WHO ARE PLAYING THE GAMES?

Anonymous authors
Paper under double-blind review

ABSTRACT

The Shapley value has been widely used as the measures of feature importance of a predictive model, by treating a model as a cooperative game (N, v). There have been many discussions on what the correct characteristic function v should be, but almost all literature will take the player set N as the set of features. While in classical cooperative game scenarios, players are obvious and well defined, it is not clear whether we should treat each feature individually as a player in machine learning. In fact, adding or deleting a feature, even a redundant one, will change every feature's Shapley value and its rank among all features in a non-intuitive way. To address this problem, we introduce a new axiom called "Consistency", which characterizes the "robustness" of computed Shapley-like values against different player set identifications, and is specific to machine learning setup. We show that while one can achieve Efficiency and Consistency in special cases, such as inessential games and 2-player games, they are contradictory to each other in general. This impossibility theorem signifies a conundrum of applying Shapley values in the feature selection process: The Shapley value is only axiomatically desirable if the players(features) are correctly identified, however, this prerequisite is exactly the purpose of the feature selection task. We then introduce the GroupShapley value to help address this dilemma, and as an additional bonus, GroupShapley valuess have a computational advantage over the classical Shapley values.

1 INTRODUCTION AND BACKGROUND

1.1 COALITION GAME AND MACHINE LEARNING

There have been many efforts to utilize concepts from cooperative game theory to machine learning to quantify the feature importance, many are focused on Shapley values(see, for example, Lundberg & Lee (2017); Lundberg et al. (2018); Sundararajan & Najmi (2020); Merrick & Taly (2020); Frye et al. (2020b)). More precisely, consider a machine learning model $f : \mathbb{R}^p \to \mathbb{R}$, that is used to predict a target Y using the p-dimensional feature $X = [X_1, X_2, \dots X_p]$, we can associate a cooperative game (N, v) to it. In almost all the aforementioned literature, the player set is chosen to be the set of features $N = [p] = \{1, 2, \dots p\}$, i.e. $i \in N$ is thought of as the feature X_i. Then, we can compute Shapley valuess $\text{Sh}_v : [p] \to \mathbb{R}$ of the game $([p], v)$, and use $\text{Sh}_v(i)$ as the measure of the importance of the feature X_i.

Depending on the choice of v, Sh_v can be interpreted as *global* feature importance or *local* feature importance. Most of our paper's discussion doesn't depend on the choice of v, and applies to both, thus, we don't distinguish them. In the case of local feature importance, the choice of v has been discussed by many(see, for example, Sundararajan & Najmi (2020); Janzing et al. (2020); Frye et al. (2020a), but fewer on how to "correctly" choose the player set N. In fact, Kumar et al. (2020) points out that $N = [p]$ is problematic, and along with other oppositions, they argue that "Shapley-value-based explanations for feature importance fail to serve their desired purpose in general". In this paper, we try to address the problem of correctly identifying the set of players. We will also discuss the oppositions to Linearity in Remark 2.

1.2 MOTIVATION

To motivate the importance of "correctly" identifying the player set in a machine learning model, let's start with an example of business.

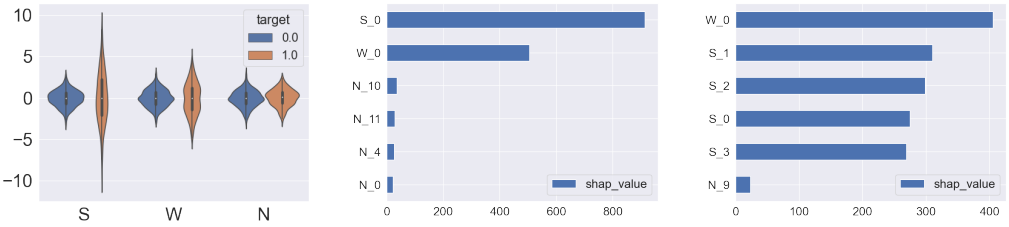
Example 1 (Vanishing Effect in Game Theory)**.** In a game $(\{o, w_1\}, v)$, where an owner o provides the capital, and a worker w_1 provides the labor, and $v(S) = \max(|S| - 1, 0)$. Each player has a Shapley value of $\frac{1}{2}$, same as the intuition. Now, as the global economy gradually falls into a recession, there are more workers on the labor market, and subsequently joining the team, and we end up with the player set $\{o, w_1, w_2, \ldots, w_m\}$. However, it turns out that our business has such a small capacity, that the increase of labor doesn't improve its profits: $v(S) = 1$ if and only if $|S| > 1, o \in S$, and $v(S) = 0$ otherwise. The owner will have Shapley value $\text{Sh}(o) = 1 - \frac{1}{m+1}$, and each of the worker $\text{Sh}(w_i) = \frac{1}{m(m+1)}$.

Example 1 makes perfect economical sense, as the owner is the veto player, and all the workers are replaceable, hence the share of the worker drops significantly as more of its peer joining the team. However, we will see it is not desirable in machine learning, especially when we use Shaply values to conduct feature selection.

