Supplementary Materials :
Accurate and robust Shapley Values for explaining predictions and identifying group of Important Variables

Contents

Preprint. Under review.

A Proofs

This section gathers all the proofs of the propositions and claims of the main paper.

2. Coalition and Invariance for Shapley Values

2.1 Invariance under reparametrization for continuous variables

Proposition A.1. *Let f and $\tilde{f} = f \circ \boldsymbol{\varphi}^{(-1)}$ its reparametrization, then we have for all $i \in [\![1, p]\!]$, for all $\boldsymbol{x}, \boldsymbol{u} = \boldsymbol{\varphi}(\boldsymbol{x})$:*

$$\phi_i(f, \boldsymbol{x}) = \phi_i(\tilde{f}, \boldsymbol{\varphi}(\boldsymbol{x})).$$

Proof. It is a direct application of the change of variables formula. If $g(\boldsymbol{x})$ is the joint density of $X_1, \dots, X_p$ (X_i has density g_i), the transformed variable $\boldsymbol{U} = (\varphi_1(X_1), \dots, \varphi_p(X_p))$ has density $\tilde{g}(\boldsymbol{u}) = g(\boldsymbol{\varphi}^{(-1)}\boldsymbol{u}) \times \prod_i |J(\varphi_i^{(-1)})(u_i)|$. With obvious notations, we have

$$\tilde{g}(u_{\bar{S}}|u_S) = \frac{\tilde{g}(u_{\bar{S}}, u_S)}{\tilde{g}_S(u_S)} = g\left(\varphi_{\bar{S}}^{(-1)}(u_{\bar{S}}|\varphi_S^{(-1)}(u_S)\right) \times \prod_{i \in \bar{S}} |J(\varphi_i^{(-1)})(u_i)|.$$

The computation of the reducted predictor is straightforward

$$\begin{aligned} E\left[f(\boldsymbol{X})|\boldsymbol{x}_S\right] &= \int f(\boldsymbol{x}_S, \boldsymbol{x}_{\bar{S}}) g(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S) d\boldsymbol{x}_{\bar{S}} \\ &= \int f(\varphi_S^{(-1)}\left(\varphi_S(\boldsymbol{x}_S)\right), \varphi_{\bar{S}}^{(-1)} \varphi_{\bar{S}}(\boldsymbol{x}_{\bar{S}}))\, g(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S) d\boldsymbol{x}_{\bar{S}} \\ &= \int \tilde{f}(\boldsymbol{u}_S, \boldsymbol{u}_{\bar{S}}) g\left(\varphi_{\bar{S}}^{(-1)}(\boldsymbol{u}_{\bar{S}})|\varphi_S^{(-1)}(\boldsymbol{u}_S)\right) \prod_{i \in \bar{S}} \left|J\varphi^{(-1)}(u_i)\right| d\boldsymbol{u}_{\bar{S}} \\ &= E\left[\tilde{f}(\boldsymbol{U}_S, \boldsymbol{U}_{\bar{S}})|\boldsymbol{U}_S = \boldsymbol{u}_S\right]. \end{aligned}$$

The equality of Shapley Values is then a direct consequence of the equality of reduced predictors.

□

2.2 Invariance for encoded categorical variable

We recall the expression of the SV for 2 variables for all $x \in \mathbb{R}$ and $Y \in \{1, \dots, K\}$. 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 given:

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

Proposition A.2. *For all $x \in \mathcal{X}$, and if $y_{1:K-1} = \mathcal{C}(y)$ then*

$$\left\{ \begin{array}{lcl} \phi_C(\tilde{f}; x, y_{1:K-1}) & = & \phi_Y(f; x, y) \\ \phi_X(\tilde{f}; x, y_{1:K-1}) & = & \phi_X(f; x, y) \end{array} \right. \tag{A.2}$$

Proof. As we consider only doable $(x, y_{1:K-1})$, then $\exists! y \in \{1, \dots, K\}$ such that $\mathcal{C}(y) = y_{1:K-1}$. We have the coalition $C = \{1, \dots, K-1\}$, and number of variables $p = K$, meaning

$$\phi_{\{1,\dots,K-1\}}(\tilde{f}; x, y_{1:K-1}) = \frac{1}{2}\left\{\frac{1}{\binom{1}{0}}\Delta(\tilde{f}; \emptyset, C) + \frac{1}{\binom{1}{1}}\Delta(\tilde{f}; \{X\}, C)\right\}$$

where

$$\begin{aligned} \Delta(\tilde{f}; \emptyset, C) &= E_{\tilde{P}}\left[\tilde{f}(X, Y_{1:K-1})|Y_{1:K-1} = y_{1:K-1}\right] - E_{\tilde{P}}\left[\tilde{f}(X, Y_{1:K-1})|\emptyset\right] \\ &= E_P\left[\tilde{f}\left(X, \varphi\left(Y\right)\right)|Y = y\right] - E_P\left[\tilde{f}\left(X, \varphi\left(Y\right)\right)\right] \\ &= E_P\left[f\left(X, Y\right)|Y = y\right] - E_P\left[f\left(X, Y\right)\right] \end{aligned}$$

Indeed

$$\begin{aligned}
E_{\tilde{P}}\left[\tilde{f}(X, Y_{1:K-1})|Y_{1:K-1}=y_{1:K-1}\right] &= \int \tilde{f}(x, y_{1:K-1})dP(x|y_{1:K-1}) \\
&= \int \tilde{f}(x, y_{1:K-1})\frac{dP(x, y_{1:K-1})}{P(y_{1:K-1})} \\
&= \int \tilde{f}(x, \varphi(y))\frac{dP(x, \varphi(y))}{P(\varphi(y))} \\
&= \int f(x, y)\frac{dP(x, y)}{P(y)}
\end{aligned}$$

In addition,

$$\begin{aligned}
\Delta(\tilde{f}; \{X\}, C) &= E_{\tilde{P}}\left[\tilde{f}(X, Y_{1:K-1})|X=x, Y_{1:K-1}=y_{1:K-1}\right] - E_{\tilde{P}}\left[\tilde{f}(X, Y_{1:K-1})|X=x\right] \\
&= \tilde{f}(x, y_{1:K-1}) - E_P\left[\tilde{f}(X, \varphi(Y))|X=x\right] \\
&= \tilde{f}(x, \varphi(y)) - E_P\left[\tilde{f}(X, \varphi(Y))|X=x\right] \\
&= f(x, y) - E_P\left[f(X, y)|X=x\right]
\end{aligned}$$

$$\begin{aligned}
\phi_{\{1,\ldots,K-1\}}(\tilde{f}; x, y_{1:K-1}) &= \frac{1}{2}\left(E_P\left[f\left(X, Y\right)|Y=y\right] - E_P\left[f\left(X, Y\right)\right]\right) \\
&+ \frac{1}{2}\left(f(x, y) - E_P\left[f(X, y)|X=x\right]\right)
\end{aligned}$$

We can recognize that we have exactly $\phi_{\{1,\ldots,K-1\}}(\tilde{f}; x, y_{1:K-1}) = \phi_Y(f; x, y)$. From Equation 2.1, we derive that $\phi_X(\tilde{f}; x, y_{1:K-1}) = \phi_X(f; x, y)$. ☐

Proposition A.3. *If $X \sim \mathcal{N}(\mu, \Sigma)$, then $X_{\bar{S}}|X_S = x_S$ is also multivariate gaussian with mean $\mu_{\bar{S}|S}$ and covariance matrix $\Sigma_{\bar{S}|S}$ equal:*

$$\mu_{\bar{S}|S} = \mu_{\bar{S}} + \Sigma_{\bar{S},S}\Sigma_{S,S}^{-1}(x_S - \mu_S) \text{ and } \Sigma_{\bar{S}|S} = \Sigma_{\bar{S}\bar{S}} - \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}\Sigma_{S,\bar{S}}$$

B Fast Algorithm for the computation of Shapley Values with the Leaf estimator

In section 3.2. of the main paper, we have introduced a plug-in estimator of the conditional expectation

$$f_S(\boldsymbol{x}_S) = \sum_{m=1}^{M} f_m P_X\left(L_m\left|\boldsymbol{X}_S = \boldsymbol{x}_S\right.\right),$$

that is based on an approximation of the conditional expectation by event $\{\boldsymbol{X}_S = \boldsymbol{x}_S\}$ by a conditional expectation based on event $\{\boldsymbol{X}_S \in L_m^S\}$. For sake of notational simplicity, we write simply $L_m^S = L_m^S(\boldsymbol{x})$ and we remove the dependence on $\boldsymbol{x}$.

