Important variables are Game-Changers: Revisiting Shapley Values for explaining Black-Box models

Salim Ibrahim Amoukou [1] & Nicolas Brunel [2]

[1] *Stellantis et Université Paris Saclay, CNRS, Laboratoire de Mathématiques et Modélisation d'Evry, salim.ibrahim-amoukou@universite-paris-saclay.fr*
[2] *Université Paris Saclay, CNRS, ENSIIE Laboratoire de Mathématiques et Modélisation d'Evry, nicolas.brunel@ensiie.fr*

Résumé. L'explicabilité des modèles de Machine Learning est un domaine très actif, car il est un vecteur important de l'acceptabilité des algorithmes d'Intelligence Artificielle. Parmi les techniques récemment proposées, les valeurs de Shapley émergent comme un indicateur de référence, car il fournit une explication additive des prédictions. Cependant les valeurs de Shapley utilisées actuellement peuvent être empreintes d'erreurs d'estimation, et sensibles à la présence de variables peu importantes. Nous avons développé un algorithme qui permet de calculer les "same decision probabilities" qui mesurent la probabilité de garder la même décision en ne fixant qu'une partie des variables prédictives. Ceci nous permet d'introduire un nouveau jeu coopératif qui permet de montrer que les variables qui contribuent le plus à cette stabilité sont des variables importantes du modèle. Nous illustrons les concepts proposés sur un modèle graphique.

Mots-clés. Valeurs de Shapley, Explicabilité, Apprentissage Automatique, Sélection de variables, Importance de Variables.

Abstract. The explainability of Machine Learning models is a very active field, because it is an important vector of the acceptability of AI algorithms. Among the recently proposed techniques, Shapley Values have emerged as a gold-standard, as it provides an additive explanation of predictions. However, Shapley Values are often prone to estimation errors, and are sensitive to the presence of unimportant variables. We have introduced a new computation algorithm which allows us to compute also the "Same Decision Probability", which measures the probability of keeping the same decision by fixing a subset of the predictor variables. Our main contribution is the introduction of a new cooperative game that shows that the variables who acts as game-changer are the important variables of the model. We illustrate our findings on a graphical model.

Keywords. Shapley values, Explainable AI, Variable Importance, Variable Selection, Machine Learning.

1 Introduction

This work addresses the problem of interpretability of Machine Learning models. Despite a growing use of machine learning in applications and real life problems, a significant part of previous academic works was dedicated to the improvement of the prediction capabilities or computational efficiencies of ML Models until the recent years. The objective of the very active and recent field of Explainable AI (XAI) aims at developing tools that could provide better insights in the important variables, at a global or at a local level. While statistical models are often based on some testable assumptions, or might be interpretable by design, there is a need for development of model-agnostic importance measures for ML models, in order to be able to understand the differences between very diverse models and to perform some sort of variables selection. Among the most used local measures, the Shapley Values (SV) comes from cooperative game theory and evaluates the "fair" contribution of a variable $X_i = x_i$ in a prediction [1]. One of the main interest of Shapley values, is that they provide an additive (decomposition) explanation of the prediction, which makes it relatively easy to understand. While Shapley Values are considered as one of the state-of-the-art methodology, several critics have been addressed concerning the computational complexity and the approximations needed, or the difficulty to relate them to other interpretable frame work, such as causality [2, 3]. We recall in the next section the definition of the Shapley Values. In section 3, we introduce a measure of stability, called "same decision probability" that we use for computing the importance of group of variables. Finally, in order to obtain a reliable estimate of the importance of a variable we define a new cooperative game where we can define "Swing Shapley Values" that are different than the standard ones introduced in [1]. Finally, we show that the variables that perform as game-changer when we consider swinging coalitions, are important variables. An example on realistic data illustrates our findings and conclusions.

2 Feature attribution and Shapley Values

We recall in this section the definitions and main properties of Shapley Values.

2.1 Shapley values for explaining Black-Box models

For any group of variables $\boldsymbol{X}_S = (X_i)_{i\in S}$, with any subset $S \subseteq [\![1, p]\!]$, we define the reduced predictors as

$$f_S(\boldsymbol{x}_S) \triangleq E\left[f\left(\boldsymbol{X}\right)|\boldsymbol{X}_S = \boldsymbol{x}_S\right]. \tag{2.1}$$

The SV for local interpretability at $\boldsymbol{x}$ are based on a cooperative game with the value function $v(f; S) \triangleq f_S(\boldsymbol{x}_S)$ (a value function is a function from 2^p set to $\mathbb{R}$). 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 SV of the coalition C is defined as

$$\phi_C(f;\boldsymbol{x}) = \frac{1}{p-|C|+1} \sum_{k=0}^{p-|C|} \frac{1}{\binom{p-|C|}{k}} \sum_{S \in \mathcal{S}_k(C)} \left(f_{S\cup C}(\boldsymbol{x}_{S\cup C}) - f_S(\boldsymbol{x}_S)\right)) \tag{2.2}$$