Example 2 (Vanishing Effect in Machine Learning)**.** To demonstrate that the same Vanishing Effect in machine learning is not desirable, here we offer an example on synthetic data (X, y) , where the ground truth is known, and each component of X falls into 3 categories: strong predictor (S), weak predictor (W), and non predictor(N):

$$p_S(x) = \frac{\mathcal{N}[0,1](x) + \mathcal{N}[0,3](x)}{2}, \quad p_W(x) = \frac{\mathcal{N}[0,1](x) + \mathcal{N}[0,2](x)}{2}, \quad p_N(x) = \mathcal{N}[0,1](x),$$

see Figure 1a for a graphical representation of their distributions. More precisely, we generate 10000 data points, each consisting of 20 real-valued features, and we split the data into 50% class 1, and 50% class 0. We consider two cases, in the first case, class 0 follows a 20-dim standard normal distribution $p_0(x_0, \ldots x_{19}) = \prod_{i=0}^{19} \big(\mathcal{N}[0,1](x_i)\big)$, and class 1 follows $p_1(x_0, \ldots x_{19}) = \big(\mathcal{N}[0,3](x_0)\big)\big(\mathcal{N}[0,2](x_1)\big)\prod_{i=2}^{19}\big(\mathcal{N}[0,1](x_i)\big)$, that is , X_0 is a S predictor, X_1 is a W predictor, and the rest are N predictors, and there is no redundancy in the first case. In the second case, we introduce redundancy to S predictors by replacing three N predictors $\{X_i\}_{i\geq 17}$ with copies of the S predictor X_0. In both cases, we fit a RandomForestClassifier, and compute Shapley values for all features using TreeShap as in Lundberg et al. (2018). In Figure 1b, we see that S predictor gets more than twice as much attribution as that of the W predictor in the first case. But all copies of S predictor X_0 receive less attributions than the W predictor in the second case, as shown in Figure 1c. In Proposition 1, we show that this kind of vanishing effect holds for any games.



(a) Distribution of features by class (b) First case: one strong feature, one weak feature (c) Second case: four redundant strong features, one weak feature

Figure 1: Vanishing Effect of adding redundancy to the strong features as in Example 2.

As surveyed in Fryer et al. (2021), many Shapley-value based feature selection framework follows a variation of Algorithm 1. Implicitly in this framework is the identification of the player set with the feature set (which we write out explicitly), and the Vanishing Effect as in Example 2 will lead this framework to the suboptimal selection of features. Besides the Vanishing Effect, adding features (and treating them as individual players) not only will change Shapley values of features redundant to them, but also change Shapley values of all other features, in a non-intuitive way, see an example of a 3-player game in Kumar et al. (2020).

We want to stress that an appropriate preprocess needs to be done, before interpreting the absolute Shapley-like values of features as a measure of importance. Because we see a recent paper Masoomi et al. (2021) uses the absolute value of a Shapley-like quantity to measure the redundancy between

features. More precisely, Masoomi et al. (2021) introduces Bivariate Shapley value $E(v)_{ij}$, and defines directional redundancy of i with respect to j as $E(v)_{ij} = 0$, but suggests one can use $E(v)_{ij} < \gamma$ for some threhsold γ in practice. $E(v)_{ij}$ is a Shapley-like quantity, one can think of it as the Shaply value of i under the presence of j, and it also suffers from the same Vanishing Effect, that is , under the presence of many features redundant to i, we would get $E(v)_{ij} < \gamma$ regardless of whether i is truly directional redundant with respect to j. It doesn't mean that the Shapley-value-based feature selection framework or the Bivariate Shapley value $E(v)_{ij}$ is useless, rather, it emphasizes the necessity of carefully identifying the player set before utilizing these techniques, especially when facing a large pool of features. For example, it has been known that practitioners in financial machine learning would generate thousands of features ("alphas") automatically Zhang et al. (2020), many of which doesn't make sense to humans, even if they do, the sheer number of features will be overwhelming for people to study them individually. Moreover, there could be strong correlations among certain features, such as past returns over various window lengths, as well as those build on them. In such circumstances, it is imperative to make appropriate player identifications before diving into the Shapley-values-based explanation methodologies.

Algorithm 1 Shapley-value-based Feature Selection Algorithm

Input: Training Data $(X, y), X = [X_1, \dots X_p] \in \mathbb{R}^{n\times p}, y \in \mathbb{R}^n$, and a learning algorithm A.
Output: A selected subset of k features $\{X_{i_j}\}_{j=1,\dots k}$.
1: Set the player set $P = \{1, 2, \dots p\}$ to be the set of features.
2: Choose the characteristic function $v : 2^P \to \mathbb{R}$.
3: Compute the classical Shapley values $(\text{Sh}_i)_{i=1,2\dots p}$ for the game (P, v).
4: Select feature i, if Sh_i is among the highest k Shapley values in $\{\text{Sh}_i\}_{i=1,2,\dots p}$.

Our first attempt to address this problem is to introduce a Shapley-like value, that is more robust against different identifications of players, an axiom we call "Consistency" (made precise in Section 2.3). Unfortunately, while we can achieve both Efficiency and Consistency in some special cases, such as inessential games and 2-player games, Theorem 1 shows we cannot have both in general. Hence, Theorem 1 signifies the contradiction between Efficiency and Consistency: One cannot get Efficiency among players, without giving up the flexibility of determining who are the actual players. However, without either one, the other would also be rendered useless, for example, Shapley-value is Efficient, but can be misleading due to Vanishing Effect, which results from its lack of Consistency. We want to note that Consistency has never been brought up in the classical game theory, because, as Example 1 shows, the Vanishing Effect, due to lack of Consistency, is entirely natural. However, in the application of Shapley values in machine learning, both are important, and Consistency is an axiom specific to the machine learning setup. As Efficiency is highly desirable, we need to give up Consistency, as the prevalent studies have been doing. But we will do so with a twist, that is we need to carefully choose the player set first, and that's our second attempt: GroupShapley values. To motivate Consistency and GroupShapley values, let's again start with a real-world problem.