Thanks to this approximation, we can propose a straightforward estimate based on empirical frequencies, and we focus here on the computational efficiency offered by this approximation. It is well-known that the complexity of the computation of a Shapley value is exponential as we need to compute 2^p different coalitions for each observation $\boldsymbol{x}$. We show below that the complexity can be made much lower, as we derive an algorithm with complexity exponential in the depth of the tree instead of being exponential in the total number of variable p. This is very interesting as the depth of the tree is rarely above 10 in practice, while p can be very large (different order of magnitudes). We want to compute the predictor

$$\tilde{f}_S(\boldsymbol{x}_S) = \sum_{m=1}^{M} f_m P_X\left(L_m|\boldsymbol{X}_S \in L_m^S(\boldsymbol{x})\right)$$

(or its estimated version equal to $\hat{f}_S^{(Leaf)}(\boldsymbol{x}_S)$) that can be used for defining a new cooperative game based on the value function

$$S \mapsto \tilde{v}(f, S) \triangleq \tilde{f}_S(\boldsymbol{x}_S).$$

For any coalition C, our estimate of the Shapley value $\phi_C(\boldsymbol{x})$ is the Shapley value of the cooperative game $\tilde{v}(f, S)$ defined as

$$\tilde{\phi}_C(\tilde{f}; \boldsymbol{x}) = \frac{1}{p - |C| + 1} \sum_{k=0}^{p-|C|} \frac{1}{\binom{p-|C|}{k}} \sum_{S \in \mathcal{S}_k(C)} \left(\tilde{f}_{S \cup C}(\boldsymbol{x}_{S \cup C}) - \tilde{f}_S(\boldsymbol{x}_S) \right). \tag{B.1}$$

We show in the next proposition that the game $\tilde{v}$ can be split into the sum of smaller games (we consider only $C = \{i\}$ in the proposition, but it remains true for any coalition C).

Proposition B.1. *Let $f(\boldsymbol{x}) = \sum_{m=1}^{M} f_m \mathbb{1}_{L_m}(\boldsymbol{x})$ be a tree based on p variables $\boldsymbol{X} = (X_1, \dots, X_p)$. We introduce for each leaf L_m the set of variables $S_m = \{X_{N_1}, X_{N_2}, \dots, X_{N_{d_m}}\}$ used in the tree path defining the leaf L_m. For any variable X_i, the SV $\phi_i(\tilde{f}, \boldsymbol{x})$ can be decomposed into the sum of M cooperative games defined on each leaf L_m, and we have*

$$\tilde{\phi}_i(\tilde{f}, \boldsymbol{x}) = \sum_{m=1}^{M} \tilde{\phi}_i^m(\tilde{f}, \boldsymbol{x})$$

where $\tilde{\phi}_i^m(\tilde{f}, \boldsymbol{x})$ is a reweighted version of the Shapley Value of the cooperative game with value function

$$\tilde{v}(\tilde{f}, , S) = P_X\left(L_m | \boldsymbol{X}_S \in L_m^S(\boldsymbol{x})\right)$$

.

Proof. By definition, we have for a single variable

$$\begin{aligned}
\tilde{\phi}_i(\boldsymbol{x}) &= \frac{1}{p} \sum_{S \subseteq [p] \setminus \{i\}} \binom{p-1}{|S|}^{-1} \left(\tilde{f}_{S \cup i}(\boldsymbol{x}_{S \cup i}) - \tilde{f}_S(\boldsymbol{x}_S) \right) \\
&= \frac{1}{p} \sum_{S \subseteq [p] \setminus \{i\}} \binom{p-1}{|S|}^{-1} \left(\sum_{m=1}^{M} f_m \Big[P(L_m | \boldsymbol{X}_{S \cup i} \in L_m^{S \cup i}) - P(L_m | \boldsymbol{X}_S \in L_m^S \Big] \right) \\
&= \frac{1}{p} \sum_{m=1}^{M} \sum_{S' \subseteq S_m \setminus \{i\}} \left[\binom{p-1}{|S'|}^{-1} f_m \Big[P(L_m | \boldsymbol{X}_{S' \cup i} \in L_m^{S' \cup i}) - P(L_m | \boldsymbol{X}_{S'} \in L_m^{S'}) \Big] \right. \\
&\quad \left. + \sum_{Z \neq \emptyset, Z \subseteq \overline{S_m \cup i}} \binom{p-1}{|Z| + |S'|}^{-1} f_m \Big[P(L_m | \boldsymbol{X}_{S' \cup Z \cup i} \in L_m^{S' \cup Z \cup i}) - P(L_m | \boldsymbol{X}_{S' \cup Z} \in L_m^{S_m \cup Z}) \Big] \right]
\end{aligned}$$

However, if $Z \subseteq \bar{S}_m, S \subseteq S_m$:

$$P_X(L_m | \boldsymbol{X}_{Z \cup S} \in L_m^{Z \cup S}) = P_X(L_m | \boldsymbol{X}_S \in L_m^S). \tag{B.2}$$

We shall remark that the identity of eq. (B.2) is not true anymore if we consider the conditional probability $\boldsymbol{X}_S = \boldsymbol{x}_x$). Therefore, the SV $\tilde{\phi}_i(\boldsymbol{x})$ can be rewrite as:

$$\begin{aligned}
\tilde{\phi}_i(\boldsymbol{x}) &= \frac{1}{p} \sum_{m=1}^{M} \sum_{S' \subseteq S_m \setminus \{i\}} \left[\binom{p-1}{|S'|}^{-1} f_m \Big[P(L_m | \boldsymbol{X}_{S' \cup i} \in L_m^{S' \cup i}) - P(L_m | \boldsymbol{X}_{S'} \in L_m^{S'}) \Big] \right. \\
&\quad \left. + \sum_{Z \neq \emptyset, Z \subseteq \overline{S_m \cup i}} \binom{p-1}{|Z| + |S'|}^{-1} f_m \Big[P(L_m | \boldsymbol{X}_{S' \cup i} \in L_m^{S' \cup i}) - P(L_m | \boldsymbol{X}_{S'} \in L_m^{S'}) \Big] \right] \\
&= \frac{1}{p} \sum_{m=1}^{M} \sum_{S' \subseteq S_m \setminus \{i\}} \left[\binom{p-1}{|S'|}^{-1} + \sum_{Z \neq \emptyset, Z \subseteq \overline{S_m \cup i}} \binom{p-1}{|Z| + |S'|}^{-1} \right] f_m \Big[P(L_m | \boldsymbol{X}_{S' \cup i} \in L_m^{S' \cup i}) - p(L_m | \boldsymbol{X}_{S'} \in L_m^{S'}) \Big] \Big] \\
&\triangleq \sum_{m=1}^{M} \tilde{\phi}_i^m(\boldsymbol{x})
\end{aligned}$$

Each term $\tilde{\phi}_i^m(\boldsymbol{x})$ introduced in the sum is a re-weighted Shapley value of the cooperative game defined in each leaf L_m with the variables S_m and associated to the value function $P(L_m|\boldsymbol{X}_{S'} \in L_m^{S'})$, for all coalition $S' \subseteq S_m$. As shown above, it is not the standard Shapley Value as we consider p variables in the normalisation (and not $|S_m|$, and we need to take into account the additional contributions of the coalitions that partly overlaps S_m and $\bar{S}_m$: this adds the additional constant

$$\sum_{Z \neq \emptyset, Z \subseteq \overline{S_m \cup i}} \binom{p-1}{|Z| + |S'|}^{-1}.$$

$\square$

A straightforward algorithm for computing SV has a complexity $\mathcal{O}(p \times \texttt{tree-depth} \times 2^p)$ (called *Brute Force Algorithm*): we have p variables, 2^p groups of variables to consider each time, and we need to go down into the tree. Instead, we suggest to compute SV leaf by leaf thanks to equation (eq.B.1).In that case, the computation of the SV for the p variables is done by summing over M games (leafs), each of them having a number of variables $|S_m|$ lower than `tree-depth`. Consequently, the complexity is $\mathcal{O}(p \times M \times 2^{\texttt{tree-depth}})$ in worst cases.
The *Multi-Games algorithm* improves dramatically the computational complexity as `tree-depth` is often much lower than p. Moreover, the algorithm is linear in the number of observations where we want to compute the SV.

