Shapley Residuals: Quantifying the limits of the Shapley value for explanations

I. Elizabeth Kumar [1] Carlos Scheidegger [2] Suresh Venkatasubramanian [1] Sorelle A. Friedler [3]

Abstract

Popular feature importance techniques compute additive approximations to nonlinear models by first defining a cooperative game describing the value of different subsets of the model's features, then calculating the resulting game's Shapley values to attribute credit additively between the features. However, the specific modeling settings in which the Shapley values are a poor approximation for the true game have not been well-described. In this paper we utilize an interpretation of Shapley values as the result of an orthogonal projection between vector spaces to calculate a *residual* representing the kernel component of that projection. We provide an algorithm for computing these residuals, characterize different modeling settings based on the value of the residuals, and demonstrate that they capture information about model predictions that Shapley values cannot.

1. Introduction

There have been many recent efforts to quantify the importance of features to a model (Ribeiro et al., 2016; Datta et al., 2016; Adler et al., 2018; Marx et al., 2019; Lundberg and Lee, 2017; Lundberg et al., 2018a). Many of these determine the importance through estimating the Shapley value of a game designed to assign importance to sets of features (Datta et al., 2016; Lundberg and Lee, 2017; Frye et al., 2019; Lundberg et al., 2018a;b; Dhamdhere et al., 2019). These Shapley-value-based feature importance methods have been used widely in practice (e.g., (Lundberg et al., 2018b), and survey / analysis (Bhatt et al., 2020)). At the same time, there have been increasing concerns that these game theoretic values may not completely capture human or technical notions of feature importance (Kumar et al., 2020; Slack et al., 2020; Sundararajan and Najmi, 2019). One of these concerns is that the Shapley value is only a summary of the cooperative game which describes a model's dynamics, and does not fully describe that game (Kumar et al., 2020).

In this work, we introduce **Shapley Residuals**, vector-valued objects that capture information lost by Shapley values. Shapley residuals can be associated with individual variables, as well as with sets of variables. When the residual of a feature exhibits a large norm, the associated Shapley value should be taken with skepticism: the resulting importance is not just due to the variable acting by itself. On the other hand, if a residual is small, most of the effect of the variable on the model is explainable by the variable acting independently (we make these statements precise in Section 3).

Consider the following two motivating scenarios. First, suppose a practitioner uses Shapley values to determine the effect of *data interventions* on model outcomes. Consider two models f_1 and f_2. In a real-world scenario, the practitioner will often only have black-box access to such models, and the models will often be significantly more complex. Here, we use these simple models:

$$\begin{aligned} f_1(x_1, x_2, x_3) &= x_1 + x_2 + x_3 \\ f_2(x_1, x_2, x_3) &= x_1 + 2x_2x_3 \end{aligned}$$

Suppose the practitioner seeks to explain the output $f_1(1, 1, 1) = 3$ or $f_2(1, 1, 1) = 3$, using KernelSHAP (Lundberg and Lee, 2017) to compute local feature importances. For both models, the Shapley values of x_1, x_2, and x_3 are all 1. Despite that, intervening by increasing the value of x_2 changes f_2 more than increasing the value of x_1; in f_1, this clearly does not happen. The Shapley residuals for all variables in f_1 are zero, indicating that variables in f_1 do not interact (as we prove in Section 4.1). The Shapley residuals for x_2 and x_3 in f_2, on the other hand, are nonzero, while the Shapley residual of x_1 is still zero. Finally, the Shapley residual for the set of variables $\{x_2, x_3\}$ is also zero.

In the second scenario, consider a data generating distribution where α controls the correlation between two features

[1] School of Computing, University of Utah, Salt Lake City, UT, USA [2] Department of Computer Science, University of Arizona, Tucson, AZ, USA [3] Department of Computer Science, Haverford College, Haverford, PA, USA. Correspondence to: I. Elizabeth Kumar <kumari@cs.utah.edu>.

2020 ICML Workshop on Human Interpretability in Machine Learning (WHI 2020). Copyright 2020 by the author(s).

in X and a regression target y:

$$(X, y) \sim \left(N\left((0,0), \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix} \right), \langle X, (3,1) \rangle \right).$$

We examine a regression model $f(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2$ determined via linear least squares. Assume access to infinitely many IID samples from (X, y), $\beta = (3,1)$. Suppose a practitioner wanted to explain the output of $f(1,1) = \beta_1 + \beta_2$, this time using Conditional Expectation SHAP (Sundararajan and Najmi, 2019). The Shapley values are $\beta_1 + \alpha(\beta_2 - \beta_1)/2$ for x_1 and $\beta_2 + \alpha(\beta_1 - \beta_2)/2$ for x_2. When $\alpha \approx 0$, Shapley values correspond to model weights β_1, β_2, and support a (valid) interventional interpretation that changing x_1 yields a larger change to the output of f than does x_2. However, if $\alpha \neq 0$, Shapley values do not support this interpretation. A practitioner employing Shapley values alone lacks the information to distinguish these scenarios. Shapley residuals provide useful diagnostic information; the norm of the residuals for x_1 and x_2 is exactly linearly proportional to α.

