Consistent Sufficient Explanations and Minimal Local Rules for explaining regression and classification models

Anonymous Author
Anonymous Institution

Abstract

To explain the decision of any model, we extend the notion of probabilistic Sufficient Explanations (P-SE). For each instance, this approach selects the minimal subset of features that is sufficient to yield the same prediction with high probability, while removing other features. The crux of P-SE is to compute the conditional probability of maintaining the same prediction. Therefore, we introduce an accurate and fast estimator of this probability via random Forests for any data $(\boldsymbol{X}, Y)$ and show its efficiency through a theoretical analysis of its consistency. As a consequence, we extend the P-SE to regression problems. In addition, we deal with non-binary features, without learning the distribution of X nor of having the model for making predictions. Finally, we introduce local rule-based explanations for regression/classification based on the P-SE and compare our approaches w.r.t other explainable AI methods.

1 Introduction

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

Among these categories, the most popular are feature attributions approaches, in particular SHAP (Lundberg and Lee, 2017), which is based on Shapley Values (SV) and aims at indicating the importance of each feature in the decision. One of the main reasons for SHAP's success is its scalability, nice representations of the explanations, and mathematical foundations. However, SV used in SHAP does not guarantee the truthfulness of the important variables involved in a given decision. Indeed, it is possible to construct simple theoretical models (defined on a partition of the feature space) for which SV cannot distinguish between local important and non-important variables (see Appendix E in Amoukou et al. (2021)). Similar difficulties have also been highlighted by Ghalebikesabi et al. (2021) for SHAP and LIME (Ribeiro et al., 2016). This lack of guarantees is a major issue since the explanations may be used for high-stakes decisions. Moreover, the estimation of interaction effects requires extra work because of the additive nature of SV.

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

In this paper, we generalize the concept of *probabilistic Sufficient Explanations* (P-SE) introduced by Wang et al. (2020). P-SE is a relaxation of logic-based explanation: it explains the classification of an example by choosing a minimal subset of features. For a given instance, this means that if we know only this group of features, the model makes the same prediction with high probability, whatever the values of the remaining features (under the data distribution). Such a subset is called a Sufficient Explanation (also known as sufficient reason or prime implicant (Shih et al., 2018; Darwiche and Hirth, 2020).

We make several contributions. We adapt the definition of the Same-Decision Probability (SDP) to the regression setting, such that we can extend Sufficient Explanations from classification to regression. We in-

Preliminary work. Under review by AISTATS 2022. Do not distribute.

troduce a fast and efficient estimator of the SDP based on Random Forests and prove its uniform almost sure convergence. Our approach can deal with non-discrete features and does not need the estimation of the distribution of $\boldsymbol{X}$. Our method can explain the data generating process $(\boldsymbol{X}, Y)$ directly or any learnt model $(\boldsymbol{X}, f(\boldsymbol{X}))$. In particular, we don't need to have access to the model f, we need only to have access to the predictions, contrary to Wang et al. (2020).
We introduce the probabilistic local explanatory importance which is the frequency of each feature to be in the set of all Sufficient Explanations. In particular, this indicates the diversity of the Sufficient Explanations. We introduce local rule-based explanations for classification or regression which are simultaneously minimal and sufficient. We compare our approaches w.r.t other explainable AI methods and provide a Python package [1] that computes all our methods.

2 Motivations and Related works

The methods used to explain the local behavior of Machine Learning models can be organized into 5 groups: features attributions, decision rules, instance-wise feature selection, logical reasoning approaches, data generation based or counterfactual examples. The benefit of feature attribution-based explanations, e.g., SHAP (Lundberg et al., 2020) or LIME (Ribeiro et al., 2016) is that they can be applied to any model and are generally more scalable than their alternatives. On the other hand, they are very sensitive to perturbations (Ignatiev et al., 2019), can be fooled by adversarial attacks (Slack et al., 2020) and can be poor in explaining the local behavior of the model (Amoukou et al., 2021; Ghalebikesabi et al., 2021). These downsides can be caused by the local perturbations used, which make them inconsistent with the data distribution.

Quite differently, instance-wise feature selection such as L2X (Chen et al., 2018) or INVASE (Yoon et al., 2018) aims at finding the minimal subset of variables that are relevant for a given instance $\boldsymbol{x}$ and its label y. Interactions can be captured in that way. In addition, the identification of minimal subset $S = S(\boldsymbol{x})$ is well-formalized and the objective is to find S such that $\mathcal{L}(Y|\boldsymbol{X} = \boldsymbol{x}) \approx \mathcal{L}(Y|\boldsymbol{X}_S = \boldsymbol{x}_S)$. However, these methods are not reliable because they are prone to many approximation errors due to the training of several Neural Networks, and they provide no guarantees about the fidelity of the explanations (Jethani et al., 2021).

Anchors (Ribeiro et al., 2018) are local rule-based explanations that propose a solution to the reliability issue by providing an explanation with guarantees. It explains individual predictions of any classification model by finding a decision rule that reaches a given accuracy for a high percentage of the neighborhood of the instance. However, the method is only available for classification, requires discrete variables, and is inconsistent with the data distribution.

Logical Reasoning Approaches such as Sufficient Reasons (Shih et al., 2018; Darwiche and Hirth, 2020) select a minimal subset of features guaranteeing that, no matter what is observed for the remaining features, the decision will stay the same. It can be seen as an instance-wise feature selection but with guarantees of sufficiency and minimality (i.e., no subset of the set satisfies the sufficiency condition). However, since the guarantees are deterministic, it is often necessary to include many features into the explanation, making the explanation more complex and thus less intelligible. A relaxation of this method is the Sufficient Explanations (Wang et al., 2020) that gives probabilistic guarantees instead of deterministic guarantees, i.e., it requires that the prediction remains the same with high probability.