The algorithm is describes below, we use the following notation $\mathbb{1}_{L_m^{\emptyset}}(\boldsymbol{x}_{\emptyset}) = 1$.

Algorithm 1: *Multi-Games Algorithm* - Compute SV in the worst case in $\mathcal{O}(p \times M \times 2^{\texttt{tree-depth}})$

```
Inputs: $\boldsymbol{x}, f(\boldsymbol{x}) = \sum_{m=0}^{M} f_m \mathbb{1}_{L_m}(\boldsymbol{x})$;
$p = length(\boldsymbol{x})$;
$\phi = \text{zeros}(p)$;
for $m = 1$ to $M$ do
    for $i$ in $[p]$ do
        if $i$ not in $S_m$ then
            continue ;                                /* skip to next variable */
        end
        for $S \subseteq S_m$ do
            $\phi[i] += \left(\binom{p-1}{|S|}^{-1} + \sum_{k=1}^{p-|S_m|} \binom{p-|S_m|}{k} \binom{p-1}{k+|S|}^{-1}\right) \times$
            $\left(\mathbb{1}_{L_m^{S\cup i}}(\boldsymbol{x}_{S\cup i}) \hat{P}_X^{(Leaf)}\left(L_m \,\middle|\, \boldsymbol{X}_S \in L_m^{S\cup i}\right) - \mathbb{1}_{L_m^S}(\boldsymbol{x}_S) \hat{P}_X^{(Leaf)}\left(L_m \,\middle|\, \boldsymbol{X}_S \in L_m^S\right)\right)$
        end
    end
end
return $\phi$
```

Remark B.1. The algorithm can be vectorized in order to compute SV of several observations at the same time.

C Same Decision Probability with tree-based model

Same Decision Probability

Our methodology for identifying the most important features is based on the Same Decision Probability (SDP) criterion, introduced in [2]. In particular, we derive an efficient way to approximate the SDP for tree-based classifier.

Definition C.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 C.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 [3]. Consequently, approximate criterion based on expectations instead of probabilities have been introduced see [7]. 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 C.1. *Let f a probabilistic predictor and its binary classifier C with threshold T, $Q_{S,x}$ 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:*

$$SDP_S\left(f; \boldsymbol{x}\right) = \frac{\mathbb{E}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S=\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})\right] - \mathbb{E}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S=\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})|f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) < T\right]}{\mathbb{E}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S=\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})|f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) \geq T\right] - \mathbb{E}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S=\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})|f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) < T\right]} \tag{C.1}$$

Proof. First note that $SDP_S\left(f; \boldsymbol{x}\right) = \mathbb{P}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S=\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) \geq T\right]$

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S=\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})\right] = {} & \mathbb{E}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S\,=\,\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})|\, f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) < T\right] \mathbb{P}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S=\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) < T\right] \\ & + \mathbb{E}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S=\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})|\, f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) \geq T\right] \mathbb{P}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S=\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) \geq T\right]\end{aligned}$$

Rearranging the terms leads to equation C.1 □

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 2 and defines the function *returnSubsets($\boldsymbol{x}$, size)* that returns all subsets of length *size* of $\boldsymbol{x}$.

We already know how to estimate $E[f(\boldsymbol{x}_s, \boldsymbol{X}_{\bar{S}})|\boldsymbol{X}_S = \boldsymbol{x}_s]$. Therefore, we use the same idea to estimate $\mathbb{E}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S=\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})|f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) < T\right]$. We estimate each probability of the Leaf estimator of the reduced predictor with the condition $\{f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) < T\}$ as

$$\hat{P}^{(Leaf)}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S=\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})|f(\boldsymbol{x}_s, \boldsymbol{X}_{\bar{S}}) < T\right] = \frac{N(L_m, f < T)}{N(L^S_m, f < T)},$$

and for the Discrete estimator, we use

$$\hat{P}^{(D)}_{\boldsymbol{X}_{\bar{S}}|\boldsymbol{X}_S=\boldsymbol{x}_S}\left[f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}})|f(\boldsymbol{x}_s, \boldsymbol{X}_{\bar{S}}) < T\right] = \frac{N(L_m, \boldsymbol{x}_S, f < T)}{N(\boldsymbol{x}_S, f < T)},$$

where

- $N(\boldsymbol{x}_S, f < T)$: number of observations such that $\boldsymbol{X}_S = \boldsymbol{x}_S$ and $f(x) < T$ (across all the leaves of the tree)
- $N(L_m, \boldsymbol{x}_S, f < T)$: number of observations in leaf L_m that meet the condition $\boldsymbol{X}_S = \boldsymbol{x}_S$ and $f(x) < T$
- $N(L_m, f < T)$: number of observations in the leaf L_m that meet the condition $f(x) < T$,
- $N(L_m^S, f < T)$: number of observations that meet the conditions $\boldsymbol{x}_S \in L_m^S$ and $f(x) < T$ across all the leaves of the tree.

The case with the condition $\{f(\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}) \geq T\}$ is similar.

Algorithm 2: Find Sufficient Coalition $S^\star_\pi(\boldsymbol{x})$

```
Inputs: $\boldsymbol{x}, \pi$;
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 $\geq$ bestSdp then
            bestSdp = sdp;
            $S^\star_\pi = S$;
            find = True;
        end
    end
end
if find == True then
    return $S^\star_\pi$
end
```

D Link between the Algorithm 1 (TreeSHAP with path-dependent) and $\hat{f}^{SHAP}$

In section 3.1, we have said that the recursive algorithm 1 introduced in [6] and shows in figure 2 assumes that the probabilities can be factored with the decision tree as:

$$P^{SHAP}_X\left(\prod_{k=1}^{d_m} I_{N_k}|\boldsymbol{X}_S = \boldsymbol{x}_S\right) = \prod_{i=2|N_i \notin S}^{d_m} P(X_{N_i} \in I_{N_i}|X_{N_{i-1}} \in I_{N_{i-1}}) \times \delta_S(N_1) \quad \text{(D.1)}$$

with $\delta_S(N_1) = P(X_{N_1} \in I_{N_1})$ if $N_1 \notin S$, and 1 otherwise.

To show the link between between $\hat{f}^{SHAP}$ and the Algorithm 1, let choose an observation $x_{ref} = [2, 3, 0.5, -1]$ and compute $E[f(\boldsymbol{X})|\ X_0 = 2, X_2 = 0.5]$ where f is the tree in figure 1. The condition $X_0 = 2, X_2 = 0.5$ is compatible with the leaves $6, 7, 11, 13, 14$, we denote $f_6, f_7, f_{11}, f_{13}, f_{14}$ the value of each leaf respectively. The output of the algorithm 1 (described in figure 2) is on step 4, 5 of the table D.1 below and its corresponds to $\hat{f}^{(SHAP)}_{\{0,2\}}(x_{ref})$.