Example 3 (Confusing Cost-Sharing/Profit-Attribution)**.** Suppose at the end of the year, a fund made 10M dollars in net profits, utilizing a production model f, a real-valued function, that takes in a vector of alphas (features) $X = [X_1, X_2, \dots X_n]$, which are contributed by different researchers, i.e., each researcher contribute a subset of $\{X_i\}_{i=1,2,\dots n}$ to the production model. The fund owner, with a heart of gold, would like to distribute the profits fairly. This cost-sharing scenario is a perfect problem for Shapley values, as all of the axioms, even Linearity, is non-controversial. However, the problem remains, on what set of players should we apply Shapley values? It could be the set of features (alphas); or the set of individual researchers, which corresponds to a group of alphas it contributed, since none of these alphas would exist without the presence of the researcher; or the set of teams consisting of researchers, that would correspond to a even coarser group of alphas.

Due to the lack of Consistency, these Shapley-value-based attributions would look different (for example, a researcher's Shapley value may not be the sum of its alphas' Shapley values). However, this does motivate us to consider a group of features as a player, and compute the GroupShapley value as the contribution of the group of features. While we don't have domain knowledge as in Example 3, we can determine groups using statistical methods, such as clustering. As an additional bonus, by grouping features as players, we also decrease the number of players in the coalition game, which will help to accelerate the computations for Shapley values, another well-known challenge.

1.3 Related Literature

1.3.1 Problems with Shapley values

Kumar et al. (2020) has brought up many oppositions to the Shapley values as measures of feature importance, among which they mentioned the difficulties of identifying player set from the feature set, but they don't formalize this difficulty, nor gave any solutions to address this difficulty. In fact, they argue that Shapley values are not appropriate for feature explanations, we will discuss some of their other oppositions in Remark 2 as well. Fryer et al. (2021) discusses some danger of the Shapley-value-based feature selection framework 1, but they don't inspect this framework from the point of view of player set identification, nor do they provide the solutions. Sundararajan & Najmi (2020); Slack et al. (2020); Frye et al. (2020a) all raise concerns about certain out-of-data characterization function v when associating a model to a coalition game (N, v). None of them mentioned the first component of the game (N, v), i.e., the player set N, as they all identify the player set as the set of the features. Kumar et al. (2021) proposes Shapley Residuals to measure the limits of Shapley values for explanations, our Theorem 1 has an implication that overlaps with Shapley Residual's non-warning on the inessential games, but they do differ in other cases, see Remark 1 for more discussions.

1.3.2 Measuring the importance of a subset of features

Sundararajan et al. (2020) and Harris et al. (2021) both extend Shapley values to a subset of features, where they each propose Shapley-like values, unique to a set of axioms, that are originated from intuitions on classical game theory. In fact, they are special cases of a long line of work on the axiomatic approach on interaction indices in classical game theory, such as Grabisch & Roubens (1999); Owen (1972). Our Consistency axiom differs from axioms in all of these work, in that Consistency is specific to machine learning setup, i.e., it only makes intuitive sense when the game theory is applied to machine learning, especially in the feature selection stage (see the differences between Example 1 and Example 2, and the discussion in between). Jullum et al. (2021) studies GroupShapley values, however, their motivation is to address the inefficiency when computing Shapley values of a model with many features, without studying the conceptual motivation of GroupShapley values. While our work demonstrates, both in theories and examples, that certain preprocess, such as grouping(clustering) is not only a compromise, but also a necessity.

1.4 Our Contributions

In summary, our contributions are:

1. We demonstrate the Vanishing Effect of Shapley values both in theories and examples, stressing the problem of identifying the "correct" player set.

2. We introduce the Consistency axiom, an axiom specific to machine learning, that characterizes the robustness against different identifications of the player set. We also characterize the unique family of Shapley-like values that satisfies Efficiency or Consistency respectively, and shows while we can achieve both Efficiency and Consistency in certain special cases, they are contradictory to each other in general.

3. In the general case, we propose the Group-Shapley-value framework as a prerequisite to the Shapley value computations, due to the lack of Consistency. As an additional bonus, GroupShapley values have a computational efficiency over the classical Shapley values.

The rest of the paper is structured as follows. We formalize and prove the Vanishing Effect in Section 2.2 after reviewing game-theoretical Shapley values. In Section 2.3, we extend the Shapley values, as well as its characterizing axioms to all subsets of features, and introduce the novel Consistency axiom. We also give unique characterizations of G-Efficiency and G-Consisntecy in 2.3. Finally, we propose GroupShapley values in Section 3, and present the results on the Boston housing dataset.

2 SHAPLEY VALUES AND CONSISNTECY AXIOM

2.1 NOTATIONS

A coalition game consists of a set $N = \{1, 2, \dots n\}$ of n players, and a characteristic function $v : 2^N \to \mathbb{R}$. A player $i \in N$ is a null player if $v(S \cup \{i\}) = v(S), \forall i \notin S \subset N$, and two players i, j are symmetric, if $v(S \cup \{i\}) = v(S \cup \{j\}), \forall S \subset N \backslash \{i, j\}$. We write $j \sim i$ if j is a redundant player of i, that is, i, j are symmetric, and $v(S \cup \{i, j\}) = v(S \cup \{j\}) = v(S \cup \{i\}), \forall S \subset N \backslash \{i, j\}$. When N is clear from the context, we write $\bar{S} = N \backslash S$.