Sufficient Explanations addresses all of the issues raised above by giving a simple local explanation with guarantees while considering feature interactions and the data distribution. However, its area of application is very restricted. It is available in the case of classification with binary features and requires knowing/learning the distribution of the features. Moreover, we can obtain numerous Sufficient Explanations, which causes a selection problem as the whole set of explanations is not interpretable.

In this work, we propose a consistent method that efficiently finds the Sufficient Explanations of any data generating process $(\boldsymbol{X}, Y)$ or any model $(\boldsymbol{X}, f(\boldsymbol{X}))$, without learning the distribution of $\boldsymbol{X}$. We also summarize the diversity of the Sufficient Explanations. In addition, we propose local rule-based explanations for regression and classification models based on the Sufficient Explanations. To the best of our knowledge, it is the first local rule-based explanations for regression. Note that our method does not make any assumption about the model and do not need the model. We only need the observation $\boldsymbol{X}$ and its output Y.

3 Probabilistic Sufficient Explanations for Regression

Let us assume that we have a i.i.d sample $\mathcal{D}_n = (\boldsymbol{X}_i, Y_i)_{i=1,\dots,n}$ such that $(\boldsymbol{X}, Y) \sim P_{(\boldsymbol{X},Y)}$ where $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. We use uppercase letters for ran-

[1] github.com/aistats2022exp/ConsistentExplanations

dom variables and lowercase letters for their value assignments. For given a subset $S \subset [p]$, $\boldsymbol{X}_S = (X_i)_{i \in S}$ denotes a subgroup of the features.

We define the explanations of an instance $\boldsymbol{x}$ as the minimal subsets $\boldsymbol{x}_S, S \subset [p]$ such that given only those features, the model yields "almost" the same prediction $\boldsymbol{y}$ as on the complete example with high probability, under the data distribution $p(\boldsymbol{X})$. The main probabilistic reasoning tool that we use for our explanations is the Same-Decision Probability (SDP) (Chen et al., 2012). In earlier works (Wang et al., 2020), the SDP was defined only for classification, which, intuitively, gives us the probability that the classifier has the same output by ignoring some variables. To explain also regression models, we propose a definition of the SDP in the regression setting:

Definition 3.1. ***(Same Decision Probability of a regressor).*** *Given an instance $(\boldsymbol{x}, y)$, the Same-Decision Probability at level t of the subset $S \subset [\![1, p]\!]$, w.r.t $\boldsymbol{x} = (\boldsymbol{x}_S, \boldsymbol{x}_{\bar{S}})$ is*

$$SDP_S\left(y; \boldsymbol{x}, t\right) = P\left((Y - y)^2 \leq t \,|\boldsymbol{X}_S = \boldsymbol{x}_S\right)$$

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

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

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

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

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

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

Definition 3.4. ***(Minimal Sufficient Rule).*** *Given an instance $(\boldsymbol{x}, y)$, S a Minimal Sufficient Explanation, the rectangle $L_S(\boldsymbol{x}) = \prod_{i=1}^{|S|}[a_i, b_i], a_i, b_i \in \mathbb{R}$ is a Minimal Sufficient Rule if $L_S(\boldsymbol{x}) = \text{argmax}_{L(\boldsymbol{x})}\ Vol(L(\boldsymbol{x}))$, $\boldsymbol{x}_S \in L_S(\boldsymbol{x})$ and $\forall \boldsymbol{z}_S \in L_S(\boldsymbol{x})$, $SDP_S\left(\boldsymbol{y}; \boldsymbol{z}, t\right) \geq \pi$.*

Intuitively, the Sufficient Rule is a generalization of the Sufficient Explanation, i.e., instead of satisfying the minimality/sufficiency conditions of definition 3.2 if we fixed the values $\boldsymbol{X}_S = \boldsymbol{x}_S$, we want to satisfy these conditions on all the elements of a rectangle $L_S(\boldsymbol{x})$ that contains $\boldsymbol{x}_S$. We also want this rectangle to be of maximal volume such that it covers a large part of the input space. Thus, the Sufficient Rule captures the local behavior of the model around $\boldsymbol{x}$ while ensuring the minimality of the rule and guarantees on the outcome.

While Sufficient Rules are similar to Anchors introduced by Ribeiro et al. (2018), we emphasize two ma-

jor types of differences. The first one is that our framework for constructing rules can address regression problems, deal with continuous features, and do not need access to the model f. Moreover, if we have a model f and an instance $\boldsymbol{x}$, Anchors search the largest rule (or rectangle) $L_S(\boldsymbol{x})$ such that $P_Q(f(\boldsymbol{x}) = Y|\ \boldsymbol{X}_S \in L_S(\boldsymbol{x})) \geq \pi$ under an instrumental distribution Q. This is different from the Sufficient Rule that requires the stability of the prediction for all the observations in the rectangle i.e $\forall \boldsymbol{x}_S \in L_S(\boldsymbol{x}), P(f(\boldsymbol{x}) = Y|\ \boldsymbol{X}_S = \boldsymbol{x}_S) \geq \pi$. The second major difference is that the Sufficient Rule is based on the original distribution $P_{(X,Y)}$ as we use conditional distribution $P(Y|\boldsymbol{X}_S)$. At the contrary, anchors use local sampling perturbations (introducing another distribution Q). As we discuss in the next section, the effective computation of these rules is very different. Anchors use a heuristic approach to find the minimal rule, which might produce suboptimal minimal rules. The Sufficient Rules satisfy a minimality principle by definition.