In these simple scenarios, it is clear that Shapley residuals capture, respectively, *variable interactions* and *mismatches between dependent features in the data and independent variables in the model*. As we show in Section 5, these observations apply to real-world scenarios as well.

In summary, we:

- introduce *Shapley residuals* (Section 3), which characterize the limits of Shapley values as explanatory mechanisms for cooperative games,
- study the properties of Shapley residuals both in general and in context of existing formulations for explanatory games (Sections 3 and 4.1),
- show via a number of experiments that Shapley residuals capture meaningful information for model explanations in realistic scenarios (Section 5), and
- place Shapley residuals in context of the broader discussion of the goals of model interpretability, and caution against overinterpreting them (Section 6).

2. Preliminaries

Let V be a vector space and let L be a linear mapping from V to V. We denote $\mathcal{R}(L) = \{w \mid \exists v \in V, L(v) = w\}$ as the *range space* of L and $\text{Null}(L) = \{v | L(v) = 0\}$ as the *null space* of L.

Games. A *cooperative* game consists of d players and a *value function* $v : 2^{[d]} \to \mathbb{R}$ where as usual $[d] = \{1, \dots, d\}$. The quantity $v(S)$ represents the value of the game for a coalition of players $S \subset 2^{[d]} \triangleq N$. Without loss of generality we will assume that $v(\emptyset) = 0$, and that we can identify the game with v. Let the space of games be denoted by $\mathcal{G}$.

Definition 1 (Shapley values(Shapley, 1952))**.** *The Shapley values $\phi_i(v), i \in [d]$ are the unique values satisfying the properties*

Efficiency: $\sum_{i=1}^{d} \phi_i(v) = v(N)$.

Dummy: *If $v(S \cup \{i\}) = v(S)$ for all $S \subset N \setminus \{i\}$, then $\phi_i(v) = 0$.*

Symmetry: *If $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subset N \setminus \{i, j\}$, then $\phi_i(v) = \phi_j(v)$.*

Linearity: *If v, v' are two games on d players, then $\phi_i(\alpha v + \alpha' v') = \alpha \phi_i(v) + \alpha' \phi_i(v')$.*

The Shapley values are given by the equation

$$\phi_v(i) = \sum_{S \subseteq [d]} \frac{|S|!(d - |S| - 1)!}{d!} \left(v(S \cup i) - v(S) \right) \quad (1)$$

Let $\mathcal{I}$ denote the space of games v such that for all $S \subseteq N$, $v(S) = \sum_{i \in S} v(\{i\})$. $\mathcal{I}$ is called the space of *inessential games*. Note $\mathcal{I}$ is a subspace of $\mathcal{G}$. Intuitively an inessential game is one in which the player interactions are simple and additive: every player adds a fixed value $v(\{i\})$ to a coalition S independent of the composition of S.

We can define a local variant of inessentiality. We say that the game v is inessential *relative to* S if $v(C) = v(S) + v(C \setminus S)$ for all S and C such that $S \subset C \subset N$. That is, each coalition containing S obtains a value equal to the subcoalition S working separately from $C \setminus S$. An interesting special case is when the set S is the singleton $\{i\}$ – in this case we say that v is inessential *relative to* i.

Gradients on the hypercube. We can think of the set N as the d-dimensional hypercube $G = (V = N, E)$ with each vertex labeled by a set $S \subseteq N$ and edges between sets S and $S \cup \{i\}$ for all $i \in [d], S$. The differential operator $d : \ell_2(V) \to \ell_2(E)$ is then defined as

$$du(S, S \cup \{i\}) = u(S \cup i) - u(S)$$

Essentially d is a discrete gradient operator on G, with a corresponding adjoint operator d^*. Finally, we will define a partial gradient $d_i : \ell_2(V) \to \ell_2(E)$:

$$d_i u(S, S \cup \{j\}) = \begin{cases} u(S \cup j) - u(S) & i = j \\ 0 & \text{otherwise} \end{cases}$$

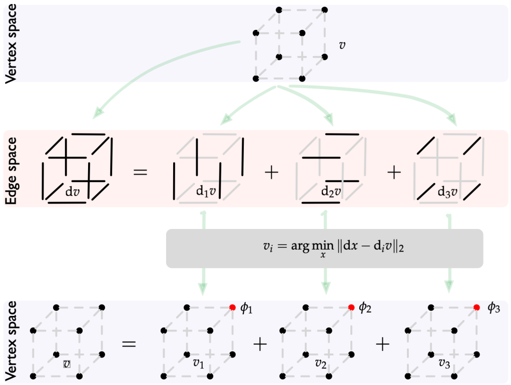


Figure 1. The decomposition of a game proposed by Stern and Tettenhorst (2019)

Intuitively, d_i only evaluates a gradient for edges corresponding to the insertion of i.

