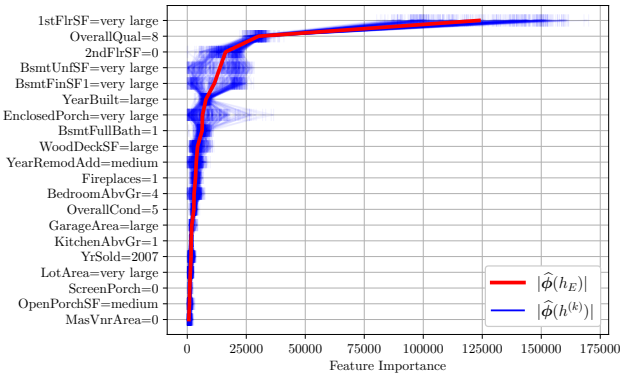
Explain A total of 990 were selected and their aggregated explanations are presented in Figure 12.



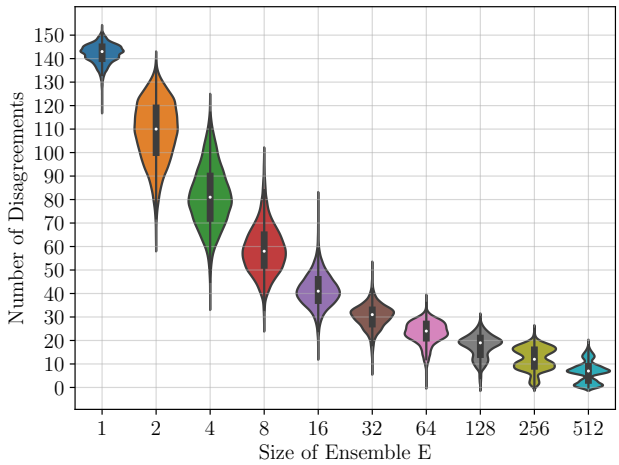
(a) Hasse diagram, trustworthiness ($R_M = 34/190, R_C = 34/34$).



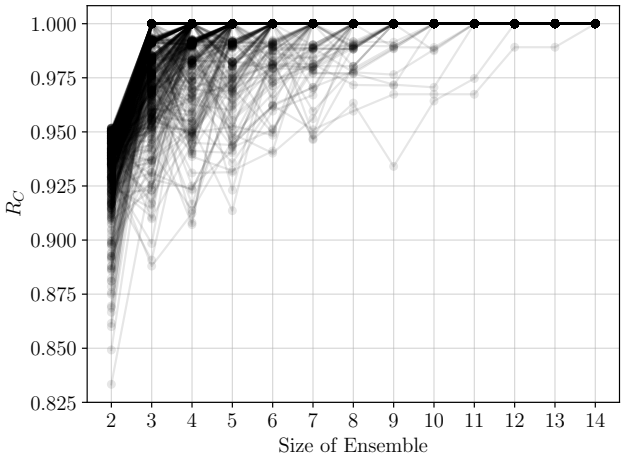
(b)



(c)

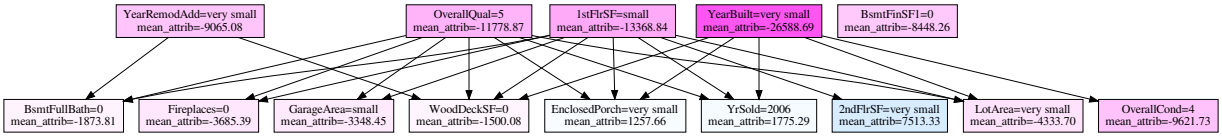


(d) Convergence of the partial order.

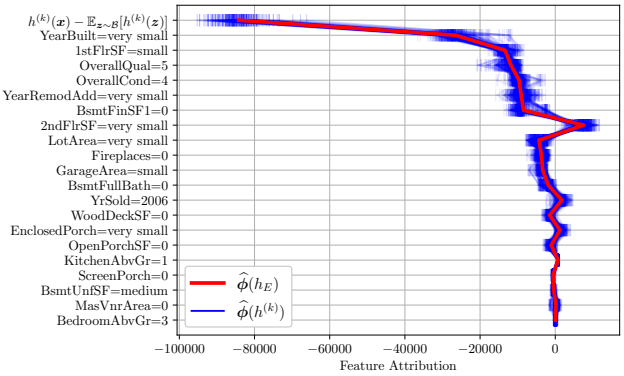


(e) Trustworthiness ratio of the partial order.