4 SDP, Sufficient Explanations and Sufficient Rules via Random Forest

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

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

$$\begin{aligned} SDP_S(\boldsymbol{x}, t) &= P((Y-y)^2 \leq t|\boldsymbol{X}_S = \boldsymbol{x}_S) \qquad (1) \\ &= F_S(y + \sqrt{t}\ |\boldsymbol{X}_S = \boldsymbol{x}_S) - F_S(y - \sqrt{t}\ |\boldsymbol{X}_S = \boldsymbol{x}_S). \end{aligned}$$

Equation 1 shows that the only challenge is to estimate the Projected (or Conditional) CDF $F_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S) = P(Y \leq y|\boldsymbol{X}_S = \boldsymbol{x}_S)$. The variant of the original Random Forest proposed by Meinshausen and Ridgeway (2006) that estimates the CDF $F(y|\boldsymbol{X} = \boldsymbol{x}) = P(Y \leq y|\boldsymbol{X} = \boldsymbol{x})$ is not of interest to us because we want to estimate the Projected CDF $F_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S)$ for any S. The recent works by Bénard et al. (2021a,c) are much more relevant as they permit to estimate $E[Y|\boldsymbol{X}_S = \boldsymbol{x}_S]$ from a classical Random Forest that has learned to predict $E[Y|\boldsymbol{X} = \boldsymbol{x}]$. The idea is to extract a new Forest called Projected Forest from the original Forest, which is a projection of the original Forest along the S-direction.

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

4.1 Random Forest and Condition Distribution Function (CDF) Forest

A Random Forest (RF) is grown as an ensemble of k trees, based on random node and split point selection based on the CART algorithm (Breiman et al., 1984). The algorithm works as follows. For each tree, a_n data points are drawn at random with replacement from the original data set; then, at each cell of every tree, a split is chosen by maximizing the CART-criterion; finally, the construction of every tree is stopped when the total number of cells in the tree reaches the value t_n. For each new instance $\boldsymbol{x}$, the prediction of the l-th tree is:

$$m_n(\boldsymbol{x}, \Theta_l, \mathcal{D}_n) = \sum_{i=1}^{n} \frac{B_n(\boldsymbol{X}^i; \Theta_l)\ \mathbb{1}_{X^i \in A_n(\boldsymbol{x}; \Theta_l, \mathcal{D}_n)}}{N_n(\boldsymbol{x}; \Theta_l, \mathcal{D}_n)} Y^i \quad (2)$$

where:

- $\Theta_l, l = 1, \ldots, k$ are independent random vectors, distributed as a generic random vector $\Theta = (\Theta^1, \Theta^2)$ and independent of $\mathcal{D}_n$. Θ^1 contains indexes of observations that are used to build the tree, i.e. the boostrap sample and Θ^2 indexes of splitting candidate variables in each node. $\Theta_{1:k}$ denotes the sequence of Θ_l's.
- $A_n(\boldsymbol{x}; \Theta_l, \mathcal{D}_n)$ is the tree cell (leaf) containing $\boldsymbol{x}$
- $N_n(\boldsymbol{x}; \Theta_l, \mathcal{D}_n)$ is the number of boostrap elements that fall into $A_n(\boldsymbol{x}; \Theta_l, \mathcal{D}_n)$
- $B_n(\boldsymbol{X}^i; \Theta_l)$ is the boostrap component i.e. the number of times that the observation has been chosen from the original data.

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

$$m_{k,n}(\boldsymbol{x}, \Theta_{1:k}, \mathcal{D}_n) = \frac{1}{k} \sum_{l=1}^{k} m_n(\boldsymbol{x}; \Theta_l, \mathcal{D}_n) \qquad (3)$$

The Random Forest estimator (Eq. 3) can also be seen as an adaptive neighborhood procedure (Lin and Jeon, 2006). For every instance $\boldsymbol{x}$, the observations in $\mathcal{D}_n$ are weighted by $w_{n,i}(\boldsymbol{x}; \Theta_{1:k}, \mathcal{D}_n)$, $i = 1, \ldots, n$. Therefore, the prediction of Random Forests can be rewritten as

$$m_{k,n}(\boldsymbol{x}, \Theta_{1:k}, \mathcal{D}_n) = \sum_{i=1}^{n} w_{n,i}(\boldsymbol{x}; \Theta_{1:k}, \mathcal{D}_n) Y^i$$

where the weights are defined by

$$w_{n,i}(\boldsymbol{x}; \Theta_{1:k}, \mathcal{D}_n) = \sum_{l=1}^{k} \frac{B_n(X^i; \Theta_l)\, \mathbb{1}_{X^i \in A_n(\boldsymbol{x}; \Theta_l, \mathcal{D}_n)}}{k \times N_n(\boldsymbol{x}; \Theta_l, \mathcal{D}_n)} \quad (4)$$