A key insight of Stern and Tettenhorst (2019) was to express inessentiality of games in terms of gradients on the hypercube.

Proposition 1 ((Stern and Tettenhorst, 2019, Prop 3.3))**.** *The game v is inessential if and only if $d_i v \in \mathcal{R}(d)$ for all $i \in N$.*

We can extend their result to relative inessentiality.

Proposition 2. *The game v is inessential* relative to S *if and only if $d_S v \in \mathcal{R}(d)$, where $d_S = \sum_{i \in S} d_i$*

2.1. Characterizing Shapley values

The main result by Stern and Tettenhorst (2019) is a decomposition of an arbitrary game v into games that are "close to being inessential" and allow extraction of Shapley values. Since v is not inessential, we cannot be sure to find v_i such that $\mathrm{d}_i v = \mathrm{d}v_i$, but we can find the "closest" such v_i. By the fundamental theorem of linear algebra, we can write

$$d_i v = dv_i + r_i$$

where $r_i \in \mathrm{Null}(\mathrm{d}^*)$ and $dv_i \in \mathcal{R}(\mathrm{d})$. Moreover, r_i is orthogonal to $\mathcal{R}(\mathrm{d})$ and so we can write v_i as the solution to the least squares problem

$$\min_{x \in \ell_2(V)} \|dx - d_i v\|$$

Stern and Tettenhorst (2019) show that for v_i defined as above, $v_i(N)$ is the ith Shapley value of v. We illustrate the construction in Figure 1.

3. Shapley Residuals

We are now ready to introduce our main contribution. Note that inessentiality of a game is key to having meaningful Shapley values. This is because when a game is inessential, each player i contributes precisely $v(\{i\})$ to each coalition it is part of, and therefore $\phi_i(v) = v(\{i\})$. This is true even locally: if v is inessential with respect to S, then S always contributes $v(S)$ to any coalition it participates in. Therefore, to understand the limits of Shapley values, we must quantify the degree of deviation from inessentiality.

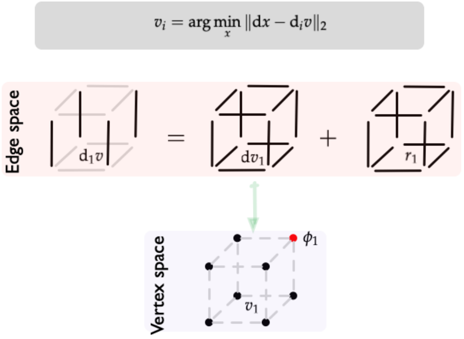


Figure 2. The construction of Shapley residuals

Definition 2 (Shapley Residuals)**.** *We call $r_i = d_i v - dv_i$ the* Shapley Residual *of player i. Analogously, $r_S = \sum_{i \in S} r_i$ is the* Shapley Residual *of set S.*

Shapley Residuals are a novel diagnostic tool for feature importance, and enjoy a number of relevant properties (which we prove in the supplementary material):

- v is inessential iff $r_i = 0$ for each $i \in N$ or, equivalently, iff $||r_i||^2 = 0$ for each $i \in N$. We use $\sum_i ||r_i||^2$ to characterize how far v is from being inessential.
- v is inessential relative to a set S iff $r_S = 0$ or, equivalently, iff $||r_S||^2 = 0$. We use $||r_S||^2$ to characterize how far v is from being inessential relative to S.
- $\sum_{i \in N} r_i = 0$. In words, v is always inessential relative to N and (vacuously) also always inessential relative to $\emptyset$.

Inessentiality implies inessentiality relative to all players and subsets, but the converse does not hold. If $r_S = 0$ for some subset S, this merely implies that players in S do not interact with players in $N \setminus S$. Figure 2 illustrates the construction of residuals.

4. Feature Importance and Residuals

We now apply this geometric framework to the problem of attributing feature importance via Shapley values. As has been noted, the different methods for Shapley value-based explanation (whether local or global) all reduce to a specific choice for the game v, at which point the Shapley values of v are estimated and returned (Sundararajan and Najmi, 2019; Kumar et al., 2020; Merrick and Taly, 2019).

We will show that the notion of relatively inessential games has a natural interpretation in the context of the two most popular forms of Shapley-based feature importance, KernelSHAP and SHAP, and show that the residuals capture information about the structure of a model that the Shapley value cannot.

4.1. Feature Importance Methods

The definitions of *Shapley sampling values* (Štrumbelj and Kononenko, 2014), as well as *SHAP values* (Lundberg and Lee, 2017), are derived from defining v as the *conditional* expected model output on a data point when only the features in S are known:

$$v_{f,x}^{Cond}(S) = E[f(\boldsymbol{X})|\boldsymbol{X}_S = \boldsymbol{x}_S]$$

We call this Conditional Expectation SHAP after Sundararajan and Najmi (2019).