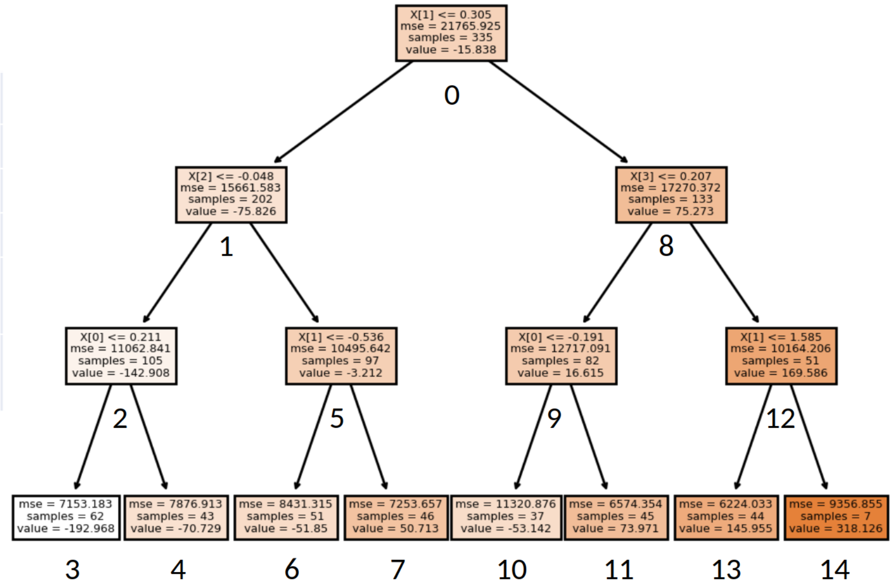
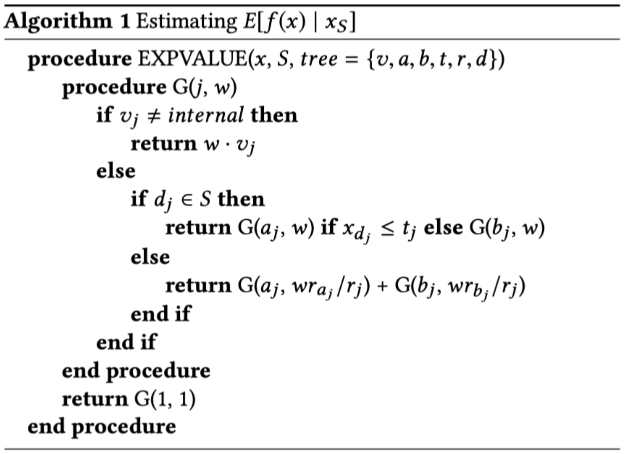


Figure 1: A simple decision tree used to illustrate the link between $\hat{f}^{(SHAP)}$ and Algorithm 1 in [6] (Tree SHAP) with numbered leaves.



```
Algorithm 1 Estimating $E[f(x) \mid x_S]$
  procedure EXPVALUE($x$, $S$, $tree = \{v, a, b, t, r, d\}$)
      procedure G($j$, $w$)
          if $v_j \neq internal$ then
              return $w \cdot v_j$
          else
              if $d_j \in S$ then
                  return G($a_j$, $w$) if $x_{d_j} \leq t_j$ else G($b_j$, $w$)
              else
                  return G($a_j$, $wr_{a_j}/r_j$) + G($b_j$, $wr_{b_j}/r_j$)
              end if
          end if
      end procedure
      return G(1, 1)
  end procedure
```

Figure 2: Algorithm 1 in [6] (Tree SHAP)

$$
\begin{aligned}
\hat{f}^{(SHAP)}_{\{0,2\}}(x_{ref}) &= P(X_1 \leq 0.305)P(X_2 > -0.048|X_1 \leq 0.305) * P(X_1 \leq -0.536|X_2 > -0.048)f_6 \\
&+ P(X_1 \leq 0.305)P(X_2 > -0.048|X_1 \leq 0.305) * P(X_1 > -0.536|X_2 > -0.048)f_7 \\
&+ P(X_1 > 0.305)P(X_3 \leq 0.207|X_1 > 0.305) * P(X_0 > -0.191|X_3 \leq 0.207)f_{11} \\
&+ P(X_1 > 0.305)P(X_3 > 0.207|X_1 > 0.305) * P(X_1 \leq 1.585|X_3 > 0.207)f_{13} \\
&+ P(X_1 > 0.305)P(X_3 > 0.207|X_1 > 0.305) * P(X_1 > 1.585|X_3 > 0.207)f_{14} \\
&= (202/335) * 1 * (51/97) * (-51.85) + (202/335) * 1 * (46/97) * (50.716) \\
&+ (133/335) * (82/133) * 1 * (73.971) + (133/335) * (51/133) * (44/51) * (145.955) \\
&+ (133/335) * (51/133) * (7/51) * (318.125) \\
&= 41.98
\end{aligned}
$$

Step	Calculus
0	G(0, 1)
1	G(1, 202/335) + G(8, 133/335)
2	G(5, 202/335) + G(9, 88/335) + G(12, 51/335)
3	G(6,(202/335)*(51/97)) + G(7,(202/335)*(46/97)) + G(11,82/335) + G(13,44/335) + G(14,7/335)
4	-(202/335)*(51/97)*51,85 + (202/335)*(46/97)*50,713 + (82/335)*73,971 + (44/335)*145,955 + (7/335)*318,126
5	= 41.98

Table 1: Steps of Algorithm 1 in [6] (TreeSHAP) for the computation of $E[f(\boldsymbol{X})|\ X_0 = 2, X_2 = 0.5]$ with $x_{ref} = [2, 3, 0.5, -1]$ and the tree in figure 1.

E Focus on influential variables in a linear model

Proposition E.1. *Let assumes that we have $\boldsymbol{X} \in \mathbb{R}^p$, $\boldsymbol{X} \in \mathcal{N}(0, I)$ and a linear predictor f defined as:*

$$f(\boldsymbol{X}) = (a_1\boldsymbol{X}_1 + a_2\boldsymbol{X}_2)\mathbb{1}_{\boldsymbol{X}_5 \leq 0} + (a_3\boldsymbol{X}_3 + a_4\boldsymbol{X}_4)\mathbb{1}_{\boldsymbol{X}_5 > 0}. \tag{E.1}$$

Even if we choose an observation $\boldsymbol{x}$ such that $x_5 \leq 0$ and the predictor only uses $\boldsymbol{X}_1, \boldsymbol{X}_2$, the SV of ϕ_3, ϕ_4 is not necessarily zero.

Proof.

$$\phi_3 = \frac{1}{p} \sum_{S \subseteq [p] \setminus \{3\}} \binom{p-1}{|S|}^{-1} \Big(f_{S\cup 3}(\boldsymbol{x}_{S\cup 3}) - f_S(\boldsymbol{x}_S) \Big) \tag{E.2}$$

$$= \frac{1}{p} \sum_{S \subseteq [p] \setminus \{3,5\}} \binom{p-1}{|S|}^{-1} \Big(f_{S\cup 3}(\boldsymbol{x}_{S\cup 3}) - f_S(\boldsymbol{x}_S) \Big) + \frac{1}{p} \sum_{S \subseteq [p] \setminus \{3,5\}} \binom{p-1}{|S|+1}^{-1} \Big(f_{S\cup\{3,5\}}(\boldsymbol{x}_{S\cup\{3,5\}}) - f_{S\cup 5}(\boldsymbol{x}_{S\cup 5}) \Big) \tag{E.3}$$

The second term is zero. Indeed, $\forall S \subseteq [p] \setminus \{3, 5\}$

$$f_{S\cup\{3,5\}}(\boldsymbol{x}_{S\cup\{3,5\}}) - f_{S\cup 5}(\boldsymbol{x}_{S\cup 5}) = 0.$$

Because, if we condition on the event $\{\boldsymbol{X}_5 = \boldsymbol{x}_5\}$ with $x_5 \leq 0$

$$\begin{aligned} f_{S\cup\{3,5\}}(\boldsymbol{x}_{S\cup\{3,5\}}) &= E\left[(a_1\boldsymbol{X}_1 + a_2\boldsymbol{X}_2)\mathbb{1}_{\boldsymbol{X}_5 \leq 0} + (a_3\boldsymbol{X}_3 + a_4\boldsymbol{X}_4)\mathbb{1}_{\boldsymbol{X}_5 > 0} \mid \boldsymbol{X}_{S\cup\{3,5\}} = \boldsymbol{x}_{S\cup\{3,5\}}\right] \\ &= E\left[(a_1\boldsymbol{X}_1 + a_2\boldsymbol{X}_2)\mathbb{1}_{\boldsymbol{X}_5 \leq 0} \mid \boldsymbol{X}_{S\cup\{3,5\}} = \boldsymbol{x}_{S\cup\{3,5\}}\right] && \text{because } x_5 \leq 0 \\ &= E\left[(a_1\boldsymbol{X}_1 + a_2\boldsymbol{X}_2) \mid \boldsymbol{X}_{S\cup 5} = \boldsymbol{x}_{S\cup 5}\right] && \text{independent of } \boldsymbol{X}_3 \\ &= f_{S\cup 5}(\boldsymbol{x}_{S\cup 5}). \end{aligned}$$