Viewing a Random Forest as an adaptive nearest neighbor predictor offers natural estimates of more complex quantities (Cumulative hazard function (Ishwaran et al., 2008), Treatment effect (Wager and Athey, 2017), conditional density (Du et al., 2021). Therefore, just as $E[Y|\boldsymbol{X} = \boldsymbol{x}]$ is approximated by a weighted mean over observation of Y^i, $E[\mathbb{1}_{Y \leq y}|\boldsymbol{X} = \boldsymbol{x}]$ is approximated by the weighted mean over the observations of $\mathbb{1}_{Y^i \leq y}$ using the same weights $w_{n,i}(\boldsymbol{x}; \Theta_{1:k}, \mathcal{D}_n)$ as in the original RF defined in Equation 4. The approximation is

$$\widehat{F}(y|\boldsymbol{X} = \boldsymbol{x}, \Theta_{1:k}, \mathcal{D}_n) = \sum_{i=1}^{n} w_{n,i}(\boldsymbol{x}; \Theta_{1:k}, \mathcal{D}_n) \mathbb{1}_{Y^i \leq y} \quad (5)$$

To simplify the notations, we omit $\Theta_1, \ldots, \Theta_k, \mathcal{D}_n$ and we write directly $\widehat{F}_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S)$ for any S.

4.2 Projected Forest and Projected CDF Forest

We describe the idea of Projected Forest (PRF) and show how it is combined with the Quantile Regression Forest to build the estimator of the Projected CDF. The idea of PRF originally comes from Bénard et al. (2021c,a). The projection eliminates the variables not contained in S from the tree predictions, thus we can estimate $E[Y|\boldsymbol{X}_S]$ instead of $E[Y|\boldsymbol{X}]$. The general principle is to project the partition of each tree of the Forest on the subspace spanned by the variables in S. PRF uses an algorithmic trick that computes the projected partition without modifying the initial tree structures. We simply drop observations down in the initial trees, ignoring the splits which use a variable outside of S: when a split involving a variable outside of S is met, the observations are sent both to the left and right children nodes. Consequently, each observation falls in multiple terminal leaves of the tree. We drop the new query point $\boldsymbol{x}_S$ down the tree, following the same procedure, and retrieve the set of terminal leaves where $\boldsymbol{x}_S$ falls. Next, we collect the training observations which belong to every terminal leaf of this collection, in other words, we intersect the collection of leaves where $\boldsymbol{x}_S$ falls. Finally, we average the outputs Y^i of the selected training points to generate the estimation of $E[Y|\boldsymbol{X}_S = \boldsymbol{x}_S]$. Notice that such set of selected observations can be empty if $\boldsymbol{x}_S$ belongs to a large collection of terminal leaves. To avoid this issue, the algorithm is stopped before reaching a tree level where $\boldsymbol{x}_S$ falls in an empty cell. Recall that a partition of the input space is associated with each tree level, and consequently, the algorithm is stopped when it reaches a level where the minimal number of observations in the leaf where $\boldsymbol{x}_S$ falls is above $t_n > 0$. An efficient implementation of the PRF algorithm is detailed in the Supplementary Material. The associated PRF is $m_{k,n}^{(\boldsymbol{x}_S)}(\boldsymbol{x}_S) = \sum_{i=1}^{n} w_{n,i}(\boldsymbol{x}_S) Y^i$ where the weights are defined by

$$w_{n,i}(\boldsymbol{x}_S) = \sum_{l=1}^{k} \frac{B_n(X^i; \Theta_l)\, \mathbb{1}_{X^i \in A_n^{(\boldsymbol{x}_S)}(\boldsymbol{x}_S; \Theta_l, \mathcal{D}_n)}}{k \times N_n^{(\boldsymbol{x}_S)}(\boldsymbol{x}; \Theta_l, \mathcal{D}_n)}, \quad (6)$$

where $A_n^{(\boldsymbol{x}_S)}(\boldsymbol{x}_S; \Theta_l, \mathcal{D}_n)$ is the leaf of the associated Projected l-th tree where $\boldsymbol{x}_S$ falls and $N_n^{(\boldsymbol{x}_S)}(\boldsymbol{x}; \Theta_l, \mathcal{D}_n)$ is the number of boostrap observations that falls in $A_n^{(\boldsymbol{x}_S)}(\boldsymbol{x}_S; \Theta_l, \mathcal{D}_n)$. Therefore, we approximate the Projected CDF $F_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S) = P(Y \leq y|\boldsymbol{X}_S = \boldsymbol{x}_S)$ as in Equation 5 by using the weights of the Projected Forest defined in 6. The estimator of the Projected CDF is defined as:

$$\widehat{F}_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S) = \sum_{i=1}^{n} w_{n,i}(\boldsymbol{x}_S) \mathbb{1}_{Y^i \leq y} \quad (7)$$

4.3 Consistency of the Projected CDF Forest

In this section, we state our main result, which is the uniform a.s. convergence of the estimator $\widehat{F}_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S)$ to $F_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S)$. Meinshausen and Ridgeway (2006) showed the uniform convergence in probability of a simplified version of the estimator of the CDF defined in Eq. 5, where the weights $w_{n,i}(\boldsymbol{x}_S; \Theta_{1:k}, \mathcal{D}_n)$ are in fact considered to be non-random while they are indeed random variables depending on $(\Theta_l)_{l=1,\ldots,k}$, $\mathcal{D}_n$. Moreover, the boostrap was replaced by subsampling without replacement as in most studies that analyse the asymptotic properties of Random Forests (Scornet et al., 2015; Wager and Athey, 2017; Goehry, 2020). However, Elie-Dit-Cosaque and Maume-Deschamps (2020) showed the almost everywhere uniform convergence of both estimators (the simplified and the one defined in 5) under realistic assumptions with random boostrap samples. We follow their works to prove the consistency of the PRF CDF $\widehat{F}_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S, \Theta_{1:k}, \mathcal{D}_n)$ based on the following assumptions.