The reduced game with respect to $T \subset N$ of $G = (N, v)$ is defined as $G_{[T]} = (N \backslash T \cup [T], v_{[T]})$, where $v_{[T]}(S) = v(S)$, and $v_{[T]}(S \cup [T]) = v(S \cup T), \forall S \cup T = \emptyset$, i.e., we group the subset of players T as a single player $[T]$. The reduced game with respect to a partition $\Pi : \bigsqcup_{i=1}^{k} S_i = N$ is defined as $G_\Pi = (\{[S_i]\}_{i=1}^{k}, v_\Pi)$, where $v(\{[S_{i_1}], \dots [S_{i_l}]\}) = v(\cup_{j=1}^{l} S_{i_j})$. For any bijection π between N and N', we define the induced game on N' as $G^\pi = (N', v^\pi)$, where $v^\pi(\pi(S)) = v(S)$ for all $S \subset N$. For two games (N, v) and (N, w), we define their linear combination as $a(N, v) + b(N, w) = (N, av + bw), \forall a, b \in \mathbb{R}$. It is well known that $\{(N, v)\}_{v:2^N \to \mathbb{R}}$ is a linear space with a basis consisting of simple games $\{G_{N,T} = (N, v_{N,T})\}_{T \subset N}$, where $v_{N,T}(S) = 1$ if $T \subset S$, and $v_{N,T}(S) = 0$ otherwise. We denote Shapley values of a game $G = (N, v)$ as Sh_G or Sh_v when N is clear from the context.

For any finite N, we define $\mathcal{G}_N$ to be the set of all games whose player set is N, hence we can identify $\mathcal{G}_N$ with the set of characterization functions $v : 2^N \to \mathbb{R}$. For $n \in \mathbb{N}$, we define $\mathcal{G}_n$ as the set of all games whose player set $|N| = n$, and $\mathcal{G} = \bigsqcup_{i \in \mathbb{N}} \mathcal{G}_i$. A game (N, v) is inessential if $v(S \cup i) = v(S) + v(i)$ for all $i \notin S \subset N$. Inessentiality is like the game-theoretical "independence" among the players, there is no synergy created by any coalition of players. We denote $\mathcal{G}_{\text{in}}$ to be the set of all inessential games. We say a subset $\mathcal{H} \subset \mathcal{G}$ is closed (under reductions and linear combinations), if $G_\Pi \in \mathcal{H}$ for all partitions Π of N, whenever $G = (N, v) \in \mathcal{H}$, and $aG_1 + bG_2 \in \mathcal{H}$ for all $a, b \in \mathbb{R}$, whenever $G_1, G_2 \in \mathcal{H} \cap \mathcal{G}_N$, for some N. Examples of closed subsets of $\mathcal{G}$ include $\bigsqcup_{i=1}^{n} \mathcal{G}_i$ and $\mathcal{G}_{\text{in}}$. For any game $G = (N, v)$, we define the closure of G as $\bar{G} = \bigcap_{\substack{G \in \mathcal{H} \subset \mathcal{G} \\ \mathcal{H} \text{ closed}}} \mathcal{H}$. The closedness is needed in defining the Consistency axiom.

For a machine learning model $f : \mathbb{R}^p \to \mathbb{R}$, we denote $G_f = (P, u)$ to be the game corresponding to f, by choosing $N = P$, and some characterization function u. In this paper, we don't care about which u is chosen.

2.2 GAME-THEORETICAL SHAPLEY VALUES

Shapley values $\text{Sh}_v : N \to \mathbb{R}$ of (N, v) is the unique "fair" allocation of the value $v(\{1, 2, \dots N\})$ of the grand coalition. More precisely, $\{\text{Sh}_v : N \to \mathbb{R}\}_{v \in \mathcal{G}_N}$ is the unique set of functions that satisfies the following axioms:

1. **Efficiency**: $\sum_{i=1}^{n} \text{Sh}_v(i) = v(\{1, 2, \dots n\})$;
2. **Anonymity** For any permutation π of N, we have $\text{Sh}_{v^\pi}(\pi(i)) = \text{Sh}_v(i)$.
3. **Symmetry** If i, j are symmetric, then $\text{Sh}_v(i) = \text{Sh}_v(j)$. Notice that Symmetry is a consequence of Anonymity by taking $\pi = (i, j)$.
4. **Null Player** $\text{Sh}_v(i) = 0$, if i is a null player of v.
5. **Linearity** $\text{Sh}_{av+bw} = a\text{Sh}_v + b\text{Sh}_w$.

In fact, $\text{Sh}_v(i)$ is computed as a weighted average of its marginal contributions $\Delta_v(i, S) = v(S \cup \{i\}) - v(S)$ of player i with respect to all the subsets $S \subset N \backslash \{i\}$:

$$\text{Sh}_v(i) = \sum_{S \subset N} \frac{s!(n - s - 1)!}{n!} \Delta_v(i, S),$$

where s is the cardinality of S. One can think of $\text{Sh}_v(i)$ as the expected marginal contribution of player i if all players arrive randomly, where each order is equally likely.

The following Proposition makes the Vanishing Effect in Section 1 precise.

Proposition 1 (Vanishing Effect)**.** In a game (N, v), if player i's marginal contribution is bounded by γ, i.e., $|\Delta_v(i, S)| \leq \gamma$, for all $i \notin S \subset N$, and there are k redundant features of i in the player set N, denote $R_i = \{j | j \in N, j \sim i, j \neq i\}$. Then $\text{Sh}_v(i) \leq \frac{\gamma}{k+1}$.