□

The first term of 3.3 is the classic marginal contribution of SV in linear model. $\forall S \subseteq [p] \setminus \{3, 5\}$,

$$\begin{aligned} f_{S\cup 3}(\boldsymbol{x}_{S\cup 3}) &= E\left[a_1\boldsymbol{X}_1 + a_2\boldsymbol{X}_2 \mid \boldsymbol{X}_{S\cup 3} = \boldsymbol{x}_{S\cup 3}\right] P(\boldsymbol{X}_5 \leq 0 \mid \boldsymbol{X}_{S\cup 3} = \boldsymbol{x}_{S\cup 3}) \\ &+ E\left[a_3\boldsymbol{X}_3 + a_4\boldsymbol{X}_4 \mid \boldsymbol{X}_{S\cup 3} = \boldsymbol{x}_{S\cup 3}\right] P(\boldsymbol{X}_5 > 0 | \boldsymbol{X}_{S\cup 3} = \boldsymbol{x}_{S\cup 3}) \\ &= E\left[a_1\boldsymbol{X}_1 + a_2\boldsymbol{X}_2 | \boldsymbol{X}_S = \boldsymbol{x}_S\right] P(\boldsymbol{X}_5 \leq 0) + (E\left[a_2\boldsymbol{X}_2 | \boldsymbol{X}_S = \boldsymbol{x}_S\right] + a_3\boldsymbol{x}_3) P(\boldsymbol{X}_5 > 0) \\ &= f_S(\boldsymbol{x}_S) + P(\boldsymbol{X}_5 > 0)\Big(a_3(\boldsymbol{x}_3 - E[\boldsymbol{X}_3])\Big) \end{aligned}$$

Therefore,

$$\begin{aligned} \phi_3 &= \frac{1}{p} \sum_{S \subseteq [p] \setminus \{3,5\}} \binom{p-1}{|S|}^{-1} P(\boldsymbol{X}_5 > 0)\Big(a_3(\boldsymbol{x}_3 - E[\boldsymbol{X}_3])\Big) \\ &= K\Big(a_3(\boldsymbol{x}_3 - E[\boldsymbol{X}_3])\Big) && K \text{ is a constant} \end{aligned}$$

The computation of ϕ_4 is symmetric.

F Additional examples

F.1 Impact of quantile discretization

The table F.2 (borrowed from [1]) shows the impact of discretization on the performance of a Random Forest on UCI datasets.

Dataset	**Breiman's RF**	**q=2**	**q=5**	**q=10**	**q=20**
Authentification	0.0002	0.08	0.002	0.0005	0.0004
Diabetes	0.17	0.23	0.18	0.18	0.18
Haberman	0.32	0.35	0.30	0.32	0.30
Heart Statlog	0.10	0.10	0.10	0.10	0.10
Hepatitis	0.13	0.15	0.14	0.14	0.13
Ionosphere	0.02	0.07	0.03	0.02	0.02
Liver Disorders	0.23	0.32	0.27	0.25	0.24
Sonar	0.07	0.09	0.07	0.07	0.07
Spambase	0.01	0.14	0.03	0.02	0.01
Titanic	0.13	0.15	0.14	0.14	0.13
Wilt	0.007	0.15	0.03	0.02	0.02

Table 2: Accuracy, measured by 1-AUC on UCI datasets, for two algorithms: Breiman's random forests and random forests with splits limited to q-quantiles, for $q \in \{2, 5, 10, 20\}$. Table 5 in [1]

In the table F.3 below, we compare the different estimators (Tree SHAP, Leaf and Discrete) after the discretization of continuous variables. The model and the data used are described in Section A.2. We see that the Discrete estimator is much more sensitive to out-of-distribution than Leaf. We can also observed that the different estimators have significant variance. Indeed, a model trained on discrete variables tends to have poorly filled leaves.

Datasets	D$^{(Explain)}$		Uniform	
Metrics	MSE	TPR	MSE	TPR
Tree SHAP	2.92	55% (49%)	2.91	74% (43%)
Leaf	0.98	74% (43%)	2.64	75% (43%)
Discrete	0.20	95% (22%)	2.99	65.79% (47%)

Table 3: Metrics of the different estimators after discretization

F.2 The differences between Coalition and sum on Census Data

We use UCI Adult Census Dataset [4]. 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 Global SV by taking the coalition or sum of the modalities over N=5000 observations. Global SV are defined as:

$$I_j = \sum_{i=0}^{N} |\phi_j^{(i)}|$$

In figure 3, we see differences between the global 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.

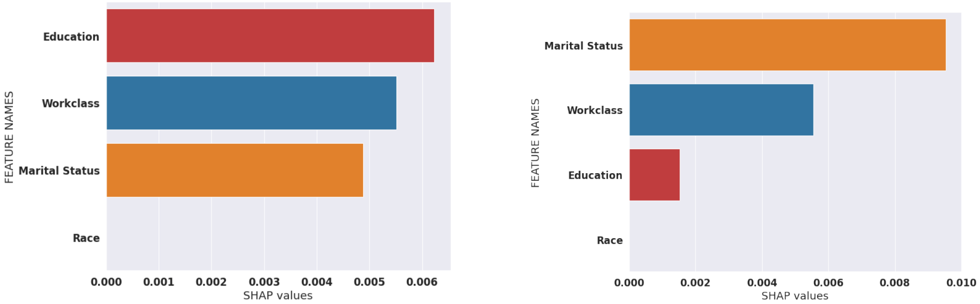


Figure 3: Difference between the global absolute value of SV: sum (left) vs coalition (right) of dummies

We also compute the proportion of order inversion over N=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.

F.3 SDP and Active SV analysis on Lucas Data

We use an accurate decision tree trained on LUCAS [5], a dataset generated by causal Bayesian networks with 12 binary variables. The graph is drawn in figure 4 and we provide the probability table in figure 5

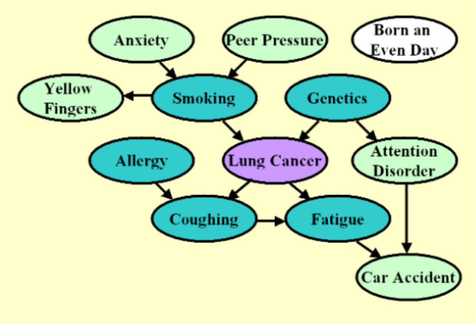


Figure 4: Bayesian network that represents the causal relationships between variables