Assumption 1. *$\forall x \in \mathbb{R}^d$, the conditional cumulative distribution function $F(y|X = x)$ is continuous.*

Assumption 1 is necessary to get uniform convergence of the estimator.

Assumption 2. *For $l \in [k]$, we assume that the variation of the conditional cumulative distribution function within any cell goes to* 0.

$$\forall x \in \mathbb{R}^d, \forall y \in \mathbb{R}, \sup_{\boldsymbol{z} \in A_n(\boldsymbol{x}; \Theta_l, \mathcal{D}_n)} |F(y|\boldsymbol{z}) - F(y|\boldsymbol{x})| \xrightarrow{a.s} 0$$

Assumption 2 allows to control the approximation error of the estimator. If for all y, $F(y|.)$ is continuous, Assumption 2 is satisfied provided that the diameter of the cell goes to zero. Note that the vanishing of the diameter of the cell is a necessary condition to prove the consistency of general partitioning estimator (see chapter 4 in Györfi et al. (2002)). Scornet et al. (2015) show that it is true in RF where the bootstrap step is replaced by subsampling without replacement and the data come from additive regression models (Stone, 1985). The result is also valid for all regression functions, with a slightly modified version of RF, where there are at least a fraction γ observations in children nodes, and the number of splitting candidate variables is set to 1 at each node with a small probability. Under these small modifications, Lemma 2 from Meinshausen and Ridgeway (2006) gives that the diameter of each cell vanishes.

Assumption 3. *Let k and $N_n(\boldsymbol{x}; \Theta_l, \mathcal{D}_n)$ (number of boostrap observations in a leaf node), then there exists $k = \mathcal{O}(n^\alpha)$, with $\alpha > 0$, and $\forall \boldsymbol{x} \in \mathbb{R}^d$, $N_n(\boldsymbol{x}; \Theta_l, \mathcal{D}_n) = \Omega^2(\sqrt{n}(ln(n))^\beta)$, with $\beta > 1$ a.s.*

Assumption 3 allows us to control the estimation error and means that the cells should contain a sufficiently large number of points so that averaging among the observations is effective.

To prove the consistency of the PRF CDF $\widehat{F}_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S)$, we only need to verify the assumptions 1, 2, 3 on the parameters of the PRF CDF and the Projected CDF $F_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S) = P(Y \leq y|\boldsymbol{X}_S = \boldsymbol{x}_S)$.

Assumptions 1 and 2 are satisfied for the Projected CDF and the PRF CDF's leaves. Since by definition $A_n^{(\boldsymbol{x}_S)}(\boldsymbol{x}_S; \Theta_l, \mathcal{D}_n) \subset A_n(\boldsymbol{x}; \Theta_l, \mathcal{D}_n)$, if the variations within the cells of the RF vanish, it also vanishes in the projected forest. In addition, if the CDF $F(y|\boldsymbol{X} = \boldsymbol{x}) = F(y|\boldsymbol{X}_S = \boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}} = \boldsymbol{x}_{\bar{S}})$ is continuous, we can show by a straightforward analysis of parameter-dependent integral that the Projected CDF $F_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S) = \int F(y|\boldsymbol{X}_S = \boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}} = \boldsymbol{x}_{\bar{S}}) p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S) d\boldsymbol{x}_{\bar{S}}$ is also continuous. Since we control the minimal number of observations in the leaf of the Projected Forest by construction, Assumption 3 is also verified. Then, the PRF CDF satisfies also Assumption 1-3 which ensures its consistency thanks to Theorem 1.

Theorem 1. *Consider a RF satisfying Assumptions 1 to 3. Then,*

$$\forall \boldsymbol{x} \in \mathbb{R}^d, \sup_{y \in \mathbb{R}} |\widehat{F}_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S) - F_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S)| \xrightarrow{a.s} 0$$

The complete proof is in the Supplementary Material.

4.4 Estimation of SDP, Sufficient Explanations and Sufficient Rules

In this section, we show how we compute the SDP, Sufficient Explanations, and Sufficient Rules using the PRF CDF estimator. We derive from (7) the following consistent estimator of any SDP:

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

However, finding the A-SE/M-SE using a greedy algorithm is computationally hard, since the number of subsets is exponential. Therefore, we propose to reduce the number of variables by focusing only on the most influential variables. We search the Sufficient Explanations in the subspace of the 10-variables frequently selected in the RF used to estimate the SDP, reducing the complexity from 2^p to 2^{10}. This preselection procedure is already used in Bénard et al. (2021b,a), and it is mainly based on Proposition 1 of Scornet et al. (2015), which highlights the fact that RF naturally splits the most on influential variables. Note that the minimum number of selected variables is a hyperparameter.

