The Shapley Value of coalition of variables provides better explanations

Anonymous Authors[1]

Abstract

While Shapley Values (SV) are one of the gold standard for interpreting machine learning models, we show that they are still poorly understood, in particular in the presence of categorical variables or of variables of low importance. For instance, we show that the popular practice that consists in summing the SV of dummy variables is false as it provides wrong estimates of all the SV in the model and implies spurious interpretations. Based on the identification of null and active coalitions, and a coalitional version of the SV, we provide a correct computation and inference of important variables. Moreover, a Python library[1] that computes reliably conditional expectations and SV for tree-based models, is implemented and compared with state-of-the-art algorithms on toy models and real data sets.

1. Introduction

The explainability and interpretability of Machine Learning (ML) models are now central topics in Machine Learning Research due to their increasing ubiquity in Industry, Business, Sciences and Society. As ML models are usually considered as black-box models, scientists, practitioners and citizens call for the development of tools that could provide better insights in the important variables in a prediction, or in identifying biases for some individuals, or sub-groups. Typically, standard global importance measures such as permutation importance measures are not sufficient for explaining individual or local predictions and new methodologies are developed in the very active field of Explainable AI (XAI). Indeed, various local importance measures have been proposed with a particular focus on model-agnostic methods that can be applied to the most successful ML models, typically ensemble methods (such as random forests, gradient boosted trees) and deep learning. The most used are for instance Partial Dependence Plot (Friedman, 2001), Individual Conditional Expectation (Goldstein et al., 2015), and local feature importance attribution measures such as Local Surrogate (LIME) (Ribeiro et al., 2016). With the same objective in mind, the Shapley Values (Shapley, 1953), a concept primarily developed in Cooperative Game Theory, has been adapted to XAI for evaluating the "fair" contribution of a variable $X_i = x_i$ in a prediction (Strumbelj & Kononenko, 2010; Lundberg & Lee, 2017). The Shapley Values (SV) are now massively used for identifying important variables at a local and a global scale. As remarked in (Lundberg et al., 2020b), a lot of importance measures aim at analyzing the behavior of a prediction model f based on p features $X_1, \ldots, X_p$ by considering reduced predictors. Typically, for any group of variables $\boldsymbol{X}_S = (X_i)_{i \in S}$, with any subset $S \subseteq [\![1, p]\!]$ and reference distribution $Q_{S,\boldsymbol{x}}$, reduced predictors are defined as

$$f_S(\boldsymbol{x}_S) \triangleq E_{Q_{S,\boldsymbol{x}}}\left[f\left(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}\right)\right]. \tag{1.1}$$

Most often, the distribution $Q_{S,\boldsymbol{x}}$ is the conditional distribution or the marginal distribution of $\boldsymbol{X}_{\bar{S}}$. The SV for local interpretability at $\boldsymbol{x}$ have been introduced in (Lundberg & Lee, 2017) and are based on a cooperative game with value function $v(f; S) \triangleq f_S(\boldsymbol{x}_S)$. Then, for any coalition of variables $C \subseteq [\![1, p]\!]$ and $k \in [\![1, p - |C|]\!]$, we denote the set $\mathcal{S}_k(C) = \{S \subseteq [\![1, p]\!] \setminus C \,|\, |S| = k\}$: the Shapley Value (SV) of the coalition C is defined as

$$\phi_C(f; \boldsymbol{x}) = \frac{1}{p - |C| + 1} \sum_{k=0}^{p-|C|} \frac{1}{\binom{p-|C|}{k}} \sum_{S \in \mathcal{S}_k(C)} \left(f_{S \cup C}(\boldsymbol{x}_{S \cup C}) - f_S(\boldsymbol{x}_S)\right) \tag{1.2}$$

The definition 1.2 of the SV is a straightforward extension of the standard SV of a single variable (or player) to a group of variables, the standard SV is recovered with $C = \{i\}$ for $i \in [\![1, p]\!]$. We remark that this more general definition of SV is rarely used in practice, and is different from the SHAP interaction values that also consider groups (pairs) of variables, but for estimating the interaction effect between features. This is not our objective and we focus on the fact that eq. (1.2) can solve problems related to the computation

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]All the experiments and simulations can be reproduced with the publicly available library *Active Coalition of Variables* https://github.com/acvicml/ACV

of the SV of categorical variables, or the computation of SV in the presence of non-important variables.

These latter shortcomings are relatively new and differ from other well-known criticisms addressed to SV in XAI. One of the current active debate is about the causal interpretation of Shapley Values, see for instance (Heskes et al., 2020; Janzing et al., 2020; Chen et al., 2020). Technically, this problem revolves around the choice of the distribution $Q_{S,\boldsymbol{x}}$ used for computing the reduced predictor: the initial definition of SV for XAI uses the conditional probability in order to define the so-called observational SV. The alternative is partly motivated and justified by causal inference, and uses the marginal distribution $Q_{S,\boldsymbol{x}} = P_{\boldsymbol{X}_{\bar{S}}}$. Our analysis covers both cases.

Another common classical criticism is about the effective estimation of the expectations needed in the SV computation that is statistically challenging and combined with a exponential complexity. Current approaches might use synthetic observations generated by a distribution different from the original dataset (Lundberg & Lee, 2017; Aas et al., 2020). This typically happens when general ML models are considered: while we acknowledged that it is an important problem, we focus in that paper on tree-based models as the computational cost can be made polynomial and the statistical problem is easier to address (Lundberg et al., 2020b). Indeed, we propose a statistically principled way of estimating the reduced predictors in that setting. In addition, we show that a better estimation of the conditional expectations implies better estimates of the SV: although it might be obvious, we show that such kind of inaccuracies can have a predominant role in the estimation of Shapley values and might cause spurious interpretations. Our paper is organized as follows.