KernelSHAP is derived from defining v by taking an expectation of f over $\bar{S}$'s joint marginal distribution while fixing the feature values from S:

$$v_{f,x}^{Kernel}(S) = E[f([\boldsymbol{x}_S, \boldsymbol{X}_{\bar{S}}])]$$

Notably, the two values are the same if the features in $\bar{S}$ are independent from those in S.

4.2. KernelSHAP Residuals

The behavior of residuals on KernelSHAP can be described with respect to the presence of interaction terms in a model.

Lemma 1. *Let $f : \boldsymbol{X} = \{X_1, X_2, ..., X_d\} \rightarrow Y$ be a multivariate function. Suppose f can be decomposed as $f(\boldsymbol{x}) = g(\boldsymbol{x}_S) + h(\boldsymbol{x}_{\bar{S}})$, for some functions $g : \{X_j : j \in S\} \rightarrow Y$ and $h : \{X_j : j \notin S\} \rightarrow Y$. Let $\boldsymbol{z} = \{z_1, z_2, ..., z_n\} \in \boldsymbol{X}$. Then $v_{f,z}^{Kernel}$ is relatively inessential with respect to the set S.*

This is important because if the model really does decompose additively for a certain variable i, the practitioner understands what to expect when variable i is perturbed. The KernelSHAP residuals thus quantify the extent to which the SHAP values describe interventional effects of the model.

4.3. Conditional Expectation SHAP Residuals

Just as the residual for KernelSHAP can be thought of as detecting feature interactions in a model, the residuals of Conditional Expectation SHAP can detect feature interactions in the data.

Lemma 2. *Let $f : \boldsymbol{X} = \{X_1, X_2, ..., X_d\} \rightarrow Y$ be a multivariate function. Suppose f can be decomposed as $f(\boldsymbol{x}) = g(\boldsymbol{x}_S) + h(\boldsymbol{x}_{\bar{S}})$, for some functions $g : \{X_j : j \in S\} \rightarrow Y$ and $h : \{X_j : j \notin S\} \rightarrow Y$. Let $\boldsymbol{z} = \{z_1, z_2, ..., z_n\} \in \boldsymbol{X}$. Suppose further that all $X_j : j \in S$ are distributed independently from all $X_j : j \notin S$. Then $v_{f,z}^{Cond}$ is relatively inessential with respect to set S.*

The residual on SHAP thus quantifies the extent to which SHAP values can be interpreted interventionally, because depending on the causal structure of the data, correlated features could imply that perturbing a feature i could result in the perturbation of a different feature as well and therefore the SHAP values cannot be interpreted interventionally.

5. Experiments

Given the theoretical justification presented in the previous sections for Shapley residuals, we focus here on examining what these residuals can help us understand about models on real-world data. Throughout, we use our own implementation of KernelSHAP to calculate the exact Shapley values and residuals.[1]

5.1. Occupancy Detection

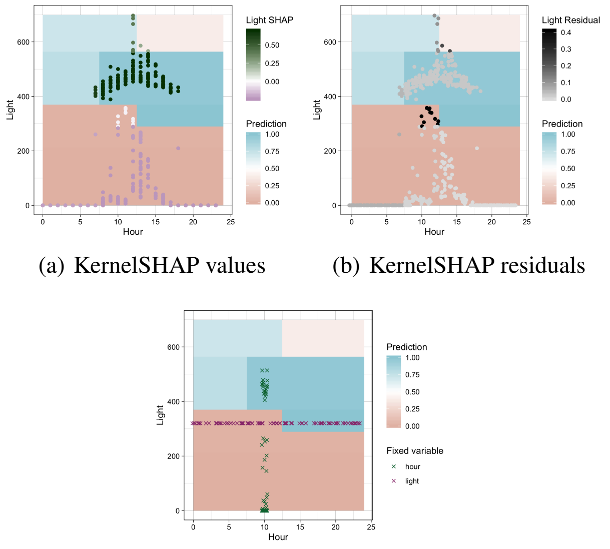
We consider the Shapley values and residuals for an occupancy detection dataset[2] with 20,560 instances used to predict whether an office room is occupied. The 7 attributes include a date stamp which is preprocessed to refer to an hour and day of the week. A decision tree model with a maximum depth of 3 is trained on 75% of this data using only the features `light` and `hour`. When evaluated on the remaining test set, the ROC-AUC for this decision tree is 0.991.

We calculate the Shapley values and residuals (using 50 randomly sampled background rows from the test set) for 1000 randomly sampled test instances. The results for the variable “light" are shown in Figure 3.