To find the Sufficient Rules, we used the SDP's estimator $\widehat{SDP}_S$. By using the fact that $\widehat{SDP}_S$ partitions the space like a tree or a Random Forest, we do not need to discretize the continuous space to find the largest rectangle. We only need to find the leaves compatible with the conditions of the Sufficient Rule defined in 3.4. Given a Minimal Sufficient Explanations S of an instance $\boldsymbol{x}$, we already a have a rectangle $L_S(\boldsymbol{x})$ defined by the PRF CDF or $\widehat{SDP}_S$ that is the largest rectangle such that $\boldsymbol{x}_S \in L_S(\boldsymbol{x})$ and $\forall z_S \in L_S(\boldsymbol{x}), \widehat{SDP}_S(\boldsymbol{z}, t) = \widehat{SDP}_S(\boldsymbol{x}, t) \geq \pi$. By definition, it is the intersection of the cell of the trees where $\boldsymbol{x}_S$ falls, namely $\cap_{l=1}^k A_n^{(\boldsymbol{x}_S)}(\boldsymbol{x}_S; \Theta_l, \mathcal{D}_n)$. Thus, starting from $\cap_{l=1}^k A_n^{(\boldsymbol{x}_S)}(\boldsymbol{x}_S; \Theta_l, \mathcal{D}_n)$, which is also a cell (leaf) of the Projected Forest, we can find all the neighboring leaf (rectangles) that we can merge with it to get the largest rectangle. We will see in the next section that it provides good insights about the local behaviour of the model.

[2] $f(n) = \Omega(g(n)) \iff \exists k > 0, \exists n_0 > 0 | \ \forall n \geq n_0 | f(n)| \geq |g(n)|$

5 Experiments

We conduct three experiments in this section. The first aims to show the convergence of the estimator of the Projected CDF. The second compares the Sufficient Explanations and Sufficient Rules approaches with state-of-art local explanations methods (SHAP, LIME, INVASE) in a regression model. Finally, we highlight the advantages of the Sufficient Rules in comparison with Anchors in classification models.

To effectively compare different explanations methods, we use synthetic data since we need the ground truth. Thus, we use the following synthetic model for the first two experiments: we have $\boldsymbol{X} \in \mathbb{R}^p$, $\boldsymbol{X} \in \mathcal{N}(0, \Sigma)$, $\Sigma = 0.8J_p + 5I_p$ with $p = 50$, I_p is the identity matrix, J_p is all-ones matrix and a linear predictor defined as:

$$Y = (X_1 + X_2)\mathbb{1}_{X_5 \leq 0} + (X_3 + X_4)\mathbb{1}_{X_5 > 0}. \quad (8)$$

The variables X_i for $i = 6 \dots 49$ are noise variables. We fit a RF with a sample size $n = 10^4$, $k = 20$ trees and the minimal number of samples by leaf node is set to $t_n = \lfloor \sqrt{n} \times \ln(n)^{1.5}/250 \rfloor$ for the original and the Projected Forest. The mean squared error of the model on the test set of size 10^4 is MSE = 0.61. The original forests is used to compute the explanations of SHAP, LIME, and the Projected Forest for the SDP approaches. We choose t of the SDP for regression as the 0.95-th quantile of the MSE on the test-set and $\pi = 0.95$. For INVASE, we use Neural Networks with 3 hidden layers for the selector model, and the predictor model as in Yoon et al. (2018).

5.1 Empirical evaluations of $\widehat{F}_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S)$

In order to compare the PRF CDF $\widehat{F}_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S)$ and $F_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S)$, we use a Monte Carlo estimator to effectively compute $F_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S)$. We use a randomly chosen point $\boldsymbol{x}_S = [-0.13, 1.29, -1.31]$ with $S = [1, 2, 5]$ from the test set. The experiment is replicated 100 times. Figure 1 shows that the estimator works well for almost all points $y \in \mathbb{R}$.

We also compute two global metrics. For a given S, we compute the average Kolmogorov-Smirnov $MKS = \frac{1}{n}\sum_{i=1}^{n} \sup_{y\in\mathbb{R}} |\widehat{F}_S(y|\boldsymbol{X}_S = \boldsymbol{x}_{S,i}) - F_S(y|\boldsymbol{X}_S = \boldsymbol{x}_{S,i})|$ and the average mean absolute deviation $MAD = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} |\widehat{F}_S(y|\boldsymbol{X}_S = \boldsymbol{x}_{S,i}) - F_S(y|\boldsymbol{X}_S = \boldsymbol{x}_{S,i})| dy$.

We have MAD = 0.008 and the MKS=0.26 on all the observations with $S = [1, 2, 3, 5]$ showing the estimator's efficiency. We also compute them with small $S = [0, 4]$, it works even better with MAD=0.068, MKS=0.0098.

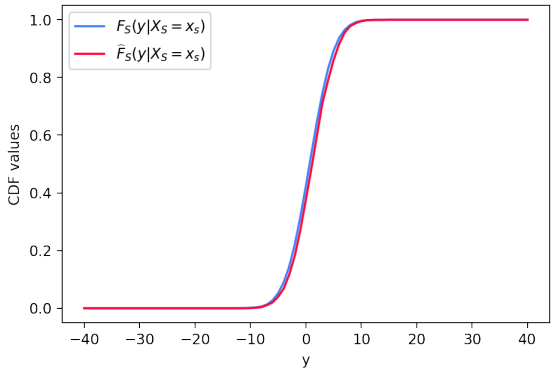

Figure 1: Comparison of $\widehat{F}_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S)$ and $F_S(y|\boldsymbol{X}_S = \boldsymbol{x}_S)$ with $S = [1, 2, 5]$ and $\boldsymbol{x}_S = [-0.13, 1.29, -1.31]$

	Metrics		
Methods	TPR	FDR	P-MSE
Sufficient Explanations	99%	2%	0.02
INVASE	99%	87%	0.006
SHAP	73%	27%	0.79
LIME	50%	49%	5.01