Proof. By definition $\text{Sh}_v(i) = \frac{1}{n!}\sum_\pi \Delta_v(i, S_{\pi,i})$, where $S_{\pi,i} = \{\pi(j) < \pi(i), j \in N\}$, and the sum runs over all permutations π of N. By symmetry, i(or any player in R_i) is equally likely to be the first one among $R_i \cup \{i\}$ in π, hence at most $\frac{1}{k+1}n!$ of $\Delta_v(i, S_{\pi,i})$ are nonzero:

$$\text{Sh}_v(i) = \frac{1}{n!}\sum_\pi \Delta_v(i, S_{\pi,i}) = \frac{1}{n!}\sum_{\substack{\pi(i)<\pi(j),\\ \forall j, j\in R_i}} \Delta_v(i, S_{\pi,i}) \leq \frac{1}{n!}\frac{n!}{k+1}\gamma = \frac{1}{k+1}\gamma.$$

□

Note that γ is finite in any finite game, Proposition 1 states that if we keep adding redundant features of i in a finite game (N, v), then $\text{Sh}_v(i)$ will tend to zero.

2.3 Shapley values in Machine Learning and Consistency

As discussed in Section 2.2, given a model $f : \mathbb{R}^p \to \mathbb{R}$, to compute Shapley values of its corresponding game, we need to first determine the player set N. As shown in Example 3, it is likely that a player is a subset of features, and they collectively form a partition of the feature set P. Our first attempt to solve the identification of the player set problem is still to identify $N = P$, but we want to find Shapley-like values, that would be robust to the different identifications of player set. More precisely, given a game $G = (N, v)$, we want to extend the function $\text{Sh}_v : N \to \mathbb{R}$ to a function $\phi_G : 2^N \to \mathbb{R}$, and interpret $\phi_G(S)$ as the feature importance of the subset of features $S \subset N$, regardless of how the rest $\bar{S}$ of features are clustered. We will extend the axioms in Section 2.2, so that they are relevant to a collection of functions $\{\phi_G : 2^N \to \mathbb{R}\}_{G\in\mathcal{H}}$ indexed by some closed subset $\mathcal{H} \subset \mathcal{G}$, and formalize Consisntency axiom. When $S = \{i\}$ is a singleton, we will write $\phi_v(S)$ as $\phi_v(i)$ for simplicity.

1. **G-Efficiency**: $\forall G \in \mathcal{H}$, and $\forall \Pi : N = \bigsqcup_{i=1}^k S_i$, we have $\sum_{i=1}^k \phi_G(S_i) = v(N)$;
2. **G-Anonymity**: $\forall G, G_\pi \in \mathcal{H}$, for some bijection $\pi : N \to N'$, we have $\phi_G(S) = \phi_{G^\pi}(\pi(S)), \forall S \subset N$.
3. **G-Null Player** $\phi_G(i) = 0$, if i is a null player in G.
4. **G-Linearity** $\forall G_1, G_2 \in \mathcal{H} \cap \mathcal{G}_N$ for some N, we have $\phi_{aG_1+bG_2} = a\phi_{G_1} + b\phi_{G_2}$.
5. **G-Consistency**. $\forall G = (N, v) \in \mathcal{H}, T \subset N$, we have $\phi_{G_{[T]}}(S) = \phi_G(S)$, and $\phi_{G_{[T]}}(S \cup [T]) = \phi_G(S \cup [T]), \forall S \subset \bar{T}$.

Consistency implies that it doesn't matter how we identify the players(partitioning the features into different groups, and treat each group as a player), we always get an attribution of each player from a consistent principle ϕ. Unlike Example 2, there is no Vanishing Effect if ϕ_v is consistent. To see that, let's assume we already added a redundant feature i_1 of i, and we end up with the game (N, v). Suppose we add another feature $i_2 \sim i$, and get $(N \cup \{i_2\}, \tilde{v})$. We can group i_1, i_2 to get the reduced game $(N \cup \{[i_1, i_2]\}\backslash\{i_1\}, \tilde{v}_{[i_1,i_2]})$, this is the same as the game (N, v), where $[i_1, i_2]$ plays the role of i_1 in (N, v). Hence, we have

$$\phi_v(i) = \phi_{\tilde{v}_{[i_1,i_2]}}(i) = \phi_{\tilde{v}}(i), \tag{1}$$

where the first equality is due to Anonymity, and the second due to Consisntecy. Equation 1 demonstrates that keep adding redundant features of i won't affect i's attribution after the first addition. Moreover, a similar argument would show that $\phi_v(j)$ is also unaffected by the addition of redundant feature of i, if $i \neq j$. It is clear that the original characteristic function $v(S)$ is consistent, we will see that there aren't many others, and it turns out, in general, we can only have one of the G-Efficiency and G-Consistency, but not both.

Theorem 1. Suppose $\mathcal{H} \subset \mathcal{G}$ is closed, and $\mathcal{G}_N \subset \mathcal{H}$ if $G = (N, v) \in \mathcal{H}$, and a collection of functions $\{\phi_G : 2^N \to \mathbb{R}\}_{G\in\mathcal{H}}$ is indexed by $\mathcal{H} \subset \mathcal{G}$, then:

$\phi_G(S) = \sum_{i\in S} \text{Sh}_G(i)$ is the only $\{\phi_G\}_{G\in\mathcal{H}}$ that satisfies Axioms 1-4.