In section 2, we address the theoretical computation of SV for categorical variables when we use standard encodings, which motivates the use of equation (1.2). In particular, we show that the true SV of the categorical variable is different from the sum of SV of encoded variables. In order to illustrate our point, we consider toy models whose exact SV are in closed-form. In section 3, we address the statistical and efficient estimation of reduced predictors and Shapley Values in order to analyze real-world models. We compare our estimators with reference estimators on simulated data. In section 4, we elaborate on the concept of Same Decision Probability (SDP) (Wang et al., 2020) as a way to identify the group of most influential variables. We show that the Shapley Values and the additive explanation can be severely altered by the presence of non-important variables. In addition, we show that the computation of the SDP enables to identify two groups of variables: the active coalition with important variables, and the null coalition that gathers non-important variables. We can define a new game which provide better and insightful additive explanations thanks to the new corresponding Shapley Values. In section 5, we show that the SDP and active coalitions can be used for defining a global feature importance score that exhibits promising results, so that we can establish a link between local and global explanations. Finally, we give a short wrap-up and some directions of future works.

2. Multiple ways of computing the Shapley value of categorical variables

2.1. The Shapley value of a categorical variable

In this part, we show how to properly handle the SV of a categorical variable after encoding. For the sake of simplicity, we consider the case of two variables $\boldsymbol{X} = (X, Y)$ where $X \in \mathcal{X} \subseteq \mathbb{R}$ is a quantitative variable and $Y = 1, \ldots, K$ is a categorical variable. The SV gives the decomposition

$$f(x,y) - E_P\left[f\left(X,Y\right)\right] = \phi_X(f;x,y) + \phi_Y(f;x,y) \tag{2.1}$$

A direct application of eq. (2.1) gives for all $x \in \mathbb{R}$ and $Y \in \{1, \ldots, K\}$.

$$\left\{\begin{array}{rcl} \phi_X(f;x,y) & = & \frac{1}{2}\left(E\left[f(X,Y)|X=x\right] - E\left[f\left(X,Y\right)\right]\right) \\ & + & \frac{1}{2}\left(f\left(x,y\right) - E\left[f\left(X,Y\right)|Y=y\right]\right) \\ \phi_Y(f;x,y) & = & \frac{1}{2}\left(E\left[f(X,Y)|Y=y\right] - E\left[f\left(X,Y\right)\right]\right) \\ & + & \frac{1}{2}\left(f\left(x,y\right) - E\left[f\left(X,Y\right)|X=x\right]\right) \end{array}\right. \tag{2.2}$$

The role of variable X, Y are symmetric and the categorical or quantitative nature of the variable does not have any impact on the computation of SV. Nevertheless, categorical variables are very often not processed in this way during the training of statistical or machine learning models. Indeed, the variable Y is encoded in a more tractable way by introducing the dummy variables:

$$Y_k = \begin{cases} 1, & \text{if } Y = k \\ 0 & \text{otherwise} \end{cases}. \tag{2.3}$$

When the dummy variables are introduced in the model, we can compute their Shapley Values. Difficulties appear when we want to relate the SV of the dummies Y_k, $k = 1, \ldots, K-1$ to the SV of the variable Y as defined as in eq. (2.2). In order to establish this link, we introduce more notations. Let $\mathcal{C} : y \mapsto (y_1, \ldots, y_{K-1})$ be the encoding transformation, $Y_{1:K-1} = \mathcal{C}(Y)$, $\tilde{P}$ the distribution of $(X, Y_{1:K-1})$ on $\mathcal{X} \times \prod_{k=1}^{K-1}\{0,1\}$: $\tilde{P}$ is the image probability of P induced by $\mathcal{C}$. The function $f : \mathcal{X} \times \{1, \ldots, K\} \longrightarrow \mathbb{R}$ defines a function $\tilde{f} : \mathcal{X} \times \prod_{k=1}^{K-1}\{0,1\} \longrightarrow \mathbb{R}$ such that

$$f(X,Y) \triangleq \tilde{f}(X, Y_1, \ldots, Y_{K-1}).$$

The function $\tilde{f}$ is not completely defined for all $(y_1 \ldots, y_{K-1}) \in \prod_{k=1}^{K-1}\{0,1\}$ and is only defined $\tilde{P}$-almost everywhere because of the deterministic dependence

$\sum_{k=1}^{K-1} Y_k \leq 1$ induced by the dummy transformation $\mathcal{C}$ (2.3). Consequently, we need to extend $\tilde{f}$ to the whole space $\mathcal{X} \times \prod_{k=1}^{K-1} \{0,1\}$ by setting $\tilde{f}(x, y_1, \dots y_{K-1}) = 0$ as soon as $\sum_{k=1}^{K-1} y_k > 1$.

For the predictor $\tilde{f}(X, Y_1, \dots, Y_{K-1})$, we can compute the SV of $X, Y_1, \dots, Y_{K-1}$ and obtain the decomposition

$$\begin{aligned} &\tilde{f}(x, y_{1:K-1}) - E_{\tilde{P}}\left[\tilde{f}\left(X, Y_{1:K-1}\right)\right] \\ &= \phi_X^{indiv}(\tilde{f}; x, y_{1:K-1}) + \sum_{k=1}^{K-1} \phi_k(\tilde{f}; x, y_{1:K-1}) \end{aligned} \quad (2.4)$$

where $\phi_k(\tilde{f}; x, y_{1:K-1})$ are the SV of the variable Y_k computed with distribution $\tilde{P}$. By definition of $\tilde{f}$, we have $\tilde{f}(x, y_{1:K-1}) = f(x, y)$ and $E_{\tilde{P}}\left[\tilde{f}\left(X, Y_{1:K-1}\right)\right] = E_P\left[f\left(X, Y\right)\right]$. This implies the equality

$$\begin{aligned} \phi_X(f; x, y) + \phi_Y(f; x, y) = \phi_X^{indiv}(\tilde{f}; x, y_{1:K-1}) \\ + \sum_{k=1}^{K-1} \phi_k(\tilde{f}; x, y_{1:K-1}) \end{aligned} \quad (2.5)$$

In general, $\phi_Y(f; x, y) \neq \sum_{k=1}^{K-1} \phi_k(\tilde{f}; x, y_{1:K-1})$, in particular because SV depends on the number of variables. Indeed, we show in the next proposition that $\phi_Y(f; x, y) = \phi_C(\tilde{f}; x, y_{1:K-1})$ where C is the coalition of variables $(Y_1, ..., Y_{K-1})$.