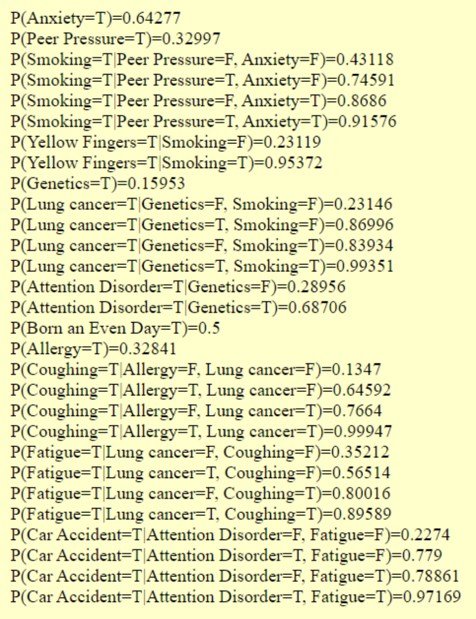

P(Anxiety=T)=0.64277
P(Peer Pressure=T)=0.32997
P(Smoking=T|Peer Pressure=F, Anxiety=F)=0.43118
P(Smoking=T|Peer Pressure=T, Anxiety=F)=0.74591
P(Smoking=T|Peer Pressure=F, Anxiety=T)=0.8686
P(Smoking=T|Peer Pressure=T, Anxiety=T)=0.91576
P(Yellow Fingers=T|Smoking=F)=0.23119
P(Yellow Fingers=T|Smoking=T)=0.95372
P(Genetics=T)=0.15953
P(Lung cancer=T|Genetics=F, Smoking=F)=0.23146
P(Lung cancer=T|Genetics=T, Smoking=F)=0.86996
P(Lung cancer=T|Genetics=F, Smoking=T)=0.83934
P(Lung cancer=T|Genetics=T, Smoking=T)=0.99351
P(Attention Disorder=T|Genetics=F)=0.28956
P(Attention Disorder=T|Genetics=T)=0.68706
P(Born an Even Day=T)=0.5
P(Allergy=T)=0.32841
P(Coughing=T|Allergy=F, Lung cancer=F)=0.1347
P(Coughing=T|Allergy=T, Lung cancer=F)=0.64592
P(Coughing=T|Allergy=F, Lung cancer=T)=0.7664
P(Coughing=T|Allergy=T, Lung cancer=T)=0.99947
P(Fatigue=T|Lung cancer=F, Coughing=F)=0.35212
P(Fatigue=T|Lung cancer=T, Coughing=F)=0.56514
P(Fatigue=T|Lung cancer=F, Coughing=T)=0.80016
P(Fatigue=T|Lung cancer=T, Coughing=T)=0.89589
P(Car Accident=T|Attention Disorder=F, Fatigue=F)=0.2274
P(Car Accident=T|Attention Disorder=T, Fatigue=F)=0.779
P(Car Accident=T|Attention Disorder=F, Fatigue=T)=0.78861
P(Car Accident=T|Attention Disorder=T, Fatigue=T)=0.97169

Figure 5: Probabilities table used to generate Data

We want to explain an observation with a well-defined ground truth. The observation choose has all the variables at false except Born an even day and Car accident. 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 table F.4.

We have also computed the Active SV and the standard SV. The figure 6 shows that the Active SV are indeed sparse giving importance to the local active SV while standard SV found that Fatigue, Yellow Fingers, Anxiety is more important than Genetic for this observation.

Active and Null coalition	SDP
$S^\star_\pi(x)$ = [Smoking, Genetics, Coughing]	0.96
$N_\pi(x)$ = [Yellow Fingers, Anxiety, Peer Pressure, Attention Disorder, Born an Even Day, Car Accident, Fatigue, Allergy]	0.77

Table 4: The Sufficient coalition found with $\pi = 0.9$

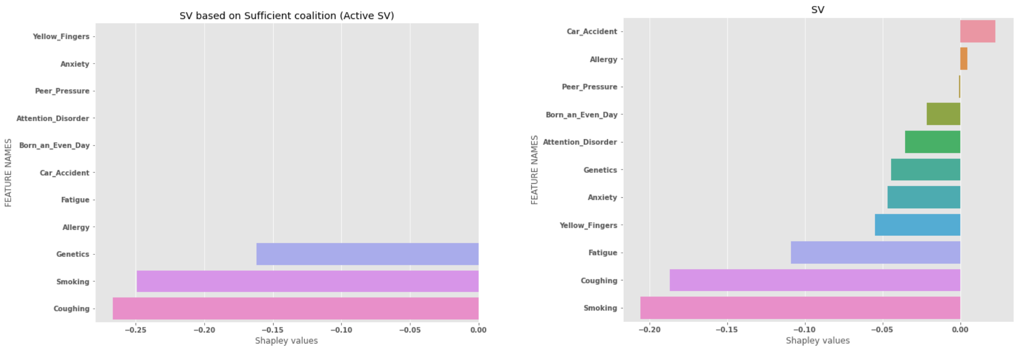


Figure 6: Left figure: SV $\phi^\star_i$ computed with the Sufficient Coalition given in figure 6. Right figure: SV ϕ_i computed with all the variables.

G Individual Shapley values for indicator variables in Dummy Encoding

We give some partial results for the Shapley Values of the modalities $Y = k$, based on the dummy encoding considered in section 2. Indeed equation 2.4 introduces $\phi_k(\tilde{f}, x, y_{1:K-1})$, and proposition 2.1 claims that their sum is different in all generality of the SV of Y. In this section, we give a deeper insight into these values and show that are related multiple comparisons between modalities.

We compute the Shapley Value at point $(x, y = i) = (x, 0, 0, \dots, 1, \dots, 0) = (x, \mathcal{C}(y))$: for ease of notation, we set $Y_0 = X$, and we compute also the Shapley values $\phi_k(\tilde{f}; x, y_{1:K-1})$ for $k = 1, \dots, K-1$. We recall that we need to compute

$$\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{\binom{K-1}{k}}\sum_{\substack{Z\subseteq [\![1..K]\!]\setminus i\\ |Z|=k}}\Delta(\tilde{f}; Z, i).$$

where Δ denotes the difference between the value function evaluated at $Z \cup \{i\}$ and Z. If we examine the terms $\Delta(\tilde{f}; Z, i)$, the computation needs to take into account if $X = Y_0$ is part of the conditioning variable of not. Indeed, we have for each $k \geq 1$,

$$\sum_{\substack{Z\subseteq [\![0..K-1]\!]\setminus i\\ |Z|=k}}\Delta(\tilde{f}; Z, i) = \sum_{\substack{Z\subseteq [\![1..K-1]\!]\setminus i\\ |Z|=k}}\Delta(\tilde{f}; Z, i) + \sum_{\substack{Z'\subseteq [\![1..K-1]\!]\setminus i\\ |Z'|=k-1}}\Delta(\tilde{f}; Z'\cup\{0\}, i). \tag{G.1}$$

We start by computing the first term in the right hand side, and it involves only the dummies, and not the quantitative variable.

Proposition G.1 (Computation of Contributions in Shapley without X)**.** *We compute the Shapley values of the variable Y_i, when we have the observations $(x, y_{1:K-1}) = (x, \mathcal{C}(i))$ for $i \in \{1, \dots, K\}$. We consider any $Z' \subseteq [\![1..K-1]\!]\setminus i$, with $|Z'| = k \geq 1$ and $Z' = \{j_1, \dots, j_k\}$. In that case,*

$$\Delta(\tilde{f}; Z, i) = E_P\left[f(X, Y)|Y = i\right] - E_P\left[f(X, Y)|Y \notin \{j_1, \dots, j_k\}\right] \tag{G.2}$$

Proof. We have $Y_i = 1 \Leftrightarrow Y = i$, and for $Z' \subseteq [\![1..K-1]\!] \setminus \{0, i\}$, we consider $Z' = \{j_1, \dots, j_k\}$, with $1 \leq j_1 < \cdots < j_k \leq K-1$,

$$\begin{aligned} E_{\tilde{P}}\left[\tilde{f}(Y_0, Y_{1:K-1})|Y_{j_1} = 0, \dots, Y_{j_k} = 0, Y_i = 1\right] &= E_{\tilde{P}}\left[\tilde{f}(Y_0, Y_{1:K-1})|Y_i = 1\right] \\ &= E_{\tilde{P}}\left[\tilde{f}(Y_0, \mathcal{C}(Y))|Y_i = 1\right] \\ &= E_P\left[f(Y_0, Y)|Y = i\right] \end{aligned}$$

because for all $j_1, \dots, j_{k-1} \neq i$, we have $\{Y_{j_1} = 0, \dots, Y_{j_k} = 0, Y_i = 1\} = \{Y_i = 1\}$.
Moreover,

$$E_{\tilde{P}}\left[\tilde{f}(Y_0, Y_{1:K-1})|Y_{j_1} = 0, \dots, Y_{j_k} = 0\right] = E_P\left[\tilde{f}(Y_0, \mathcal{C}(Y))|Y \neq j_1, \dots, j_k\right]$$

Hence for $Z \subseteq [\![1..K-1]\!] \setminus i$, we have

$$\Delta(\tilde{f}; Z, i) = E_P\left[f(X, Y)|Y = i\right] - E_P\left[f(X, Y)|Y \notin \{j_1, \dots, j_k\}\right].$$

$\square$

The second term of the right hand side is given below.