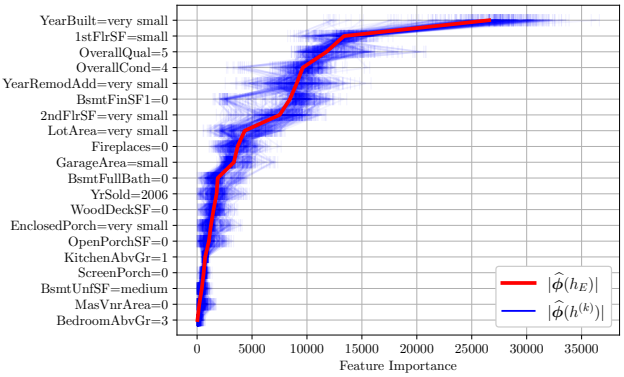
Figure 9: Results of explaining the test instance A of Houses with SHAP.



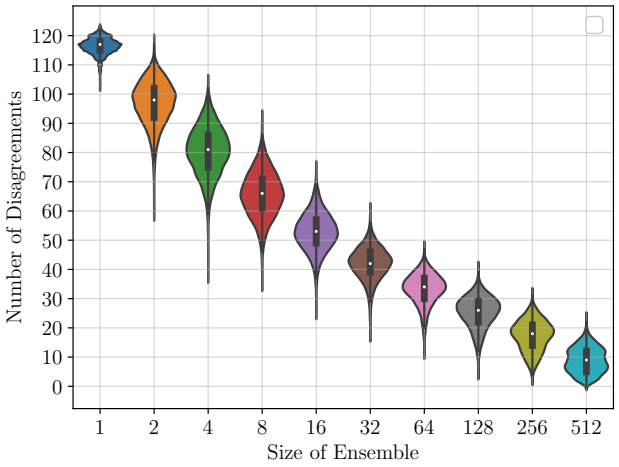
(a) Hasse diagram, trustworthiness ($R_M = 65/190, R_C = 65/65$).



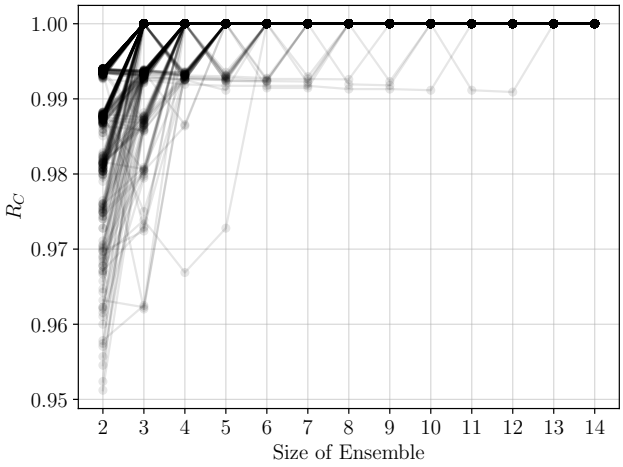
(b)



(c)

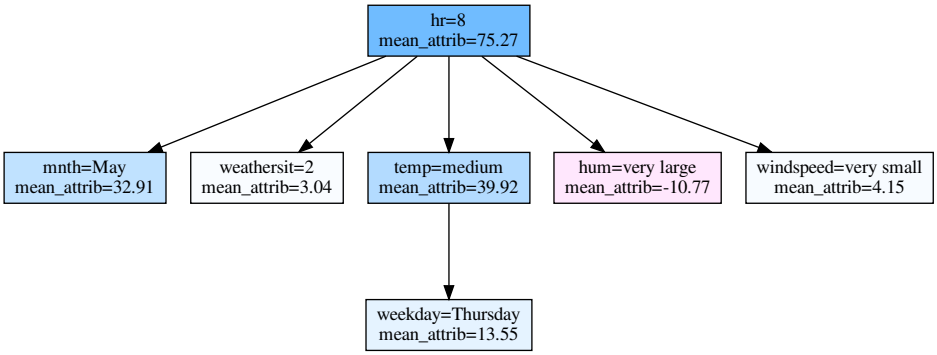


(d) Convergence of the partial order.