Proposition 2.1. *For all $x \in \mathcal{X}$, and if $y_{1:K-1} = \mathcal{C}(y)$ then*

$$\begin{cases} \phi_C(\tilde{f}; x, y_{1:K-1}) &= \phi_Y(f; x, y) \\ \phi_X^{coal}(\tilde{f}; x, y_{1:K-1}) &= \phi_X(f; x, y) \end{cases} \quad (2.6)$$

We refer to Appendix A for detailed derivations. In general, for cooperative games, the SV of a coalition $\phi_C(\tilde{f}; x, y_{1:K-1})$ is different from the sum of individual Shapley effects $\sum_{k \in C} \phi_k(\tilde{f}; x, y_{1:K-1})$. We remark that we can compute two different SV for X when we use the encoded predictor $\tilde{f}$: $\phi_X^{coal}(\tilde{f}; x, y_{1:K-1})$ and $\phi_X^{indiv}(\tilde{f}; x, y_{1:K-1})$. These two SV are different in general as they involve different number of variables and different conditional expectations. Proposition 2.1 shows that we should prefer $\phi_X^{coal}(\tilde{f}; x, y_{1:K-1})$ to $\phi_X^{indiv}(\tilde{f}; x, y_{1:K-1})$, as ϕ_X^{coal} is equal to the theoretical SV given in eq. (2.2). For this reason, we denote for simplicity $\phi_X(\tilde{f}; x, y_{1:K-1})$.

2.2. Coalition or Sum: numerical comparisons

We give numerical examples illustrating the differences between coalition or sum and corresponding explanations.

2.2.1. Toy model

We consider a linear predictor f, with categorical and 3 continuous variables (X_0, X_1, X_2), defined as $f(X, Y) = B_Y X$ with $X|Y = y \sim \mathcal{N}(\mu_y, \Sigma_y)$ and $\mathbb{P}(Y = y) = \pi_y$, $Y \in \{a, b, c\}$. The values of the parameters used in our experiments are found in Appendix D.

In figure 1, we remark that the SV change considerably for a single observation. The sign changes given the encoding (dummies, One-Hot-Encoding - OHE) and is often different from the sign of the true SV of Y without encoding. We can also note important differences in the SV of the quantitative variable X.

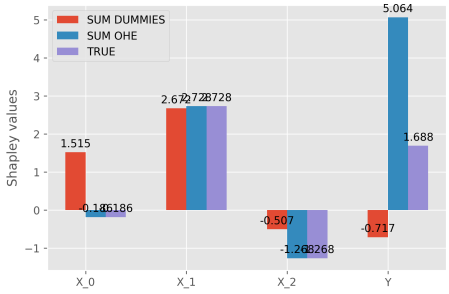

Figure 1. SV with or without encoding (OHE - Dummies) for observation $x = [-1.15, 3.50, -1.28, 1, 0]$.

In order to provide a global overview of the difference, we compute the SV of Y on 100 observations, by varying the first variable X_0 on a regular grid in $[-4, 2]$. We observe in figure 2 very high discrepancies between the different encoding and the true SV of Y. Moreover, the distribution of discrepancies is highly skewed and non-uniform on the grid, which makes it unpredictable, and difficult to anticipate in practice.

Figure 2. SV of the categorical variable of different observations with $X_0 \in [-4, 2]$ and $Y = a$.

2.2.2. Census Income Data

We use UCI Adult Census Dataset (Dua & Graff, 2017). We keep only 4 highly-predictive categorical variables: Marital Status, Workclass, Race, Education and use a Random Forest which has a test accuracy of 86%. We compare the SV by taking the coalition or sum of the modalities. SV are computed with algorithm 1 described in (Lundberg

et al., 2020a). In the rest of the paper, this estimator is denoted SHAP as it is the backbone implementation used in the Python library SHAP[2]. In figure 3, we see differences

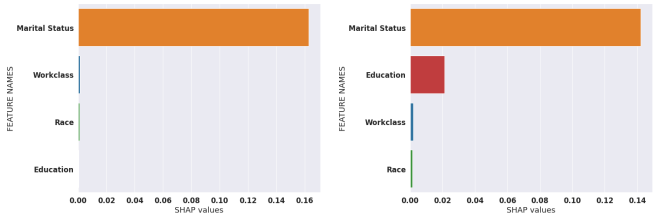

Figure 3. Difference between the absolute value of SV: coalition (left) vs sum (right) of dummies of individual (Married, local gov, others, 1rst-4th).

between the absolute value of SV with coalition and sum. The ranking of the variables changes, e.g. Education goes from important with sum to not important with the coalition. We also compute the proportion of order inversion over 5000 observations choose randomly. The ranking of variables is changed in 10% of the cases. Note that this difference may increase or diminish depending on the data.

3. Shapley values in tree-based models

The computation of the SV uses all the conditional expectations $E\left[f(\boldsymbol{X})|\boldsymbol{X}_S\right], S \subseteq [\![1,p]\!]$. While it is difficult in general, the paper (Lundberg et al., 2020b) introduce a recursive algorithm that reads sequentially and recursively the different nodes. In section 3.2, we show that the estimates and results obtained can be far from the true expectations. Hence, we suggest a straightforward Plug-In estimator of the true conditional expectations that is more accurate.

3.1. Closed-form expressions for reduced predictors

With tree-structured models, we can have efficient algorithms for computing in closed-form conditional expectations and SV. We assume that f is estimated from the data $\mathcal{D} = (x_i, z_i)_{1\leq i\leq n}$, and we show how we compute $f_S(\boldsymbol{x}_S)$. We assume that we have a regression tree with M leafs $L_1, \ldots, L_M$ based on the variables $X_1, \ldots, X_p$ (continuous or qualitative), the function f can be a predictor or any function estimated with a tree.