Table 1: TPR, FDR, P-MSE for each method

5.2 Comparisons with state-of-art

In this section, we analyze the capacity of each method to discover the important local variables of the model defined in Eq. 8. Indeed, Eq. 8 shows that if $x_5 \leq 0$, the model uses only the variables x_1, x_2 otherwise it uses the variables x_3, x_4. Thus, we try to find the top $K = 3$ relevant features for each sample. Note that K is not a required input for SDP and INVASE, but K must be given for SHAP and LIME. The performance metrics we use are the true positive rate (TPR) (higher is better) and false discovery rate (FDR) (lower is better) to measure the performance of the methods on discovery (i.e., discovering which features are relevant). In addition, as one of the objectives of each method is to find the minimal subset $\boldsymbol{x}_S$ that is relevant to the corresponding target y, we also use a predictive performance metrics that shows how well the projected predictor $E[Y|\boldsymbol{X}_S = \boldsymbol{x}_S]$ selected by each method is close to the predictor on the full set of features $E[Y|\boldsymbol{X} = \boldsymbol{x}]$, under the data distribution. Formally, for a given subset S, we denote it as P-MSE $= E_Z\left[\left(E[Y|\boldsymbol{X} = Z] - E[Y|\boldsymbol{X}_S = Z_S]\right)^2\right]$ where $Z \sim P_{\boldsymbol{X}}$. We observe in table 1 that the Minimal Sufficient Explanation estimated by SDP find the top K relevant variables and outperforms the other methods by a significant margin. SHAP and LIME obtain the worst discovery rate. INVASE succeeds in finding the relevant variables, but it has a high FDR

(87%), which means we cannot distinguish between the relevant and irrelevant variables since 87% of the selected variables are irrelevant. We also see that the P-MSE of INVASE is the lowest, which is not surprising as it selects all the relevant variables despite its high FDR. Indeed, this metric is not much affected by the FDR. The P-MSE of Sufficient Explanation is also almost zero, and as above, SHAP and LIME perform worse than the other methods.

However, even if the Sufficient Explanation find effectively the top K relevant variables, it cannot provide a complete understanding of the local behavior of the model, i.e., that it's the sign of x_5 that matters. Thus, by extending the Sufficient Explanation into Sufficient Rule we can retrieve the complete story. We choose an observation $\boldsymbol{x}$ such that its Sufficient Explanations found is $S = [3, 4, 5]$, with $\boldsymbol{x}_S = [-3.64, -4.41, 0.68]$. Although the Sufficient Explanation shows that fixing the value $\boldsymbol{x}_S$ permit to maintain the prediction with high probability, the Sufficient Rule gives the additional information that we can also maintain the prediction by satisfying the rule $L_S(\boldsymbol{x}) = \{X_5 > 0 \text{ and } -4,45 \leq X_4 \leq -4,06 \text{ and } -3.67 \leq X_3 \leq -3.58\}$. The Sufficient Rule $L_S(\boldsymbol{x})$ catches perfectly the local behaviour of the model which says that despite the values of x_3, x_4, it's the sign of x_5 that matters.

5.3 Anchors vs Sufficient Rules

To demonstrate the advantages of our method w.r.t. Anchors, we have to consider a classification problem. We ran both algorithms on a toy dataset and evaluated their capacity of providing good minimal rules. We use the moon dataset $(Z_1, Z_2, Y) \in \mathbb{R}^2 \times \{0, 1\}$, see figure 2, and we add gaussian features $\boldsymbol{X} \in \mathbb{R}^{100}$ with the μ, Σ of section 5 such that the final data is $(Z_1, Z_2, \boldsymbol{X}, Y)$. In addition, if $X_1 > 0$, we flip the label Y of the observation.

We train a RF as in section 5, with AUC=99% on the test set (of size 10^4 observations). We use Anchors with threshold $\tau = 0.95$, tolerance $\delta = 0.05$, and the Minimal Sufficient Rules with $\pi = 0.95$ to explain 1000 observations of the test set. We observe that, on average anchors tend to give much longer rules. The mean size for Sufficient Rule is 2, and for Anchors it is 10. In addition, the Minimal Sufficient Explanations detect local relevant variables more accurately. It has FDR=3%, TPR=100% and Anchors has FDR=48%, TPR=80%. Finally, we observe qualitatively the rules on a given example $\boldsymbol{x}$ (black star in figure 2). We test the stability of the explanations by comparing the rules of $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ a nearby observation such that $\max_i |x_i - \tilde{x}_i| \leq 0.05$ (yellow star in figure 2). The rule given by Anchors for $\boldsymbol{x}, \tilde{\boldsymbol{x}}$ are $L_{\text{Anchors}}(\boldsymbol{x}) = \{Z_1 > 1.03 \text{ AND } X_1 > 0.02 \text{ AND } Z_2 \leq -0.21\}$ and $L_{\text{Anchors}}(\tilde{\boldsymbol{x}}) = \{X_8 > -1.61 \text{ AND } X_{92} > 1.68 \text{ AND } Z_1 > -0.03 \text{ AND } X_1 > 0.02 \text{ AND } Z_2 \leq -0.21\}$. We find that the rules are very different, showing instability. Moreover, we also note that Anchors is very sensitive to random seeds. However, the SDP approach gives the same explanation for $\boldsymbol{x}, \tilde{\boldsymbol{x}}$. The observations have two Minimal Sufficient Explanations $S_1^\star = [\boldsymbol{z}_1, \boldsymbol{x}_1]$, $S_2^\star = [\boldsymbol{z}_2, \boldsymbol{x}_1]$. Thus, they have two Sufficient Rules, we can observe them given the axis Z_1, Z_2 in figure 2.