$\{\{\phi_G(S) = av(S) + b\bar{v}(S)\}_{G\in\mathcal{H}} | a, b \in \mathbb{R}\}$, where $a, b \in \mathbb{R}$, and $\bar{v}(S) = v(N) - v(\bar{S})$, is the only family of $\{\phi_G\}_{G\in\mathcal{H}}$ that satisfies Axioms 2-5.

Proof. It is straightforward to check that $\phi_G(S) = \sum_{i\in S} \text{Sh}_G(i)$ satisfies Axioms 1-4, and $\phi_G(S) = av(S) + b\bar{v}(S)$ satisfies Axioms 2-5, hence, only the "uniqueness" part needs proof. For the first part, suppose $\{\phi_G : 2^N \to \mathbb{R}\}_{G\in\mathcal{H}}$ satisfies Axioms 1-4, then the restriction $\{\phi_G|_N : N \to \mathbb{R}\}_{G\in\mathcal{G}_N}$ would be the classical Shapley values, for Axioms 1-4 reduces to their classical counterpart, hence $\phi_G(i) = \text{Sh}_G(i)$. Then for any $S \subset N$, we consider the partition $N = S \bigsqcup (\bigsqcup_{i\in\bar{S}} \{i\})$, by G-Efficiency of ϕ_G, and Efficiency of Sh_G, we have:

$$\sum_{i\in\bar{S}} \phi_G(i) + \phi_G(S) = v(N) = \sum_{i\in N} \text{Sh}_G(i) = \sum_{i\in N} \phi_G(i).$$

This completes the proof that $\phi_G(S) = \sum_{i\in S} \text{Sh}_G(i)$. To prove the second part, we follow the common technique by noting that the simple games $\{G_{N,T} = (N, v_{N,T})\}_{T\subset N}$ form a basis of $\mathcal{G}_N$. Let's start by showing the second part for the simple games $G_1 = G_{N,T} = (N, v_1)$. For any $S \subset N$, such that $T \subset S$, we can reduce G_1 to a 2-player game $G_2 = (\{[S], [\bar{S}]\}, v_2)$. By G-Null Player, we know that $\phi_{G_1}(\bar{S}) = \phi_{G_2}([\bar{S}]) = 0$, and there exists $a \in \mathbb{R}$, such that $\phi_{G_1}(S) = \phi_{G_2}([S]) = a$. Note that we cannot claim $a = 1$, due to the lack of G-Efficiency, but we do know the existence of a, i.e., $\phi_{G_1}(S_1) = \phi_{G_1}(S_2) = a$ for any $S_1, S_2 \supset T$, due to G-Anonymity. Similarly, there exists $b \in \mathbb{R}$, such that $\phi_{G_1}(S) = b$ for all $S \subset N$, such that $S \cap T \notin \{\emptyset, T\}$, due to the Anonymity and Symmetry. Hence, we have $\phi_{G_1} = (a - b)v_1 + b\bar{v_1}$. By reducing any simple game $G_{N,T}$ to the 2-player simple game as in the above proof, we can show $\phi_{G_{N,T}} = (a - b)v_{N,T} + b\bar{v_{N,T}}$ for all $G_{N,T}$ by Consistency and Anonymity. This completes the proof. □

In general, $\sum_{i\in S} \text{Sh}_G(i)$ is not a linear combination of v and $\bar{v}$, hence Theorem 1 demonstrates that there is no $\{\phi_G\}_{G\in\mathcal{G}}$ that satisfies Axioms 1-5. But there are some special cases, that we can achieve both.

Corollary 1. $\mathcal{H} \subset \mathcal{G}$ is a closed subset, if there are $a, b \in \mathbb{R}$, such that $\sum_{i\in S} \text{Sh}_G(i) = av(S) + b\bar{v}(S)$ for all $G = (N, v) \in \mathcal{H}, S \subset N$, then $\phi_v(S) = av(S) + b\bar{v}(S)$ satisfies Axioms 1-5. This Assumption holds when:

$\mathcal{H} = \mathcal{G}_{\text{in}}$, take $a = 1, b = 0$;

$\mathcal{H} = \mathcal{G}_1 \bigsqcup \mathcal{G}_2$, take $a = b = \frac{1}{2}$.

Note that Axioms 1-5 are not stated for all of the games in $\mathcal{G}$, because, as shown in Corollary 1, there are certain special $\mathcal{H} \subset \mathcal{G}$, where Axioms 1-5 can all be achieved. Similarly, we don't state the uniquess of Shapley values in Section 2.2 for $\mathcal{G}$, but only to its "minimal domain" $\mathcal{G}_N$. We don't do this for the most generality, instead, we have a practical concern: When working with a machine learning model f, we are mainly concerned about G_f, as well as its reductions in $\bar{G_f}$. We don't care if G_f is consistent with some irrelevant classical game. But we care about its reduction $G_{f,\Pi}$, because we may treat each subset in Π as a player. Hence, the interesting question is, can we find $\{\phi_H\}_{H\in\bar{G_f}}$, that satisfies Axioms 1-5. Corollary 1 says yes if G is inessential or a 2-player game. One difficulty to solve this question in general is that the mathematical convenience brought by G-Linearity disappears, as the simple games may not lie in $\bar{G_f}$. We will consider this as an interesting future work:

Question 1. Given a game $G_f = (P, v)$, can we find Shapley-like values $\{\phi_H\}_{H\in\bar{G_f}}$, that statisfies Axioms 1-5?