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} f_m \mathbb{1}_{L_m}(\boldsymbol{x}).$$

For a regression, we have $f_m = \frac{1}{N(L_m)} \sum_{i\in L_m} z_i$ where $N(L_m)$ is the number of observations that fall in leaf L_m.

[2]https://github.com/slundberg/shap, TreeExplainer uses an accelerated version of algorithm 1.

The reduced predictor is

$$f_S(\boldsymbol{x}_S) = \sum_{m=1}^{M} f_m Q_{S,\boldsymbol{x}}\left(L_m\right)$$

showing that the only challenge is the computation of the probabilities $Q_{S,\boldsymbol{x}}\left(L_m\right)$.

Estimation of conditional probability For simplicity, we assume that $Q_{S,\boldsymbol{x}} = P\left(\cdot\,|\boldsymbol{X}_S = \boldsymbol{x}_S\right)$ with computable conditional densities $p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S)$

$$Q_{S,\boldsymbol{x}}\left(L_m\right) = P_X\left(L_m\,|\boldsymbol{X}_S = \boldsymbol{x}_S\right) = \int_{L_m} p\left(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S\right) dx_{\bar{S}}$$

The leaf can be written as $L_m = \prod_{i=1}^{p}\left[a_i^m, b_i^m\right]$ (with $-\infty \leq a_i^m < b_i^m \leq +\infty$). It is useful to partition the leaf according to the coalition S: $L_m = L_m^S \times L_m^{\bar{S}}$ with $L_m^S = \prod_{i\in S}\left[a_i^m, b_i^m\right]$ and $L_m^{\bar{S}} = \prod_{i\in \bar{S}}\left[a_i^m, b_i^m\right]$. For each condition $\boldsymbol{X}_S = \boldsymbol{x}_S$, we introduce the set of compatible leaves for each $\boldsymbol{x} = (\boldsymbol{x}_S, \boldsymbol{x}_{\bar{S}})$

$$C\left(S, \boldsymbol{x}\right) = \left\{m \in [1\ldots M]\,|\boldsymbol{x}_S \in L_m^S\right\}$$

meaning that we only need to compute the conditional probabilities for the compatible leafs, and we have

$$P_X\left(L_m\,|\boldsymbol{X}_S = \boldsymbol{x}_S\right) = \begin{cases} 0 & \text{if } m \notin C\left(S, \boldsymbol{x}\right) \\ \int_{L_m^{\bar{S}}} p\left(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S\right) dx_{\bar{S}} & \text{if } m \in C\left(S, \boldsymbol{x}\right) \end{cases}$$

If we make no assumption on the shape of the distribution $p(\boldsymbol{x})$ nor $p\left(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S\right)$ as (Aas et al., 2020), we need to estimate $P_X\left(L_m\,|\boldsymbol{X}_S = \boldsymbol{x}_S\right)$ directly from the data set $\mathcal{D}$.

When all the variables are categorical, we can estimate directly by empirical conditional frequencies. Indeed, a straightforward estimation is based on $N(\boldsymbol{x}_S)$: the number of observations such that $\boldsymbol{X}_S = \boldsymbol{x}_S$ (across all the leaves of the tree) and $N(L_m, \boldsymbol{x}_S)$: the number of observations in leaf L_m that satisfies the condition $\boldsymbol{X}_S = \boldsymbol{x}_S$. We have

$$P_X\left(L_m\,|\boldsymbol{X}_S = \boldsymbol{x}_S\right) = \frac{N(L_m, \boldsymbol{x}_S)}{N(\boldsymbol{x}_S)}.$$

However, this estimator can have a high variance when there is a small number of observations in the leaf satisfying these conditions. In order to avoid this problem, we replace the condition $\left\{\boldsymbol{X}_S = \boldsymbol{x}_S\right\}$ by the condition $\left\{\boldsymbol{X}_S \in L_m^S\right\}$. We have

$$P_X\left(L_m\,\middle|\boldsymbol{X}_S \in L_m^S\right) = \frac{N(L_m)}{N(L_m^S)}$$

where

- $N(L_m)$ is the number of observations in the leaf L_m
- $N(L_m^S)$ is the number of observations satisfying the conditions $\boldsymbol{x}_S \in L_m^S$ across all the leaves of the tree.

When the variables $\boldsymbol{X}_S$ are quantitative, the computation becomes more challenging, and a standard approach is to use kernel smoothing estimators (with Parzen-Rosenblatt kernels). The main drawback of this approach is the low rate of convergence in high dimensions, and the need to choose adaptively an appropriate bandwidth, which might add complexity and instability to the whole estimation procedure. We propose to overcome this by replacing the condition $\{\boldsymbol{X}_S = \boldsymbol{x}_S\}$ by the condition $\left\{\boldsymbol{X}_S \in L_m^S\right\}$. As in the categorical case, we introduce a bias for the estimation of the conditional in order to improve the variance. In both cases, we use the same empirical estimator of the conditional probability, i.e

$$P_X\left(L_m \,|\boldsymbol{X}_S = \boldsymbol{x}_S\right) \simeq P_X\left(\boldsymbol{X}_{\bar{S}} \in L_m^{\bar{S}} \,\middle|\boldsymbol{X}_S \in L_m^S\right) \simeq \frac{N(L_m)}{N(L_m^S)} \tag{3.1}$$

Consequently, the reduced predictor could be approximated by $\sum_{m\in C(S,\boldsymbol{x})} f_m \frac{N(L_m)}{N(L_m^S)}$. However, our estimator of the conditional probabilities does not sum to one, i.e

$$\sum_{m\in C(S,\boldsymbol{x})} \frac{N(L_m)}{N(L_m^S)} \neq 1$$

because the denominator $N(L_m^S)$ varies with m. We correct the estimated probabilities with a softmax function $\sigma = (\sigma_m)_{m\in C(S,x)}$, applied to the vector $\left(\frac{N(L_m)}{N(L_m^S)}\right)_{m\in C(S,\boldsymbol{x})}$.