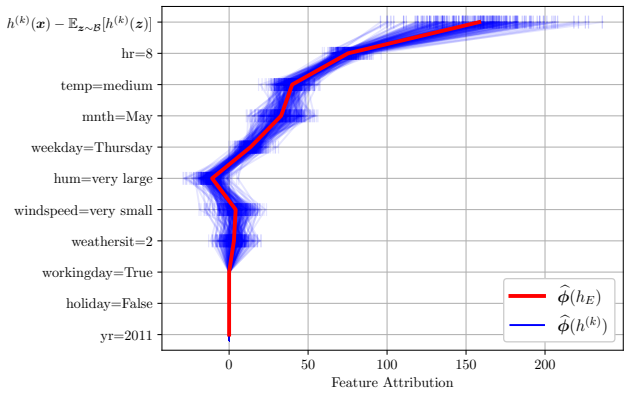


(e) Trustworthiness ratio of the partial order.

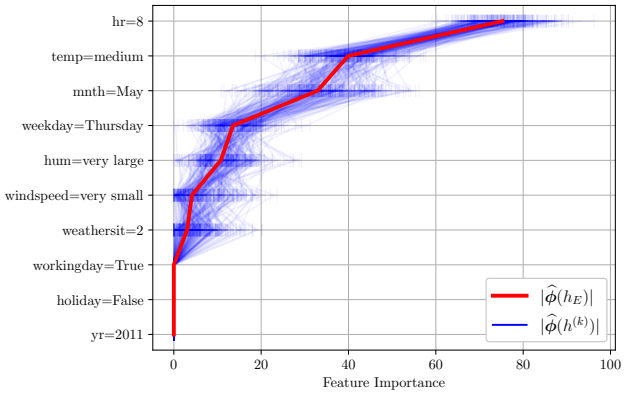
Figure 10: Results of explaining the instance B of Houses with IG.



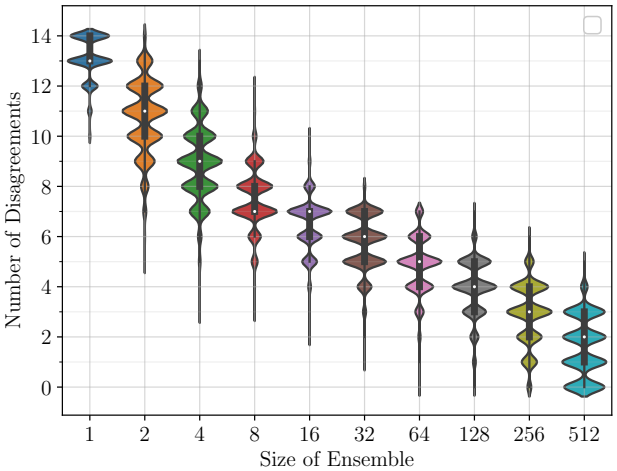
(a) Hasse diagram, trustworthiness ($R_M = 7/21, R_C = 7/7$).



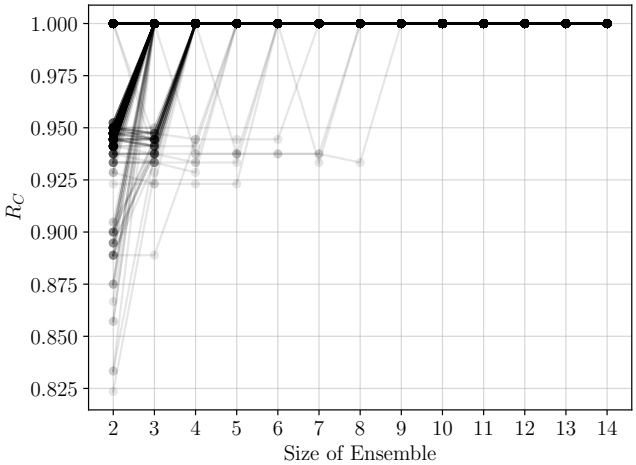
(b)



(c)

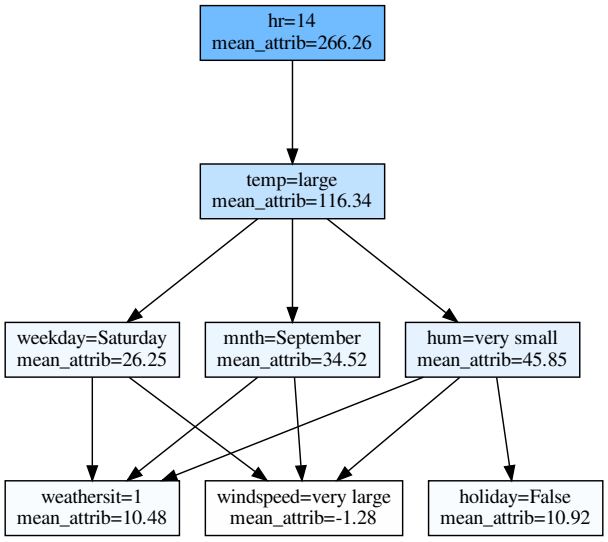


(d) Convergence of the partial order.