Proposition G.2 (Computation of Contributions in Shapley with X)**.** *We compute the Shapley values only for the variable Y_i , when we have the observations doable $(x, y_{1:K-1}) = (x, \mathcal{C}(i))$ for $i \in \{1, \dots, K\}$. We consider any $Z' \subseteq [\![1..K-1]\!] \setminus i$, with $|Z'| = k - 1 \geq 1$, and $Z' = \{j_1, \dots, j_{k-1}\}$. In that case,*

$$\Delta(\tilde{f}; Z' \cup \{0\}, i) = E_P\left[f(X, Y)|X = x, Y = i\right] - E_P\left[f(X, Y)|X, Y \notin \{j_1, \dots, j_{k-1}\}\right] \quad \text{(G.3)}$$

Proof. We assume that we have a subset $|Z'| = k - 1$, such that $Z' \subseteq [\![1..K-1]\!] \setminus i$. This means that $Z' = \{j_1, \dots, j_{k-1}\}$, with $1 \leq j_1, \dots, j_{k-1} \leq K-1$. We

$$\begin{aligned} E_{\tilde{P}}\left[\tilde{f}(Y_0, Y_{1:K-1})|Y_0 = x, Y_{j_1} = 0, \dots, Y_{j_{k-1}} = 0, Y_i = 1\right] &= E_{\tilde{P}}\left[\tilde{f}(Y_0, Y_{1:K-1})|Y_0 = x, Y_i = 1\right] \\ &= E_P\left[\tilde{f}(Y_0, \mathcal{C}(Y))|Y_0 = x, Y = i\right] \\ &= E_P\left[f(Y_0, \mathcal{C}(Y))|Y_0 = x, Y = i\right] \end{aligned}$$

and

$$\begin{aligned} E_{\tilde{P}}\left[\tilde{f}(Y_0, Y_{1:K-1})|Y_0 = x, Y_{j_1} = 0, \dots, Y_{j_{k-1}} = 0\right] &= E_P\left[\tilde{f}(Y_0, \mathcal{C}(Y))|Y_0 = x, Y \notin \{j_1, \dots, j_{k-1}\}\right] \\ &= E_P\left[f(Y_0, Y)|Y_0 = x, Y \notin \{j_1, \dots, j_{k-1}\}\right] \end{aligned}$$

$\square$

Finally, we can give several examples of the different increments involved in the Shapley values of each variable X or Y_k. If $k = 0$, then $Z' = \emptyset$ and

$$\Delta(\tilde{f}; Z', i) = \Delta(\tilde{f}; \emptyset, i) = E_P\left[f(X, Y)|Y = i\right] - E_P\left[f(X, Y)\right]$$

If $k = 1$, then $Z' = \{0\}$ or $Z' = \{j\} \neq \{i\}$,

$$\begin{aligned} \Delta(\tilde{f}; Z', i) = \Delta(\tilde{f}; 0, i) &= E_P\left[f(X, Y)|X = x, Y = i\right] - E_P\left[f(X, Y)|X = x\right] \\ \Delta(\tilde{f}; Z', i) = \Delta(\tilde{f}; \{j\}, i) &= E_P\left[f(X, Y)|Y = i\right] - E_P\left[f(X, Y)|Y \neq j\right] \end{aligned}$$

For $k = K - 1$, $Z' = \{1, \dots K - 1\}$,

$$\Delta(\tilde{f}; \{1, \dots K-1\}, i) = E_P\left[f(X, Y)|X = x, Y = i\right] - E_P\left[f(X, Y)|X = x, Y \neq i\right]$$

The propositions G.1 and G.2 show that the individual Shapley value for the variable (modality) Y_i is a weighted mean of the difference between classe i and group of classes:

$$\begin{cases} E_P\left[f(X, Y)|Y = i\right] - E_P\left[f(X, Y)|Y \notin \{j_1, \dots, j_k\}\right] \\ E_P\left[f(X, Y)|X = x, Y = i\right] - E_P\left[f(X, Y)|X, Y \notin \{j_1, \dots, j_{k-1}\}\right] \end{cases}$$

Finally, we can also compute the Shapley values of the other variables Y_j at point $(x, y = i)$, for $j \neq i$. In that case, the difference $\Delta(\tilde{f}; Z', j), j \neq i$ are of the type of

$$\begin{cases} E_P\left[f(X,Y)|Y \notin \{j, j_1, \dots, j_k\}\right] - E_P\left[f(X,Y)|Y \notin \{j_1, \dots, j_k\}\right] \\ E_P\left[f(X,Y)|Y = i\right] - E_P\left[f(X,Y)|Y = i\right] \\ E_P\left[f(X,Y)|X = x, Y \notin \{j, j_1, \dots, j_k\}\right] - E_P\left[f(X,Y)|X, Y \notin \{j_1, \dots, j_{k-1}\}\right] \\ E_P\left[f(X,Y)|X = x, Y = i\right] - E_P\left[f(X,Y)|X, Y = i\right] \end{cases}$$

The Shapley values computes a mean of the difference between different aggregation of modalities, that contains or not the variable of interest.

As a conclusion of this part, we see that the individual Shapley values $\phi_k(\tilde{f}; x, y_{1:K-1})$ perform a multiple comparison of the means obtained by aggregating the classes or modalities in various ways, looking at the presence or not of the modality k. These differences of means have weights $\frac{1}{\binom{K-1}{k}}$ where k is basically the number of classes of the variable Y that we aggregate.

Consequently the sum $\sum_{k=1}^{K} \phi_k(\tilde{f}; x, y_{1:K-1})$ is clearly different from the

$$\phi_Y(f; x, y) = \frac{1}{2}\left(E\left[f(X,Y)|Y = y\right] - E\left[f\left(X,Y\right)\right]\right) + \frac{1}{2}\left(f\left(x,y\right) - E\left[f\left(X,Y\right)|X = x\right]\right).$$

This latter has a much more global analysis that aims at measuring how the mean $E\left[f(X,Y)|Y = y\right]$ in the various classes changes w.r.t $E\left[f\left(X,Y\right)\right]$, while the individual Shapley focus on the difference between subgroups of classes.

H Plug-In estimator of Marginal expectation

As we have indicated in the paper, the Shapley Values can be computed with different probability $Q_{S,\boldsymbol{x}}$. In that section, we show that when we use the marginal distribution (as in the so-called interventional case), the previous estimators for tree-based models can be adapted straightforwardly.
We consider then decision tree

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

and remark that the Marginal Shapley coefficients involve the computations of the marginal expectations $E_P\left[\mathbb{1}_{L_m}(\boldsymbol{X}_{\bar{Z}}, \boldsymbol{x}_Z)\right]$ for any subgroup of variables Z. On real data, we need to compute the conditional expectations, but we use the Tree approximations in order to replace

$$\begin{aligned} E_P\left[\mathbb{1}_{L_m}(\boldsymbol{X}_{\bar{Z}}, \boldsymbol{x}_Z)\right] &= \int\int \mathbb{1}_{L_m}(\boldsymbol{u}_{\bar{Z}}, \boldsymbol{x}_Z) p(\boldsymbol{u}_{\bar{Z}}, \boldsymbol{u}_Z) d\boldsymbol{u}_{\bar{Z}} d\boldsymbol{u}_Z \\ &= \int\int \mathbb{1}_{L_m}(\boldsymbol{u}_{\bar{Z}}, \boldsymbol{x}_Z) p(\boldsymbol{u}_Z|\boldsymbol{u}_{\bar{Z}}) p(\boldsymbol{u}_{\bar{Z}}) d\boldsymbol{u}_Z d\boldsymbol{u}_{\bar{Z}} \\ &= \int \left\{\int p(\boldsymbol{u}_Z|\boldsymbol{u}_{\bar{Z}}) d\boldsymbol{u}_Z\right\} \mathbb{1}_{L_m}(\boldsymbol{u}_{\bar{Z}}, \boldsymbol{x}_Z) p(\boldsymbol{u}_{\bar{Z}}) d\boldsymbol{u}_{\bar{Z}} \\ &= \int \mathbb{1}_{L_m}(\boldsymbol{u}_{\bar{Z}}, \boldsymbol{x}_Z) p(\boldsymbol{u}_{\bar{Z}}) d\boldsymbol{u}_{\bar{Z}} \end{aligned}$$