Hence, our final estimate of the reduced predictor is

$$\hat{f}_S(\boldsymbol{x}_S) = \sum_{m\in C(S,\boldsymbol{x})} f_m \sigma_m\left(\frac{N(L_m)}{N(L_m^S)}\right) \tag{3.2}$$

and it should be compared to the true $f_S(\boldsymbol{x}_S)$. We see that the essential challenge in the computation of the restricted estimator is the identification of the set of compatible leaves $C(S, \boldsymbol{x})$.

3.2. Comparisons of the different estimators

To compare the different estimators, we need a model where conditional expectations can be calculated exactly. If $X \sim \mathcal{N}(\mu, \Sigma)$ then $X_{\bar{S}}|X_S$ is also multivariate gaussian with explicit mean vector $\mu_{\bar{S}|S}$ and covariance matrix $\Sigma_{\bar{S}|S}$, see Appendix A.

Let assume we have a sample $\mathcal{D} = \{(\boldsymbol{x}_i, z_i)\,, i = 1, \ldots, n\}$ generated by a linear regression model with $X \in \mathbb{R}^3$ a Gaussian variable and target $Z = B^t X$.

We use a highly accurate decision tree regressor trained on $\mathcal{D}$ with $n = 10000$, parameters can be found in Appendix D. Since we know the law of $\boldsymbol{X}$, we compute

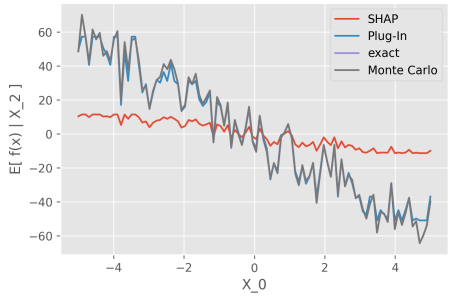

Figure 4. Estimation of expectation $\mathbb{E}\left[f(\boldsymbol{X})|X_2\right]$ with different observations with $X_0 \in [-4, 4]$.

$E\left[f(\boldsymbol{X})|\boldsymbol{X}_2 = x_2\right]$ exactly over a regular grid of 100 for $x_0 \in [-5, 5]$ and a Monte-Carlo estimator (MC) by resampling 10000 observations. We compare this two estimators with SHAP and the Plug-In estimator defined in equation 3.2. Not surprisingly, we observe in figure 4 that the MC estimator and the true expectation are both equal. The SHAP estimator has very large discrepancies, while the Plug-In estimator is very close to the truth.

Since the computation of SV uses all the conditional expectations $E\left[f(\boldsymbol{X})|\boldsymbol{X}_S = \boldsymbol{x}_S\right], S \subseteq [\![1, p]\!]$, we can use it to highlight the overall differences between the estimators. We give in figure 5 the SV of 200 obervations for the different estimators and we compare them with the Shapley values computed with the exact probabilities and plot the ℓ_1 norm of their differences. This emphasizes the results in figure 4.

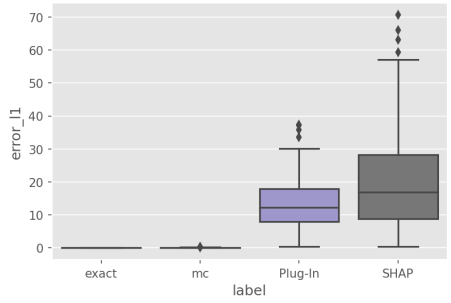

Figure 5. Errors distribution between the true SV with exact expectation and the corresponding algorithm (SHAP, Plug-In, MC) over 200 new observations generated by $P_{\boldsymbol{X}}$.

4. Identifying Active and Null coalitions

In general, we are not only interested in computing feature importance $\phi_i(f, \boldsymbol{x})$, we also want to identify the group of variables $X_i, i \in S$ that best explains $\boldsymbol{x}$ and the group of uniformative variables $\boldsymbol{X}_i, i \in \bar{S}$. Therefore, several papers (Zaeri-Amirani et al., 2018; Cohen et al., 2007; Sun et al., 2012) suggest to use SV as a heuristic for feature selection,

but as proved in (Ma & Tourani, 2020), the magnitude of SV of variables do not necessarily correspond to relevant variables. Here, we split the set of variables as an active coalition (composed of important variables) and the null coalition, composed of the variables with a low influence of the prediction $f(\boldsymbol{x})$. Then we show that the SV computed only for the variables in the active set are better estimates of the effect of the corresponding variables on the prediction.

4.1. Same Decision Probability

Our methodology for identifying the most important features is based on the Same Decision Probability (SDP) criterion, introduced in (Chen et al., 2012). In particular, we extend the work of (Wang et al., 2020) to the regression case, and we derive an efficient way to approximate the SDP for tree-based models.

Definition 4.1. ***(Same Decision Probability of a classifier).*** *Let $f : \mathcal{X} \longrightarrow [0,1]$ a probabilistic predictor and its classifier $C(\boldsymbol{x}) = \mathbb{1}_{f(\boldsymbol{x})\geq T}$ with threshold T, the Same Decision Probability of coalition $S \subset [\![1,p]\!]$, w.r.t $\boldsymbol{x} = (\boldsymbol{x}_S, \boldsymbol{x}_{\bar{S}})$ is*

$$SDP_S\left(C;\boldsymbol{x}\right) = P\left(C(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) = C(\boldsymbol{x}) \,|\boldsymbol{X}_S = \boldsymbol{x}_S\right)$$

SDP gives the probability to keep the same decision $C(\boldsymbol{x})$ when we do not observe the variables $\boldsymbol{X}_{\bar{S}}$. The higher is the probability, the better is the explanation based on S. Therefore, we want to identify the **minimal** subset of features such that the classifier makes the same decision with high probability π, given only them. More formally:

Definition 4.2. ***(Sufficient Coalition).*** *Given C a binary classifier, an observation $\boldsymbol{x} = (\boldsymbol{x}_S, \boldsymbol{x}_{\bar{S}})$, $S \triangleq S^\star_\pi(\boldsymbol{x})$ is a Sufficient Coalition for probability π if:*

1. $SDP_{S^\star_\pi(\boldsymbol{x})}\left(C;\boldsymbol{x}\right) \geq \pi$
2. *No subset Z of $S^\star_\pi(\boldsymbol{x})$ satisfies $SDP_Z\left(f;\boldsymbol{x}\right) \geq \pi$*

In order to find the coalition $S^\star_\pi(\boldsymbol{x})$, we need to be able to compute the SDP for any subset S. However, computing the SDP is known to be computationally hard. Even for a simple Naive Bayes model and classifier, computing SDP is NP-hard (Chen et al., 2013). Consequently, approximate criterion based on expectations instead of probabilities have been introduced see (Wang et al., 2020). In that section, we show that we can compute exactly and efficiently the Same Decision Probability in tree-based model by relying on reduced predictors.