The definition (2.2) 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 $i \in [\![1,p]\!]$. A reference code for computing SV is the Python Open source library SHAP[1], that implements various approximation algorithms.
The great benefit of the Shapley Values is the so-called additive explanation:

$$f(\boldsymbol{x}) - E[f(\boldsymbol{X})] = \sum_{i=1}^{p} \phi_i(f,\boldsymbol{x}) \tag{2.3}$$

which permits to measure directly the influence of the variable X_i on the prediction. A common classical criticism is about the effective estimation of the expectations needed in the SV computation that is statistically challenging and combined with an exponential complexity. 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 [4].

2.2 Closed-form expressions for reduced predictors

The computation of the SV uses all the conditional expectations $E\left[f(\boldsymbol{X})|\boldsymbol{X}_S = \boldsymbol{x}_S\right], S \subseteq [\![1,p]\!]$. While it is difficult in general, the paper [4] introduce a recursive algorithm that reads sequentially and recursively the different nodes. In practice, the conditional expectations need to be estimated from the training or the test set. With tree-structured models, we can have efficient algorithms for computing in closed-form conditional expectations and SV. We assume that we have a tree with M leafs $L_1, \ldots, L_M$ based on the variables $X_1, \ldots, X_p$ (continuous or qualitative), the predictor f is a tree $f(\boldsymbol{x}) = \sum_{m=1}^{M} f_m \mathbb{1}_{L_m}(\boldsymbol{x})$. The reduced predictor is $f_S(\boldsymbol{x}_S) = \sum_{m=1}^{M} f_m P_X\left(L_m \,|\boldsymbol{X}_S = \boldsymbol{x}_S\right)$ showing that the only challenge is the computation of the conditional probabilities. We have implemented in a Python package *Active Coalition of Variables*[2] an efficient algorithm for computing more accurate conditional probabilities and Shapley Values for tree-based models.

3 A new game for Variable Importance

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

[1] https://github.com/slundberg/shap
[2] https://github.com/salimamoukou/acv00

uninformative variables $\boldsymbol{X}_i, i \in \bar{S}$. Therefore, several papers [5, 6, 7] suggest to use SV as a heuristic for feature selection, but as proved in [2], the magnitude of SV of variables do not necessarily correspond to relevant variables. Indeed, a variable can have a low influence but paradoxically, it can have at the same time a high $\phi_i(f, \boldsymbol{x})$. So we need to filter the noisy variables.

3.1 Same Decision Probability and game changer

Our methodology for identifying the most important features is based on the Same Decision Probability (SDP) criterion, introduced in [8], and that can be computed for tree-based models in the *Active Coalition of Variables* library.

Definition 3.1 (**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. We introduce a new cooperative game that will put emphasis on the game-changers i.e on the variables that make the decision becoming stable when they enter into a coalition.

Definition 3.2 (**Swing Shapley Values**). *Let f a model, $\boldsymbol{x}$ an instance, $SDP_S(f; \boldsymbol{x})$ the same decision probability of coalition S. We define the new cooperative game with value function:*

$$v_{SDP,\pi}(f; S) = \begin{cases} 1, & \text{if } SDP_S(f; \boldsymbol{x}) \geq \pi \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

For this game, we can also compute the corresponding Shapley Value (denoted Swing-SV) in order to compute the overall contribution of a variable to the game induced by the value function (3.1). The Swing-SV $\phi_i^{SDP,\pi}$ of variable X_i is then computed by replacing $f_S(\boldsymbol{x}_S)$ by $v_{SDP,\pi}(S)$ in the standard definition of a Shapley Value (2.2). A coalition with value zero is called a "losing coalition" and with value one a "winning coalition". If a player's entry into a coalition changes the value from losing to winning, then the player's contribution is one, otherwise zero. A coalition S is said to be a *swing* for player i if S is losing but $S \cup i$ is winning. Therefore, a high Swing-SV $\phi_i^{SDP,\pi}$ implies that the variable X_i generates a lot of swings and is a game-changer; i.e this variable permits to retrieve significantly the original prediction. However, it should be noted that the SV $\phi_i^{SDP,\pi}$ can be negative, especially when the variable is not very important. In that latter case, the variable is not important enough to make a lot of swings, while correlations with other variables and local over-fitting induce a lot of reverse-swings (i.e adding the variable transforms a winning coalition into a loosing coalition).