Remark 1 (Shapley Residuals)**.** There is an interesting overlap between the Corollary 1 and the results of Kumar et al. (2021), where Shapley Residual, a characterization of inessential games, is introduced as a "warning" to the interpretation of the Shapley values. When G is inessential, Corollary 1 shows that a Shapley-value-based $\{\phi_H\}_{H\in\bar{G}}$ satisfies Axioms 1-5, while Shapley Residual in Kumar et al. (2021) is zero, hence no warning is issued. However, Corollary 1 and Shapley Residual do diverge, for example, in the case of a 2-player simple game $G_{2,N} = (N, v)$, where $v(S) = (N == S) * 1$, and $\text{Sh}(1) = \text{Sh}(2) = \frac{1}{2}$. Corollary 1 demonstrates that $\{\phi_H\}_{H\in\bar{G_{2,N}}}$

satisfies Axioms 1-5. However, $G_{2,N}$ is not inessential, and hence Shapley Residual is nonzero. It is hard to interpret what the warning is in this simple game where two symmetric players each get half of the credit. Shapley values are consistent not only in $\mathcal{G}_{\text{in}}$, but also in other cases, such as $\mathcal{G}_1 \cup \mathcal{G}_2$. We feel that the answer to Question 1 provides a measure of limit of Shapley values.

Remark 2 (Linearity Axiom)**.** In Subsection 3.2 of Kumar et al. (2020), it argues that the Linearity axiom in Shapley values is not as innocent as the other three axioms, and hence Shapley values are not entirely model-agnostic due to the restrictions posed by Linearity. This opposition to Linearity is also partly reflected in Question 1, where we don't want to restrict ourselves by asking for linear connections with possibly irrelevant simple games. However, the "unnatural example" in Subsection 3.2 of Kumar et al. (2020) is a consequence of Symmetry. To see that, the example is $f(x) = \prod_{j=1}^{d} x_d$, where the features are independent and centered at 0, but their magnitude(for example std) could differ. Using the local characteristic function $v_{f,x} = \mathbb{E}[f(X_S, X_{\bar{S}})|X_S = x_S]$, it is straightforward to see that the local explanation game $(D, v_{f,x})$ is a simple game, and $\text{Sh}(i) = \frac{1}{d} f(x)$ for all $i \in D$. Since the game $(N, v_{f,x})$ is a simple game, where all the features are symmetric, it doesn't require Linearity axiom or any other axioms other than Symmetry to deduce that all features have the same Shapley values.

Remark 3 (Single Inclusion and Permutation Importance)**.** In the Shapley-value-based feature seletion Algorithm 1, we choose the features with the highest $\text{Sh}(i)$, assumping $\text{Sh}(i)$ is the "correct" measure of importance. By Theorem 1, we can interpret $\text{Sh}(i)$ as choosing Efficiency over Consistency, if we choose Consistency over Efficiency, Algorithm 1 would turn into two most basic feature selection frameworks : Single Inclusion (v), and Permutation Importance ($\bar{v}$).

3 GROUPSHAPLEY VALUES

GroupShapley values is not entirely new, Marichal et al. (2007) already defines the concept of generalized Shapley values for a set of players, and Jullum et al. (2021) studies the GroupShapley values, to address the computational inefficiencies. We hope our results in Section 2 has convinced the readers that an appropriate identification of players, such as grouping(clustering) the features is not only a compromise to the computation inefficiency, but also a conceptual prerequisite.

Due to Theorem 1, in general, we cannot avoid the task of determining the set of players before applying Shapley-value techniques. How to determine the player set? This is where the domain knowledge of the problem comes in, for example, if one has the access to the causal graph of the underlying features, one can group all causal descendants to their ancestors. This is an approach suggested in Frye et al. (2020b). However, in may realistic problems, such as researchers with automatically generated financial features, we don't have informative prior knowledge about features. We can only rely on the data themselves, statistical methods such as clustering , PCA, can all be applied. One can also find the clustering of features by first computing Shapley values of G_f, as in Aas et al. (2021). Although this paper focuses on the theory, we will follow the clustering method suggested in Chapter 4 of de Prado (2020) to compute GroupShapley values for the Boston housing dataset Harrison Jr & Rubinfeld (1978).

Algorithm 2 GroupShapley values computation

Input: Training Data $(X, y), X = [X_1, \dots X_p] \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n$, and a learning algorithm A.
Output: A partition Π of feature set P, and the Shapley values Sh_{G_Π} computed for G_Π.

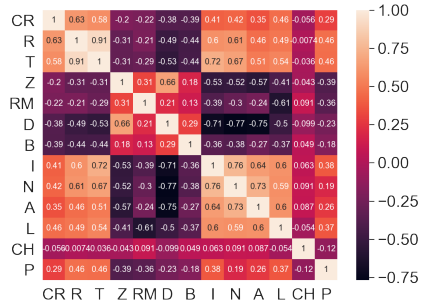
1: Compute the distance matrix of p features in X.
2: Choose the characteristic function $v : 2^P \to \mathbb{R}$.
3: Run K-means clustering algorithm for $k \in [minClusters, maxClusters]$, and use silhouette coefficient to find the optimal k, and the partition Π of P.
4: Compute Shapley Values for $G_{f,\Pi}$.