The reason that the cluster of points in the middle has a high residual is illustrated in Figure 3(c). Calculating the expected prediction while fixing a light value of 320, unlike most other possible values, results in a mix of low and high predictions. These average to 0.4, while both the overall expectation and particular prediction for occupancy probability for those points are 0.25. Specifically, the KernelSHAP game for $f(H, L) = P(\text{occupant} = T)$ for $L = 320$ and $H = 10$ is shown in the inset diagram. $L = 320$ is a positive indicator of occupancy if H is unknown (+.16) but is a “negative” indicator of occupancy is H is known to be 10 (-.24), due to the interactions in the model in this area of the feature space. The light Shapley value is close to 0

[1] Code is provided in the supplementary material

[2] https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+



(a) KernelSHAP values

(b) KernelSHAP residuals

(c) KernelSHAP sampling for (10,320)

Figure 3. Shapley values and residuals on a decision tree for the Occupancy Detection task

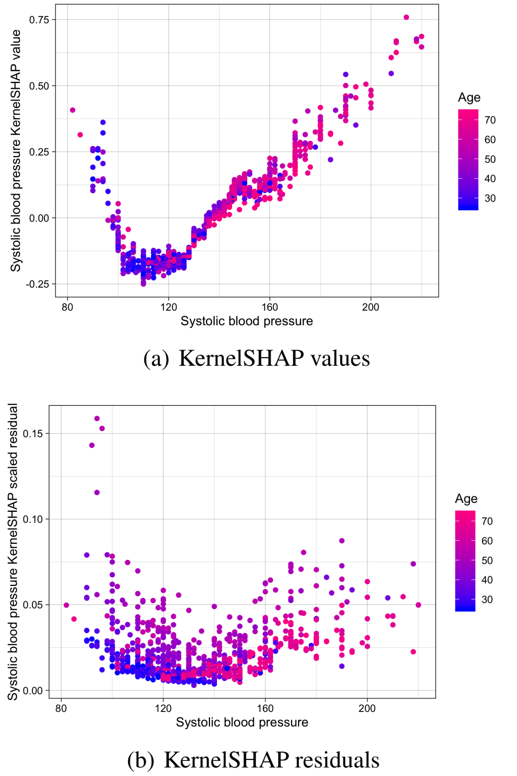
$$\begin{tikzcd}
E[f(H,L)] = .24 \arrow[r, "+.01"] \arrow[d, "+.16"'] & E[f(10,L)] = .25 \arrow[d, "-.24"] \\
E[f(H,320)] = .40 \arrow[r, "-.39"] & f(10,320) = .01
\end{tikzcd}$$

for points in this range, then, because it is the average of a positive and negative number – not because it is of “low importance” – and the non-inessentiality of this feature is what is being captured by the residual.

5.2. NHANES

The NHANES data, made available via the SHAP package[3], contains 9,932 instances of right-censored mortality data. We use the preprocessing of the data from the SHAP package and train an XGBoost Cox survival model with 5000 estimators on 7 variables (‘Age’, ‘Diastolic BP’, ‘Sex’, ‘Systolic BP’, ‘Poverty index’, ‘White blood cells’, and ‘BMI’). The resulting Harrell’s C-statistic on the test set is 0.825. We then explain its marginal predictions on 1000 randomly chosen test instances with KernelSHAP on 100 background samples. The resulting KernelSHAP values and residuals for some features are given in Figures 4 and 5.

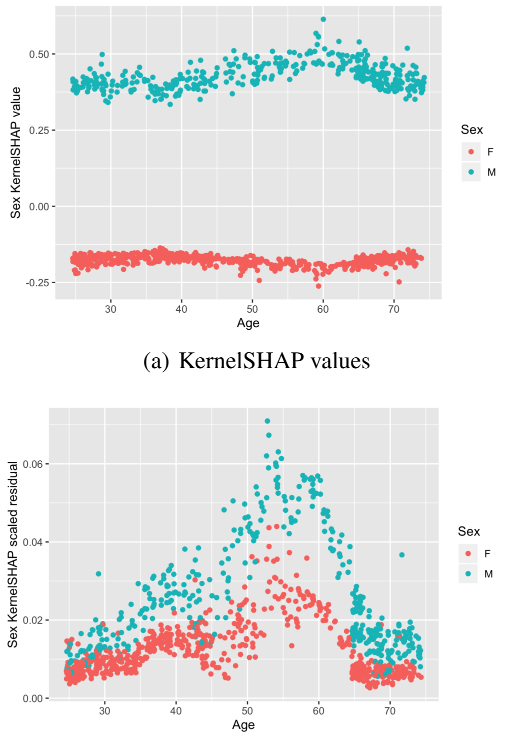
Considering the feature importance of blood pressure (Figure 4(a)), we find (as also found in Lundberg and Lee (2017)) that as blood pressure increases, the importance of blood pressure to mortality also increases, and this effect



(a) KernelSHAP values

(b) KernelSHAP residuals

Figure 4. Shapley values and residuals on an XGBoost mortality model for Systolic blood pressure and age.