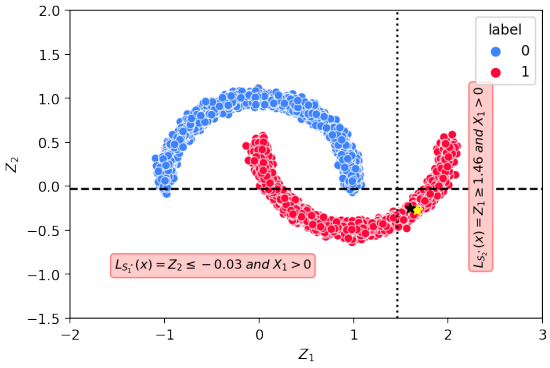

Figure 2: Explanations of $\boldsymbol{x}$, $\tilde{\boldsymbol{x}}$ by the two Sufficient Rules, the horizontal/vertical rectangle is associate with $S_1^\star = [\boldsymbol{z}_1, \boldsymbol{x}_1], S_2^\star = [\boldsymbol{z}_2, \boldsymbol{x}_1]$ respectively. The background samples are the observations with $\boldsymbol{x}_1 > 0$.

6 Conclusion

In this paper, we introduced three local explanations methods: Minimal Sufficient Explanations, Local Explanatory Importance, and Minimal Sufficient Rules. We proved that these methods considerably improve local variable detection over state-of-the-art algorithms while ensuring minimality, sufficiency, and stability. Our generalization of SDP and Minimal Sufficient Rules are tightly related. They are linked by a Random Forest, which is a computationally and statistically efficient estimator of the SDP and gives the partition that is translated into an interpretable rule. A remarkable feature of our extension of SDP is the new parameter t that corresponds to the level of variations around the prediction. The choice of t is essential as it can change the Sufficient Explanations. It adds additional complexity, but it should be dependent on the use case. A relevant default principle is to relate the value of t with the level of uncertainty of the prediction of $f(\boldsymbol{x})$ as we suggest in our experiments. Of course, better choices taking into account uncertainty or stability still needs to be explored. Besides, we believe that the extension of the Sufficient Rules into a simple global model is a promising research direction.

References

Amoukou, S. I., Brunel, N. J., and Salaün, T. (2021). Accurate and robust shapley values for explaining predictions and focusing on local important variables. *arXiv preprint arXiv:2106.03820*.

Bénard, C., Biau, G., Da Veiga, S., and Scornet, E. (2021a). Shaff: Fast and consistent shapley effect estimates via random forests. *arXiv preprint arXiv:2105.11724*.

Bénard, C., Biau, G., Veiga, S., and Scornet, E. (2021b). Interpretable random forests via rule extraction. In *International Conference on Artificial Intelligence and Statistics*, pages 937–945. PMLR.

Bénard, C., Da Veiga, S., and Scornet, E. (2021c). Mda for random forests: inconsistency, and a practical solution via the sobol-mda. *arXiv preprint arXiv:2102.13347*.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees. wadsworth int. *Group*, 37(15):237–251.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. (2018). Learning to explain: An information-theoretic perspective on model interpretation. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 882–891. PMLR.

Chen, S., Choi, A., and Darwiche, A. (2012). The same-decision probability: A new tool for decision making.

Chen, S., Choi, A., and Darwiche, A. (2013). An exact algorithm for computing the same-decision probability. IJCAI '13, page 2525–2531. AAAI Press.

Choi, Y., Vergari, A., and Van den Broeck, G. (2020). Probabilistic circuits: A unifying framework for tractable probabilistic models. Technical report, Technical report.

Darwiche, A. and Hirth, A. (2020). On the reasons behind decisions. *arXiv preprint arXiv:2002.09284*.

Du, Q., Biau, G., Petit, F., and Porcher, R. (2021). Wasserstein random forests and applications in heterogeneous treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 1729–1737. PMLR.

Elie-Dit-Cosaque, K. and Maume-Deschamps, V. (2020). Random forest estimation of conditional distribution functions and conditional quantiles. *arXiv preprint arXiv:2006.06998*.

Ghalebikesabi, S., Ter-Minassian, L., Diaz-Ordaz, K., and Holmes, C. (2021). On locality of local explanation models. *arXiv preprint arXiv:2106.14648*.

Goehry, B. (2020). Random forests for time-dependent processes. *ESAIM: Probability and Statistics*, 24:801–826.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*, volume 1. Springer.

Ignatiev, A., Narodytska, N., and Marques-Silva, J. (2019). On validating, repairing and refining heuristic ml explanations. *arXiv preprint arXiv:1907.02509*.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The annals of applied statistics*, 2(3):841–860.

Jethani, N., Sudarshan, M., Aphinyanaphongs, Y., and Ranganath, R. (2021). Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In *International Conference on Artificial Intelligence and Statistics*, pages 1459–1467. PMLR.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774.

Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(6).

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " 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.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

Shih, A., Choi, A., and Darwiche, A. (2018). A symbolic approach to explaining bayesian network classifiers. *arXiv preprint arXiv:1805.03364*.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). 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.

Stone, C. J. (1985). Additive Regression and Other Nonparametric Models. *The Annals of Statistics*, 13(2):689 – 705.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests.

Wang, E., Khosravi, P., and Van den Broeck, G. (2020). Towards probabilistic sufficient explanations. In *Extending Explainable AI Beyond Deep Models and Classifiers Workshop at ICML (XXAI)*.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). Invase: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*.