3.1 EXPERIMENT ON BOSTON HOUSE

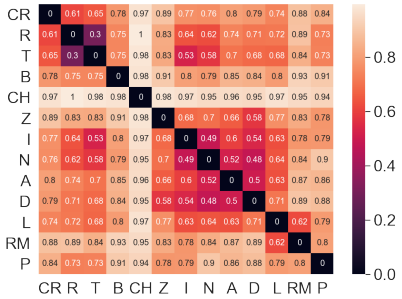
We split the data into 75% training data, and 25% test data. We follow Algorithm 2, in which we choose RandomForestRegressor as our learning algorithm A, and we compute distance matrix as $\sqrt{1 - |Corr|}$, where we compute $Corr$ in 2 different ways, either as the correlation matrix

$Corr_1$ of feature values in training data directly, or the correlation matrix $Corr_2$ of local Shapley explanations $\text{Sh}_{v_x}(i)$ on the training set, i.e., we first run TreeShap for G_f. Computing correlation this way doesn't have the bonus advantage on computational efficiency, but would be interesting to compare these two methods. We choose the characteristic function as R^2 on the test data, and set $minClusters = 3, maxClusters = 10$ for the K-means algorithm. Here are the results:

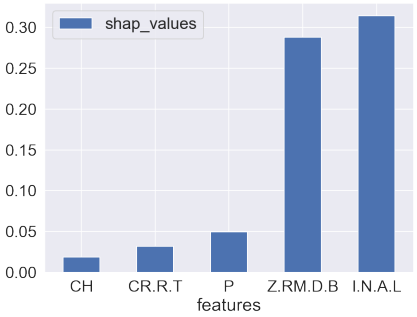
To fit in the graph, we rename features to their initials (or the first two letters in case of collisions), and a group of features is joined by a dot, for example, CR.R.T represents the group(subset) of features $\{CRIM, RAD, TAX\}$. Denote Π_i as the output partitions computed from $Corr_i$, for $i = 1, 2$. We see that while $Corr_1$ has clearer blocks than $Corr_2$, they actually give very similar partitions.



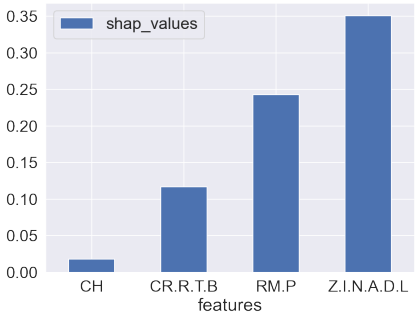
(a) Distance matrix of feature values



(b) Distance matrix of the local Shapley explanations on training data



(c) Shapley values for G_{f,Π_1}



(d) Shapley values for G_{f,Π_2}

Figure 2: Distance matrix and GroupShapley values for Boston housing data, computed in 2 ways.

4 Conclusion

We show the Vanishing Effect for Shapley values both in theories and examples, and demonstrate that it is imperative to properly identify the set of players before computing Shapley values for a machine learning model. To address this player set identification problem, we introduce a novel Consistency axiom, that characterizes the robustness against different identifications of player set, and is specific to the machine learning setup, especially the feature selection stage. We extend Shapley values to all subsets of features, and characterizes the unique family that satisfy G-Efficiency or G-Consistency respectively. Our results show that while G-Efficiency and G-Consistency can be achieved in special cases, such as inessential games, and 2-player games, but they are contradictory to each other in general, and an explicit identification of the player set is unavoidable. Finally, we suggest GroupShapley values as a partial solution to the player identification problem, and it also has a bonus advantage on the computational efficiency.

5 ETHICS STATEMENT

This paper contributes to the literature on explainable AI. It does not use human subjects; it only draws on publicly available datasets; the work has not been sponsored, and does not seek to promote any third organisations; none of the authors face any conflict of interest.

REFERENCES

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

Marcos M López de Prado. *Machine learning for asset managers*. Cambridge University Press, 2020.

Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. *arXiv preprint arXiv:2006.01272*, 2020a.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33:1229–1239, 2020b.

Daniel Fryer, Inga Strümke, and Hien Nguyen. Shapley values for feature selection: the good, the bad, and the axioms. *IEEE Access*, 9:144352–144360, 2021.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4):547–565, 1999.

Chris Harris, Richard Pymar, and Colin Rowat. Joint shapley values: a measure of joint feature importance. *arXiv preprint arXiv:2107.11357*, 2021.

David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pp. 2907–2916. PMLR, 2020.

Martin Jullum, Annabelle Redelmeier, and Kjersti Aas. groupshapley: efficient prediction explanation with shapley values for feature groups. *arXiv preprint arXiv:2106.12228*, 2021.

I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pp. 5491–5500. PMLR, 2020.

Indra Kumar, Carlos Scheidegger, Suresh Venkatasubramanian, and Sorelle Friedler. Shapley residuals: Quantifying the limits of the shapley value for explanations. *Advances in Neural Information Processing Systems*, 34:26598–26608, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

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

Jean-Luc Marichal, Ivan Kojadinovic, and Katsushige Fujimoto. Axiomatic characterizations of generalized values. *Discrete Applied Mathematics*, 155(1):26–43, 2007.

Aria Masoomi, Davin Hill, Zhonghui Xu, Craig P Hersh, Edwin K Silverman, Peter J Castaldi, Stratis Ioannidis, and Jennifer Dy. Explanations of black-box models based on directional feature interactions. In *International Conference on Learning Representations*, 2021.

Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 17–38. Springer, 2020.

Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5-part-2):64–79, 1972.

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*, pp. 180–186, 2020.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International conference on machine learning*, pp. 9259–9268. PMLR, 2020.

Tianping Zhang, Yuanqi Li, Yifei Jin, and Jian Li. Autoalpha: an efficient hierarchical evolutionary algorithm for mining alpha factors in quantitative investment. *arXiv preprint arXiv:2002.08245*, 2020.