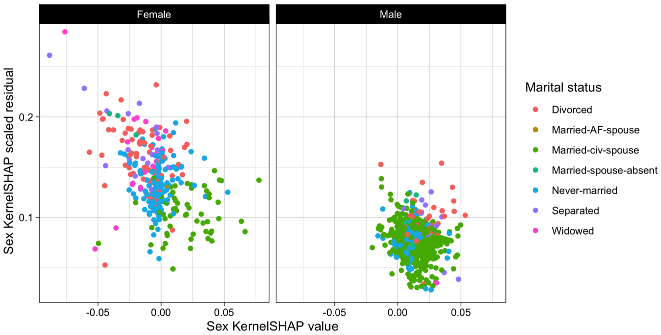


(a) KernelSHAP values

(b) KernelSHAP residuals

Figure 5. Shapley values and residuals on an XGBoost mortality model for age and sex.

[3] https://slundberg.github.io/shap/notebooks/NHANES%20I%20Survival%20Model.html



(a) KernelSHAP values

Figure 6. Shapley values and residuals on a random forest income model

is correlated with an increase in age. Examining the residuals (Figure 4(b)) additionally allows us to see that there is a missing importance associated with blood pressure for middle aged people across all blood pressure readings, perhaps indicating that blood pressure acts in combination with other variables to impact mortality for this age range.

When we consider the KernelSHAP feature importance of sex on mortality within this model (Figure 5(a)) we find the importance to be remarkably stable across age ranges, such that being male consistently has a much larger predictive impact on mortality risk. However, the residuals for these features and instances (Figure 5(b)) show that middle aged men may have many other interacting and contributing factors for predicting mortality. These two residual charts taken together (Figures 4(b) and 5(b)) may indicate that blood pressure, sex, and age interact within the model to increase the importance of both sex and blood pressure for mortality predictions of middle aged men.

5.3. Adult Income

The Adult Income dataset[4] contains 48,842 instances of people's census information from 1994, including 14 attributes describing their education, job, marital status, etc., and with the goal of predicting whether the person makes more or less than $50,000 per year. We preprocess the data by removing rows with missing values and train a random forest with 10 trees on all the variables (except `fnlwgt`) on 80% of the data. The ROC-AUC of the model evaluated on the remaining 20% is .857. We calculate the Shapley values and residuals using 1000 test instances and KernelSHAP with 25 background samples. The results for features sex and marital status are shown in Figure 6.

For both men and women, the distribution of Shapley values indicating the importance of sex to the income prediction model is close to a Shapley value of 0. However, in addition the Shapley values for women having a larger variance, we see with Shapley residuals that residuals for some women are also much higher than those for men. Specifically, while essentially all men have low residuals, essentially only women who are also married to civilian non-absent spouses have low residuals. This indicates that sex and marital status interact in more complex ways with the income prediction model for women than they do for men.

6. Discussion

Much of the motivation in interpretable machine learning, and especially within Shapley-value-based feature importance, is to give a rigorous theoretical foundation to interpretability notions so that practitioners can better understand the impacts of their models. This is especially important in societal contexts where models make high-stakes decisions about people, e.g., via criminal risk assessments and interview screening algorithms. We believe people have the right to understand those decisions, and particularly which features were important for the decision. Putting such feature importance measurements on solid theoretical grounds is important for the validity of these feature importance claims. Their validity is an important part of the ethics of algorithms as societal interventions.

Our motivation for this work is to contribute further to the theoretical foundation of Shapley-value-based feature importance measures and, critically, to quantify any missing importance via our introduced Shapley residuals. We believe that these Shapley residuals could have a positive societal impact by alerting practitioners to model complexities and importances that have previously gone unattended. One societal concern is that the meaning of these residuals may be hard for practitioners to understand and that errors in the interpretation of these residuals may cause unanticipated negative consequences. We encourage further research on how to best present these residuals for human understanding and interaction, and hope that even if they are only understood as a *warning* attached to specific Shapley values that may be useful for practitioners.

Additionally, we warn against Shapley residuals further encouraging the unskeptical use of Shapley values for feature importance. Previous research has shown that Shapley values are not well-aligned with the human understanding goals of feature importance (Kumar et al., 2020), as well as suffering from technical issues such as susceptibility to adversarial attacks (Slack et al., 2020). Given the already widespread use of Shapley values for feature importance, we felt that the increased scrutiny enabled by Shapley residuals outweighs the risk of further increased popularity.

[4] `http://archive.ics.uci.edu/ml/datasets/Adult`

Acknowledgments. This research was supported in part by the National Science Foundation under grants IIS-1633724, IIS-1633387, DMR-1709351, IIS-1815238, the DARPA SD2 Program, and the ARCS Foundation.

References

Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657, 2020.