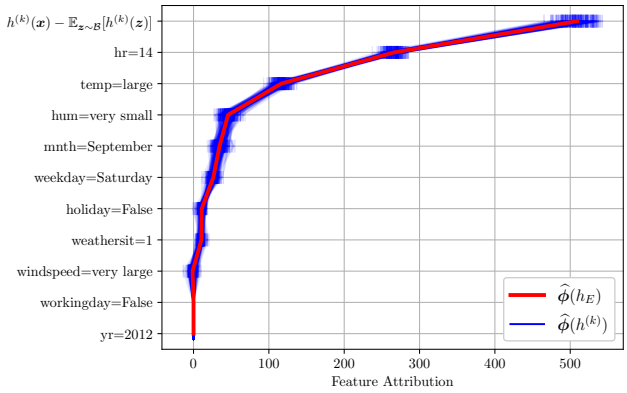


(e) Trustworthiness ratio of the partial order.

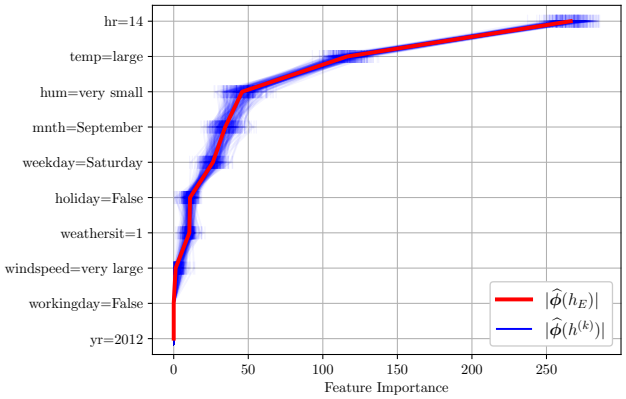
Figure 11: Results of explaining instance B from Bikes with SHAP.



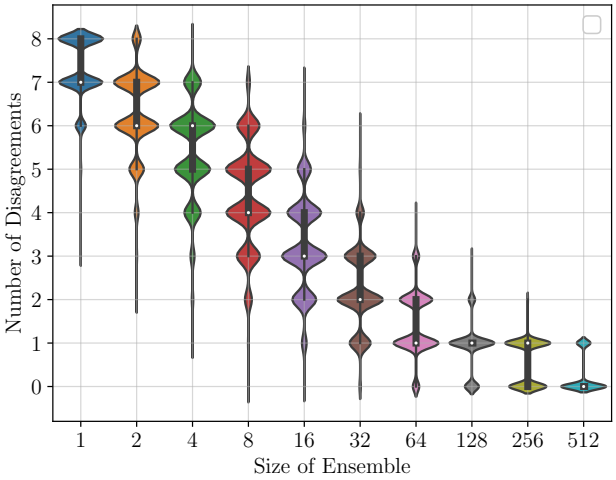
(a) Hasse diagram, trustworthiness ($R_M = 20/28, R_C = 20/20$).



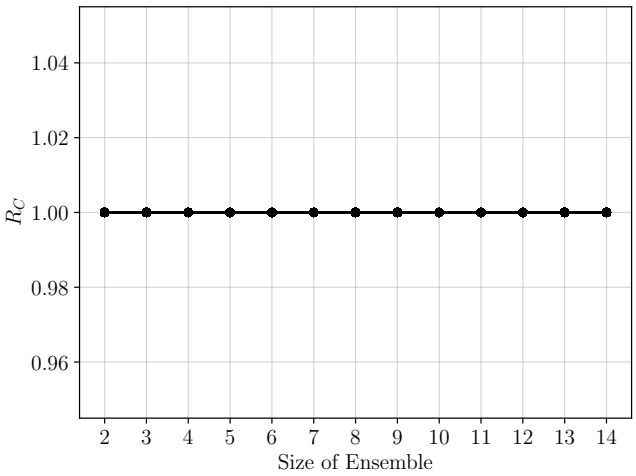
(b)



(c)



(d) Convergence of the partial order.



(e) Trustworthiness ratio of the partial order.

Figure 12: Results of explaining instance C from Bikes with SHAP.