Proposition 4.1. *Let f a probabilistic predictor and its binary classifier C with threshold T, $Q_{S,\boldsymbol{x}}$ the law of $\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S = \boldsymbol{x}_S$, then the $SDP_S\left(f;\boldsymbol{x}\right)$ can be written explicitly with the reduced predictor:*

$$SDP_{S,\boldsymbol{x}} = \frac{d - d^-}{d^+ - d^-} \tag{4.1}$$

Where $d = E_{Q_{S,\boldsymbol{x}}}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})\right]$, $d^+ = E_{Q_{S,\boldsymbol{x}}}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})|f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) \geq T\right]$ and $d^- = E_{Q_{S,\boldsymbol{x}}}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})|f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) < T\right]$.

The proof can be found in Appendix A.
In the case of the regression, the SDP and sufficient coalition can be defined straightforwardly and we can obtain the same kind of decomposition as in proposition 4.1.

Definition 4.3. ***(Same Decision Probability of a regressor).*** *Let $f : \mathcal{X} \longrightarrow \mathcal{Z}$ a regressor, the Same Decision Probability at level t of coalition $S \subset [\![1,p]\!]$, w.r.t $\boldsymbol{x} = (\boldsymbol{x}_S, \boldsymbol{x}_{\bar{S}})$ is*

$$SDP_S\left(f;\boldsymbol{x},t\right) = P\left(d\left(f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}), f(\boldsymbol{x})\right) \leq t \,|\boldsymbol{X}_S = \boldsymbol{x}_S\right)$$

In a regression setting, the SDP gives the probability to stay close to the same prediction $f(\boldsymbol{x})$ at level t, when we do not observe variables $\boldsymbol{X}_{\bar{S}}$.

Proposition 4.2. *Let f be a regressor, $Q_{S,\boldsymbol{x}}$ the law $\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S = \boldsymbol{x}_S$, then the $SDP_S\left(f;\boldsymbol{x},t\right)$ can be written explicitly as:*

$$SDP_S\left(f;\boldsymbol{x},t\right) = \frac{d^+ - d}{d^+ - d^-} \tag{4.2}$$

Where $d = E_{Q_{S,\boldsymbol{x}}}\left[d(f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}), f)\right]$, $d^+ = E_{Q_{S,\boldsymbol{x}}}\left[d(f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}), f) \mid d(f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}), f) > t\right]$ and $d^- = E_{Q_{S,\boldsymbol{x}}}\left[d(f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}), f) \mid d(f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}), f) \leq t\right]$.

The proof can be found in Appendix A.
Propositions 4.1 and 4.2 are useful for computing effectively the SDP for a model. In particular, prop. 4.1 shows that it can be accurately estimated with the reduced predictors and the Plug-In estimators (3.2). In the case of the regression, the computation of the SDP is not as straightforward for general distance $d(\cdot,\cdot)$. Nevertheless, if d is the Euclidean distance and f a tree-based model, we can compute efficiently the SDP with eq. (4.2) by exploiting the tree structure of the model that partition the feature space.

Proposition 4.3. *Let $f(\boldsymbol{x}) = \sum_{m=1}^{M} f_m \mathbb{1}_{L_m}(\boldsymbol{x})$ be a tree function, $Q_{S,x}$ is the conditional law given $\boldsymbol{X}_S = \boldsymbol{x}_S$. The conditional expectation of the Euclidean distance can be computed with the Plug-In estimator 3.2:*

$$E_{Q_{S,\boldsymbol{x}}}\left[d(f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}), f)\right] = E_{Q_{S,\boldsymbol{x}}}\left[(\sum_m f_m \mathbb{1}_{L_m}(\boldsymbol{x}) - f)^2\right]$$
$$= \sum_m f_m^2 P_{Q_{S,\boldsymbol{x}}}\left[L_m(\boldsymbol{x})\right] + f^2 - 2f\sum_m f_m P_{Q_{S,\boldsymbol{x}}}\left[L_m(\boldsymbol{x})\right] \tag{4.3}$$

The probability $\mathbb{P}_{Q_{S,\boldsymbol{x}}}\left[L_m(\boldsymbol{x})\right]$ is approximated by $\frac{N(L_m)}{N(L_m^S)}$.

We can adapt eq. (4.3) to compute all the terms in the decomposition (4.2) of $SDP_S\left(f;\boldsymbol{x},t\right)$. We can also extend

this result to tree-ensemble models like Random Forest, see Appendix A.

Based on the computation of the SDP of any coalition given by the previous propositions, we can derive an algorithm that finds the Sufficient Coalitions for probability π i.e $S^\star_\pi(\boldsymbol{x})$. Unlike SV computation, we don't have to compute all the conditional expectations for all subsets in order to find the coalition $S^\star_\pi$. We use a greedy algorithm that computes the SDPs for subsets of increasing sizes (starting from 1) until we find a minimal subset satisfying the Sufficient Coalition conditions. The algorithm is described in Algorithm 1 and defines the function *returnSubsets($\boldsymbol{x}$, size)* that returns all subsets of length *size* of $\boldsymbol{x}$.