Ozan Candogan, Ishai Menache, Asuman Ozdaglar, and Pablo A Parrilo. Flows and decompositions of games: Harmonic and potential games. *Mathematics of Operations Research*, 36(3):474–503, 2011.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016.

Kedar Dhamdhere, Ashish Agarwal, and Mukund Sundararajan. The shapley taylor interaction index. *arXiv preprint arXiv:1902.05622*, 2019. 37th International Conference on Machine Learning, to appear.

Christopher Frye, Ilya Feige, and Colin Rowat. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*, 2019.

Katsushige Fujimoto, Ivan Kojadinovic, and Jean-Luc Marichal. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55(1):72–99, 2006.

Norman L Kleinberg and Jeffrey H Weiss. Weak values, the core, and new axioms for the shapley value. *Mathematical Social Sciences*, 12(1):21–30, 1986.

I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. *arXiv preprint arXiv:2002.11097*, 2020. 37th International Conference on Machine Learning, to appear.

Scott M Lundberg. shap: A game theoretic approach to explain the output of any machine learning model, 2020. https://github.com/slundberg/shap/blob/master/shap/explainers/kernel.py#L L154.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018a.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018b.

Charles Marx, Richard Phillips, Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. Disentangling influence: Using disentangled representations to audit model predictions. In *Advances in Neural Information Processing Systems*, pages 4498–4508, 2019.

Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models with cooperative game theory, 2019.

Christopher C Paige and Michael A Saunders. Lsqr: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1):43–71, 1982.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Lloyd S Shapley. A value for n-person games. Technical report, Rand Corp Santa Monica CA, 1952.

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

Ari Stern and Alexander Tettenhorst. Hodge decomposition and the shapley value of a cooperative game. *Games and Economic Behavior*, 113:186–198, 2019.

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

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. *arXiv preprint arXiv:1908.08474*, 2019. 37th International Conference on Machine Learning, to appear.

7. Appendix

7.1. Proofs from Section 3

Stern and Tettenhorst define $C_{\sigma,i} = \{j \in N : \sigma(j) < \sigma(i)\}$. We will need the slight extension to sets $C_{\sigma,S} = \bigcap_{i \in S} C_{\sigma,i}$. Stern and Tettenhorst define and prove their main claims in Definition 3.1, 3.2, and Proposition 3.3. We generalize them to the many-player setting, providing Definition 3.1_S, 3.2_S, and Proposition 3.3_S.

Definition $\mathbf{3.1_S}$. For a subset $S \subset N$, let $\mathrm{d}_S \colon \ell^2(V) \to \ell^2(E)$ be the operator $\mathrm{d}_S = \sum_{i \in S} \mathrm{d}_i$, or

$$\mathrm{d}_S u(C, C \cup \{j\}) = \begin{cases} \mathrm{d}u\,(C, C \cup \{i\}) & \text{if } i = j \text{ and } i \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $\mathrm{d}_S v \in \ell^2(E)$ encodes the marginal value contributed by the player subset S to the game v. For any permutation σ of N, which defines a path from $\emptyset$ to N, the marginal value contributed by subset S along this path is

$$\sum_{j \in N} \mathrm{d}_S v(C_{\sigma,j}, C_{\sigma,j} \cup \{j\}) = \sum_{i \in S} \mathrm{d}_i v(C_{\sigma,j}, C_{\sigma,j} \cup \{j\}) = \sum_{i \in S} v(C_{\sigma,i} \cup \{i\}) - v(C_{\sigma,i}),$$

which can also be interpreted as a discrete "line integral" of $\mathrm{d}_S v$ along the path.

Definition $\mathbf{3.2_S}$. The game v is inessential *relative to* S if $v(C) = v(S) + v(C \setminus S)$ for all S and C such that $S \subset C \subset N$. That is, each coalition containing S obtains a value equal to the subcoalition S working separately from $C \setminus S$. In addition, inessentiality relative to a single player i is the same as inessentiality relative to the singleton set $\{i\}$.

Stern and Tettenhorst's 3.2 is a stronger condition than 3.2_i which in turn is stronger than 3.2_S. In other words, v being inessential implies that v is inessential relative to all $i \in N$ and all subsets $S \subset N$. v being inessential with respect to each player of $i, j, \ldots z$ implies that v is inessential relative to the set $\{i, j, \ldots, z\}$. The converses, on the other hand, do *not* generally hold.