This means that we just need the marginal distributions of the variables $\boldsymbol{X}_{\bar{Z}}$ in order to compute the expectations of the leaf. In the case of quantitative data, the leaf can be written $L_m = \prod_{i=1}^{p}\left[a_i^m, b_i^m\right]$, and we have by definition

$$\exists k \in Z, \, x_k \notin [a_k, b_k] \implies \mathbb{1}_{L_m}(\boldsymbol{u}_{\bar{Z}}, \boldsymbol{x}_Z) = 0$$

We define the set of leafs compatible with condition $\boldsymbol{X}_Z = \boldsymbol{x}_Z$ as

$$C(Z, \boldsymbol{x}) = \left\{ m \in [1 \dots M] \,|L_m = \prod_{i=1}^{p}\left[a_i^m, b_i^m\right], \forall k \in Z, x_k \in \left[a_k^m, b_k^m\right] \right\}$$

We write for $m \in C(Z, \boldsymbol{x})$, $L_m = L_m^{\bar{Z}} \times L_m^{Z}$, with $L_m^{\bar{Z}} = \prod_{i \in \bar{Z}} [a_i^m, b_i^m]$ and $L_m^{Z} = \prod_{i \in Z} [a_i^m, b_i^m]$, this means that for all $m \in C(Z, \boldsymbol{x})$ we have

$$E_P \left[\mathbb{1}_{L_m}(\boldsymbol{X}_{\bar{Z}}, \boldsymbol{x}_Z) \right] = E_P \left[\mathbb{1}_{L_m^{\bar{Z}}}(\boldsymbol{X}_{\bar{Z}}) \right]$$

As an approximation, the conditional probability for $m \in C(Z, \boldsymbol{x})$ is computed as

$$\begin{aligned} E_P \left[\mathbb{1}_{L_m^{\bar{Z}}}(\boldsymbol{X}_{\bar{Z}}) \right] &= P\left(X_i \in [a_i^m, b_i^m], i \in \bar{Z} \right) \\ &\simeq \frac{N(L_m^{\bar{Z}})}{N} \end{aligned}$$

where $N(L_m^{\bar{Z}})$ is the number of observations in the (partial) leaf $L_m^{\bar{Z}}$. As a consequence we have

$$\begin{aligned} E_P \left[f(\boldsymbol{X}_{\bar{Z}}, \boldsymbol{x}_Z) \right] &= \sum_{m=1}^{M} \hat{y}_m E_P \left[\mathbb{1}_{L_m}(\boldsymbol{X}_{\bar{Z}}, \boldsymbol{x}_Z) \right] \\ &= \sum_{m \in C(Z, \boldsymbol{x})} \hat{y}_m E_P \left[\mathbb{1}_{L_m}(\boldsymbol{X}_{\bar{Z}}, \boldsymbol{x}_Z) \right] \\ &= \sum_{m \in C(Z, \boldsymbol{x})} \hat{y}_m E_P \left[\mathbb{1}_{L_m^{\bar{Z}}}(\boldsymbol{X}_{\bar{Z}}) \right] \\ &\simeq \sum_{m \in C(Z, \boldsymbol{x})} \hat{y}_m \frac{N(L_m^{\bar{Z}})}{N} \end{aligned}$$

I EXPERIMENTAL SETTINGS

All our experiments are reproducible and can be found on the github repository *Active Coalition of Variables*, https://github.com/acvneurips/ACV

A.1 Toy model of Section 2.3

Recall that the model is a linear predictor with categorical variables define as $f(X, Y) = B_Y X$ with $X|Y = y \sim \mathcal{N}(\mu_y, \Sigma_y)$ and $\mathbb{P}(Y = y) = \pi_y, Y \in \{a, b, c\}$.
For the experiments in Figure 1 and 2, we set $\pi_y = \frac{1}{3}, \mu_y = 0 \forall\ y \in \{a, b, c\}$. We use a random matrices generated from a Wishart distribution. The covariance matrices are:
$\Sigma_a = \begin{bmatrix} 0.41871254 & -0.790061361 & 0.46956991 \\ -0.79006136 & 1.90865098 & -0.82571655 \\ 0.46956991 & -0.82571655 & 0.95835472 \end{bmatrix}$, $\Sigma_b = \begin{bmatrix} 0.55326081 & 0.11811951 & -0.70677924] \\ 0.11811951 & 2.73312979 & -2.94400196 \\ -0.70677924 & -2.94400196 & 4.22105088 \end{bmatrix}$,
$\Sigma_c = \begin{bmatrix} 9.2859966 & 1.12872646 & 2.4224434 \\ 1.12872646 & 0.92891237 & -0.14373393 \\ 2.4224434 & -0.14373393 & 1.81601676 \end{bmatrix}$ for $y \in \{a, b, c\}$ respectively.
The coefficients are $B_a = [1, 3, 5], B_b = [-5, -10, -8], B_c = [6, 1, 0]$ and the selected observation in figure 1 values is $x = [0.35, -1.61, -0.11, 1., 0., 0.]$

A.2 Toy model of Section 3.3

The data $\mathcal{D}_x^{(Train)} = (x_i, z_i)_{1 \leq i \leq n}$ are generated from a linear regression $Z = B^t X$ with $n = 10000$, $X \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = 0$, $\Sigma =$ 0.7 × np.ones(d, d) - (0.7-1)× np.eyes(d). d=5, $B = [6.49, -2.44, -2.11, -4.29, 3.46]$ for the continuous case and d=3, $B = [6.49, -2.44, 0]$ for the discrete case.

We used a decision tree on $\mathcal{D}^{(Train)}$ with the defaults parameters. The Mean Squared Error (MSE) are MSE $= 4.39$ for the continuous case and MSE $= 2.88$ for the discrete case.

A.3 Comparisons of calculation time of SV estimates with ACV and TreeSHAP

We show below a run-time comparison of the computation of n SV with ACV and TreeSHAP. We used 3 datasets with different shape: Boston (N=506, p=13), Adults (N=32561, p=12), Toy linear model (N=50000, p=500). The model used was XGBoost with default parameters (ntree = 100, maxdepth = 6). We compute the SV of n=1000 observations for Adults, Toy model and n=506 for Boston.

Datasets	Boston($n = 506, p = 13$)	Adults ($n = 1000, p = 12$)
Leaf	8.82 s (204 ms)	1 min 4 s (1.73 s)
Tree SHAP	129 ms (6.91 ms)	3.33 s (39.9 ms)

This difference in runtime can be partly explained by the fact that Leaf estimator has to go through all the data for each leaf, whereas TreeSHAP uses the information stored in the tree.

References

[1] Clément Bénard, Gérard Biau, Sébastien Da Veiga, and Erwan Scornet. Sirus: Stable and interpretable rule set for classification. *Electronic Journal of Statistics*, 15(1):427–505, 2021.

[2] S. Chen, Arthur Choi, and Adnan Darwiche. The same-decision probability: A new tool for decision making. 2012.

[3] Suming Chen, Arthur Choi, and Adnan Darwiche. An exact algorithm for computing the same-decision probability. IJCAI '13, page 2525–2531. AAAI Press, 2013.

[4] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[5] Dataset LUCAS. Lucas (lung cancer simple set) dataset. http://www.causality.inf.ethz.ch/data/LUCAS.html.

[6] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

[7] Eric Wang, Pasha Khosravi, and Guy Van den Broeck. Towards probabilistic sufficient explanations. In *Extending Explainable AI Beyond Deep Models and Classifiers Workshop at ICML (XXAI)*, 2020.