Algorithm 1 Find Sufficient Coalition

Input : $\boldsymbol{x}, \pi$
  n = $length(\boldsymbol{x})$
  find = `False`
  bestSdp = -1
  for size = 1 **to** n **do**
    for $S \subset$ *returnSubsets*($\boldsymbol{x}$, size) **do**
      sdp = $SDP_S(\boldsymbol{x}, f)$
      if sdp $\geq \pi$ and sdp $\geq$ `bestSdp` **then**
        bestSdp = sdp
        $S^\star_\pi = S$
        find = True
      end if
    end for
    if find = True **then**
      return $S^\star_\pi$
    end if
  end for

4.2. Active Shapley values

In general, the top ranked SV do not necessarily correspond to relevant variables. However, the notion of SDP allows us to identify these variables: $S^\star_\pi$ is the set of active features such that the model make the same decision with high probability π as on the full example $\boldsymbol{x}$ and $N_\pi(\boldsymbol{x})$ is the set of remaining variables. Since the variables in $N_\pi(\boldsymbol{x})$ are not important for the prediction, we don't have to involve them in the computation of SV. But we can still use the SV to highlight the individual effects of important variables. We introduce a new XAI game where the SV of variables in $N_\pi(\boldsymbol{x})$ are fixed to zero.

Definition 4.4. ***(Active Shapley values).*** *Let f a model, $\boldsymbol{x}$ an instance, $S^\star_\pi(x)$ the Sufficient Coalition of level π of $\boldsymbol{x}$, $N_\pi(\boldsymbol{x})$ is the set of remaining variables (the Null Coalition). We define the new cooperative game with value function $v^\star$ defined for all S in $S^\star_\pi(x)$, such that*

$$v^\star(f;S) \triangleq f_{S\cup N_\pi(\boldsymbol{x})}(\boldsymbol{x}_{S\cup N_\pi(\boldsymbol{x})})$$

and $v^\star(f;\emptyset) = E[f(X)]$. For all the variables X_i in $S^\star_\pi$, we define the new Shapley value as

$$\phi^\star_i(f;\boldsymbol{x}) = \frac{1}{|S^\star_\pi|}\sum_{k=0}^{|S^\star_\pi|-1}\frac{1}{\binom{|S^\star_\pi|-1}{k}}\sum_{S\in\mathcal{S}_k(S^\star_\pi(\boldsymbol{x}))} v^\star(S\cup i) - v^\star(S)$$

Remark 4.1. This game is different from the standard game 1.2 because we consider only the reduced predictors obtained by conditioning with the coalition $N_\pi(\boldsymbol{x})$. We keep the same reference $\phi^\star_0(f;\boldsymbol{x}) = E[f(\boldsymbol{X})]$, such that we keep explaining the same difference $f(x) - E[f(\boldsymbol{X})]$ by the additive explanation (accuracy) formula.

Examples: Let us assume that we have a predictive model built on the toy model of Section 2.2.1 with six variables $(X_0,\dots,X_4,Y)$, and we have a highly accurate decision tree. Parameters of the model can be found in Appendix D. We choose an observation in class $Y=b$ where the relevant variables are X_0, X_1, X_2, Y: indeed the coefficients B_b are $[9,5,-8,0,0]$.

In figure 6, we observe that relevant variables are indeed found in the Sufficient Coalition. In applications, we found that $\pi = 90\%$ gives relevant coalitions. Therefore, we

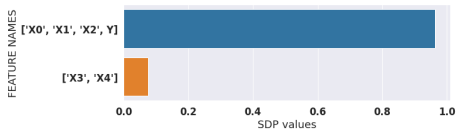

Figure 6. SDP for the Sufficient Coalition (blue) found by Algorithm 1 with $\pi = 0.9$, for observation $x = [2.06, 2.01, -0.90, -1.91, -3.70, 1., 0]$. The remaining variables are the Null Coalition with corresponding SDP in orange.

compute the Active SV based on the Sufficient Coalition. We observe in left part of figure 7 that we recover the individual effects of the important variables and they are consistent with the importance order of the variables whereas classic SV decomposition wrongly find X_3 as the most important variable, see right figure 7.

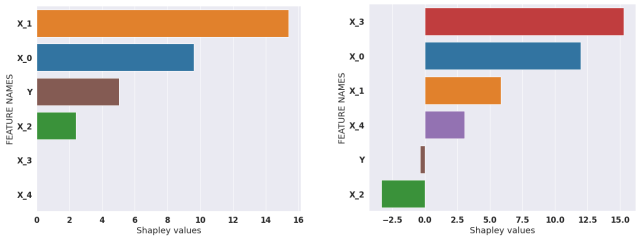

Figure 7. Left figure: SV $\phi^\star_i$ computed with the Sufficient Coalition given in figure 6. Right figure: SV ϕ_i computed with all the variables.

5. From local to global explanations with SDP

In this section, we show that the SDP gives a natural way to go from local to global variable importance scores: we interpret the task of computing a global score as the stability across the data set of the selection process of the Sufficient Coalition. By analogy with the stability selection process introduced in (Meinshausen & Bühlmann, 2010) and advocated in (Yu & Kumbier, 2020), we introduce the following definition:

Definition 5.1. ***(SDP-Global).*** *Let $X = (X_1, ..., X_p)$ be features with law P and a predictor f. The Global SDP importance of X_i is defined as:*

$$SDP_{global}(X_i) = \mathbb{E}_P\left[\mathbb{1}_{X_i \in S^\star_\pi(\boldsymbol{x})}\right]$$

Hence, a variable with a high SDP-global appears very often in the Sufficient Coalitions across all the dataset. We consider now a synthetic simple Bayesian model that possesses a causal structure where we can challenge the accuracy of the inference of important variables.