Proof. We need to show that the different paths that can be taken through the nonzero entries of d_S sum to the same value. But if $\mathrm{d}_S v \in \mathcal{R}(\mathrm{d})$, then the right-hand side of the result in the sum defined in $3.1_\mathbf{S}$ telescopes to $v(C_{\sigma,S} \cup S) - v(C_{\sigma,S})$, since $\mathrm{d}_S v \in \mathcal{R}(\mathrm{d})$ implies path independence inside S. As a result, the marginal value $v\,(S \cup C) - v(S)$ is the same for all coalitions $C \subset N \setminus S$. Taking $C = \emptyset$, we see that this value is precisely $v(S)$, and we conclude that v is inessential relative to S. Conversely, suppose that v is inessential relative to S, and define the game

$$v_S(C) = \begin{cases} v\,(S \cap C) & \text{if } S \cap C \neq \emptyset, \\ 0 & \text{if } S \cap C = \emptyset. \end{cases}$$

It follows immediately that $\left(\sum_{i \in S} di\right) v = dv_S \in \mathcal{R}(\mathrm{d})$, which completes the proof.

Corollary 1: v is inessential iff $r_i = 0$ for each $i \in N$. $\sum_{i \in N} ||r_i||^2$ characterizes the deviation from inessentiality of v.

Corollary $\mathbf{1_S}$: v is inessential relative to S iff $r_S = \sum_{i \in S} r_i = 0$. $||\sum_{i \in S} r_i||^2$ characterizes the deviation from inessentiality of game v relative to S.

Lemma: $\sum_{i \in N} r_i = 0$. Proof: by definition, $\sum_i \mathrm{d}_i = \mathrm{d}$. Since we also know that $\sum_i v_i = v$, summing $\mathrm{d}v_i + r_i = \mathrm{d}_i v$ over all i directly yields the result.

7.2. Proofs from Section 4

Lemma 1:

Proof. Using the linearity of expectation, we can rewrite this game as

$$v_{f,z}^{Kernel}(T) = E[f([\boldsymbol{z}_T, \boldsymbol{X}_{\bar{T}}])] = \begin{cases} g(\boldsymbol{z}_S) + E[h([\boldsymbol{z}_{T \setminus S}, \boldsymbol{X}_{\bar{T}}])] & S \subseteq T \\ E[g(\boldsymbol{X}_S)] + E[h([\boldsymbol{z}_T, \boldsymbol{X}_{\bar{T} \setminus S}])] & T \cap S = \emptyset \end{cases}$$

Now we can write the nonzero elements of the partial derivative $d_S v_{f,z}^{Kernel}$ as

$$\begin{aligned} v_{f,z}^{Kernel}(T \cup S) - v_{f,z}^{Kernel}(T) &= g(\boldsymbol{z}_S) + E[h([\boldsymbol{z}_{(T\cup S)\setminus S}, \boldsymbol{X}_{T\bar{\cup}S}])] - \big(E[g(\boldsymbol{X}_S)] + E[h([\boldsymbol{z}_T, \boldsymbol{X}_{\bar{T}\setminus S}])]\big) \\ &= g(\boldsymbol{z}_S) + E[h([\boldsymbol{z}_T, \boldsymbol{X}_{\bar{T}\setminus S}])] - \big(E[g(\boldsymbol{X}_S)] + E[h([\boldsymbol{z}_T, \boldsymbol{X}_{\bar{T}\setminus S}])]\big) \\ &= g(\boldsymbol{z}_S) - E[g(\boldsymbol{X}_S)] \end{aligned}$$

regardless of T, as long as $T \cap S = 0$.

□

Lemma 2:

Proof. Using the linearity of expectation, we can rewrite this game as

$$\begin{aligned} v_{f,z}^{Cond}(T) &= E[f(\boldsymbol{X})|\boldsymbol{X}_T = \boldsymbol{z}_T] \\ &= E[g(\boldsymbol{X}_S)|\boldsymbol{X}_T = \boldsymbol{z}_T] + E[h(\boldsymbol{X}_{\bar{S}})|\boldsymbol{X}_T = \boldsymbol{z}_T] \\ &= \begin{cases} g(\boldsymbol{z}_S) + E[h(\boldsymbol{X}_{\bar{S}})|\boldsymbol{X}_{T\setminus S} = \boldsymbol{z}_{T\setminus S}] & S \subseteq T \\ E[g(\boldsymbol{X}_S)] + E[h(\boldsymbol{X}_{\bar{S}})|\boldsymbol{X}_T = \boldsymbol{z}_T] & T \cap S = \emptyset \end{cases} \end{aligned}$$

Now we can write the nonzero elements of the partial derivative $d_S v_{f,z}^{Kernel}$ as

$$\begin{aligned} v_{f,z}^{Kernel}(T \cup S) - v_{f,z}^{Kernel}(T) &= g(\boldsymbol{z}_S) + E[h(\boldsymbol{X}_{\bar{S}})|\boldsymbol{X}_T = \boldsymbol{z}_T] - (E[g(\boldsymbol{X}_S)] + E[h(\boldsymbol{X}_{\bar{S}})|\boldsymbol{X}_T = \boldsymbol{z}_T]) \\ &= g(\boldsymbol{z}_S) - E[g(\boldsymbol{X}_S)] \end{aligned}$$

regardless of T, as long as $T \cap S = 0$.

□