3.2 A graphical model: LUCAS

To illustrate our method, we use a dataset generated by the Causal Bayesian network LUCAS[3], used for modeling the occurence of a Lung Cancer, based on a network of 11 binary variables. The variables "Smoking, Coughing, Allergy, Genetics, Fatigue" are Markov Blanket, the 6 other variables are not directly related to the target. We want to explain an observation with a well-defined ground truth. We know from the probability tables that if Smoking, Genetic, Coughing are True, the probability of having Cancer is very high. So, these three variables should have a high Swing-SV. To better analyze the behavior of the Swing-SV values of the new game, they are calculated for different values of π.

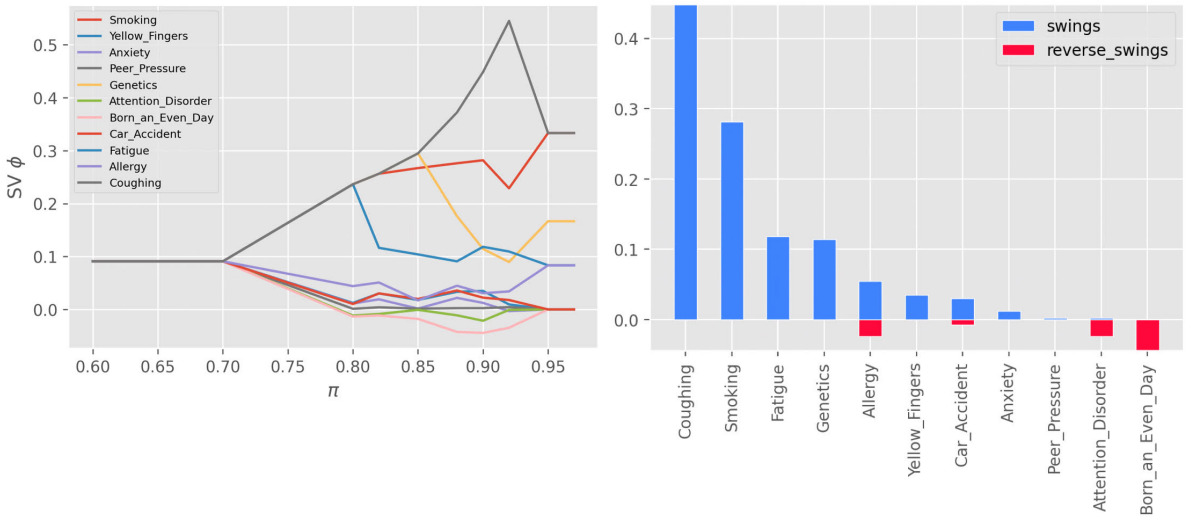

Figure 1: Left: Swing-SV given π. Right: Additive decomposition of the Swing-SV ($\pi = 0.9$).

We observe in the left of figure 1 that all the features have the same Swing-SV for low values of π (below 0.7): all the features have the same rate of swings when the condition is to give the same decision at low level π. For higher probability π, the three expected variables (Smoking, Coughing, Genetic) stand out. The variables Fatigue, Allergy seems important, but the remaining variables have almost zero contributions. In addition, we have also an additive explanation based on the Swing-SV, in order to know if its value comes essentially from the swings or the reverse-swings: we argue that we need to avoid means, as it blurs the interpretation. In the right of figure 1, we remark that important variables do not make any reverse-swings, while irrelevant variables do. Even more, reverse-swings dominate for noisy variables.

[3]http://www.causality.inf.ethz.ch/data/LUCAS.html

4 Conclusion

The Shapley Values for explainable AI are a very useful and insightful methodology for evaluating the importance of variables. While the "standard" Shapley Values can be criticized, we think that the introduction of more adapted game can give a better assessment of the impact of a variable on a decision. In particular, the same decision probability, that evaluates the stability of the decision with respect to fixed subgroups of variables, offers a promising direction for feature attribution and variable selection at a local scale, and possibly at a global scale.

References

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*.

[2] S. Ma and R. Tourani, "Predictive and causal implications of using shapley value for model interpretation," in *Proceedings of the 2020 KDD Workshop on Causal Discovery* (T. D. Le, L. Liu, K. Zhang, E. Kiciman, P. Cui, and A. Hyvärinen, eds.), vol. 127 of *Proceedings of Machine Learning Research*, pp. 23–38, PMLR, 2020.

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

[4] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[5] M. Zaeri-Amirani, F. Afghah, and S. Mousavi, "A feature selection method based on shapley value to false alarm reduction in icus a genetic-algorithm approach," vol. 2018, pp. 319–323, 07 2018.

[6] S. B. Cohen, G. Dror, and E. Ruppin, "Feature selection via coalitional game theory," *Neural Comput.*, vol. 19, no. 7, pp. 1939–1961, 2007.

[7] X. Sun, Y. Liu, J. Li, J. Zhu, X. Liu, and H. Chen, "Using cooperative game theory to optimize the feature selection problem," *Neurocomputing*, vol. 97, pp. 86–93, 2012.

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