Explanations based on SDP We use an accurate decision tree trained on LUCAS (LUCAS), a dataset generated by causal Bayesian networks with 12 binary variables. The graph is drawn in figure 8 and we provide the probability table in Appendix D. We want to explain an observation with

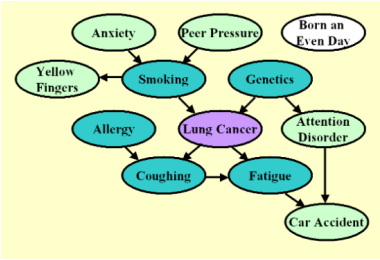

Figure 8. The target variable is shaded in purple. The nodes in dark green constitute the Markov blanket of the target variable.

a well-defined ground truth. We know from the probability tables that if Smoking, Genetic, Coughing are False, the probability of having Cancer is very low. So, we should have these three variables in the Sufficient Coalition: this is what we can observe in figure 9. The example below shows

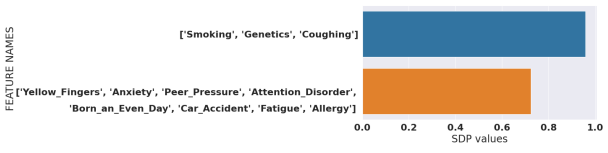

Figure 9. Sufficient coalition of an observation that has Smoking, Genetic, Coughing at False.

that we are able to identify the group of variables that drive the occurrence of a cancer. As we see feature importance as a stability property accross the dataset, we propose to compute the stability of Sufficient Coalitions defined as (5.1) in different strata. Natural strata to consider are obviously defined by the response variable $\{Y = k\}, k = 0, 1$. The corresponding global SDP are given in figure 5. For the negative class, it uses mainly Smoking, Coughing and Genetic whereas for the positive class is based on many more variables. We can have a more robust estimation of the individual explanations, while we still have a high-level overlook of the important variables across the data set. Hence we are able to aggregate the local explanations in order to get a global explanation. This aggregation is not additive but is based on the frequencies. We think that is an important benefit of SDP over Shapley values. Indeed, we have seen that the additive explanations based on SV are highly influenced by non-important variables: this makes the summation of (absolute) SV over all the data set prone to errors and misunderstandings. We have also shown that we can have better results by conditioning with the Null Coalition.

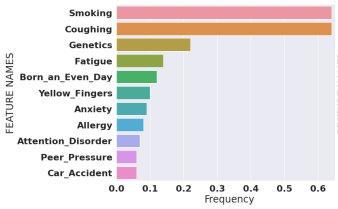

Figure 10. Global SDP given $Y = 0$.

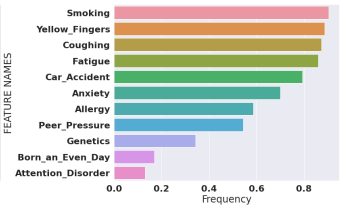

Figure 11. Global SDP given $Y = 1$.

6. Conclusion

An important finding of this work is the central role of the reduced predictors, and the ability to compute them efficiently under statistical and computational constraints. We have shown that we can estimate them in a principled way for tree-based models. Thanks to their computation, we are able to propose solutions to practical important problems in XAI. The same decision probability appears to be a useful and tractable tool for intelligibility. An interesting path to follow is to relate the different concepts in XAI with SDP and the various Shapley Values. In particular, when performing counter-factual analysis, the definition of SDP implies that it is more efficient to look for a counter-factual example by changing the variables $X_i, i \in S^\star_\pi(\boldsymbol{x})$, rather than the variables in $N_\pi(\boldsymbol{x})$. An other open question in SDP is the choice of t and π that might depend on the context and the corresponding admissible error. A natural choice is to relate t to the prediction variance, that can be estimated by resampling techniques (Barber et al., 2019).

References

Aas, K., Jullum, M., and Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to shapley values, 2020.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. *arXiv preprint arXiv:1905.02928*, 2019.

Chen, H., Janizek, J. D., Lundberg, S., and Lee, S.-I. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

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

Chen, S., Choi, A., and Darwiche, A. An exact algorithm for computing the same-decision probability. IJCAI '13, pp. 2525–2531. AAAI Press, 2013. ISBN 9781577356332.

Cohen, S. B., Dror, G., and Ruppin, E. Feature selection via coalitional game theory. *Neural Comput.*, 19(7): 1939–1961, 2007. doi: 10.1162/neco.2007.19.7.1939. URL https://doi.org/10.1162/neco.2007.19.7.1939.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232, 10 2001. doi: 10.1214/aos/1013203451. URL https://doi.org/10.1214/aos/1013203451.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24 (1):44–65, 2015.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916. PMLR, 2020.

LUCAS, D. Lucas (lung cancer simple set) dataset. http://www.causality.inf.ethz.ch/data/LUCAS.html.

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

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

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020b.

Ma, S. and Tourani, R. Predictive and causal implications of using shapley value for model interpretation. In Le, T. D., Liu, L., Zhang, K., Kiciman, E., Cui, P., and Hyvärinen, A. (eds.), *Proceedings of the 2020 KDD Workshop on Causal Discovery (CD@KDD 2020), San Diego, CA, USA, 24 August 2020*, volume 127 of *Proceedings of Machine Learning Research*, pp. 23–38. PMLR, 2020. URL http://proceedings.mlr.press/v127/ma20a.html.

Meinshausen, N. and Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Ribeiro, M. T., Singh, S., and Guestrin, C. " 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*, pp. 1135–1144, 2016.

Shapley, L. S. Greedy function approximation: A gradient boosting machine. *Contribution to the Theory of Games*, 2:307–317, 1953. URL https://www.rand.org/content/dam/rand/pubs/research_memoranda/2008/RM670.pdf.

Strumbelj, E. and Kononenko, I. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11:1–18, 01 2010. doi: 10.1145/1756006.1756007.

Sun, X., Liu, Y., Li, J., Zhu, J., Liu, X., and Chen, H. Using cooperative game theory to optimize the feature selection problem. *Neurocomputing*, 97:86–93, 2012. doi: 10.1016/j.neucom.2012.05.001. URL https://doi.org/10.1016/j.neucom.2012.05.001.

Wang, E., Khosravi, P., and Van den Broeck, G. Towards probabilistic sufficient explanations. In *Extending*

Explainable AI Beyond Deep Models and Classifiers Workshop at ICML (XXAI), 2020.

Yu, B. and Kumbier, K. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020.

Zaeri-Amirani, M., Afghah, F., and Mousavi, S. A feature selection method based on shapley value to false alarm reduction in icus a genetic-algorithm approach. volume 2018, pp. 319–323, 07 2018. doi: 10.1109/EMBC.2018.8512266.