Shapley values for feature selection: The good, the bad, and the axioms

Daniel Fryer[1], Inga Strümke[2], and Hien Nguyen[3]

***Abstract*—The Shapley value has become popular in the Explainable AI (XAI) literature, thanks, to a large extent, to a solid theoretical foundation, including four "favourable and fair" axioms for attribution in transferable utility games. The Shapley value is provably the only solution concept satisfying these axioms. In this paper, we introduce the Shapley value and draw attention to its recent uses as a feature selection tool. We call into question this use of the Shapley value, using simple, abstract 'toy' counterexamples to illustrate that the axioms may work against the goals of feature selection. From this, we develop a number of insights that are then investigated in concrete simulation settings, with a variety of Shapley value formulations, including SHapley Additive exPlanations (SHAP) and Shapley Additive Global importancE (SAGE).**

***Index Terms*—Shapley value, XAI, feature selection, interpretability, explainability, variable selection.**

I. Introduction

THE problem of feature selection in Machine Learning (ML) constitutes selecting some subset S of a set F of $|F| = d$ feature indices, such that the submodel formed from the features indexed by S will maximise some evaluation function $C(S)$ of the submodel, while minimising a cost (or complexity), which is increasing in $|S|$. The model chosen by this procedure is the *selected* model.

A similar (and more general) problem – model selection – has deep roots in computational statistics [1], where attention is paid to inferential nuances like quantification of uncertainty, significance testing, confounding predictors, collinearity, and the design of experiments. It was in this literature that the Shapley value was first applied to linear regression models, with its own history of discourse (see [2]–[7] and the more critical [8], which traces development to [9]–[11], with reinventions by [12] and [13]).

The Shapley value has, over recent years, become a popular method for interpretable feature attribution in fitted ML models (cf. [6], [14]–[26]), holding promise for the development of Explainable Artificial Intelligence (XAI). Attribution (or credit allocation), here, is the determination of the contribution by each feature to the performance of a model – often the selected model. The ML methods that stand out in terms of popularity are SHapley Additive exPlanations (SHAP) [15], [17], Shapley Effects [6] and Shapley Additive Global importancE (SAGE) [24], though the Shapley value itself carries a rich history of investigation in the context of game theory – Lloyd Shapley's 1953 seminal paper [27] has over 9000 citations, and the concept has attracted the attention of various Nobel prize winning economists [28]–[34].

Particular praise is given, in both the game theory and ML literature, to a small set of "favourable and fair" axioms, commonly known as *efficiency*, *null player*, *symmetry* and *additivity*, under which the Shapley value can be uniquely defined. We will introduce these axioms in section II-B. While much attention has been paid by game theorists and economists to interpreting and reformulating the axioms, and towards investigating axiom sets (and game formulations) that lead to Shapley value alternatives [28], [29], [31], we found comparatively little attention to these matters in ML [24], [35], [36]. There have, however, been a number of recent criticisms of the Shapley value in ML [37]–[39], suggesting to us that the Shapley value and its alternatives may be further developed considerably over the coming years.

This paper is not an exhaustive theoretical or empirical study of the worth of various Shapley value methods in ML, neither in general nor for feature selection. Our goal is to draw scrutiny towards the Shapley value axioms, and attention towards the generality of the game theoretic formulation. We do this in the specific context of feature selection, since this direction is particularly underdeveloped, and because in both academic and industrial settings we have encountered what we consider to be an over-reliance on axiomatic "guarantees" (e.g., of "fairness") when appropriating Shapley based feature *attribution* methods for feature *selection* (see, e.g., [40]–[49]). Through the arguments in this paper, the authors are convinced of two things:

- The axioms do not *in general* provide any guarantee that the Shapley value is suited to feature selection, and may, in some cases, imply the opposite.
- The relevance of the Shapley value to the feature selection task (indeed, to any ML task) is governed by the specific game formulation, and the justification of this relevance from the axioms is non-trivial.

In Section II, we introduce the Shapley value and game formulation (Section II-A), the axioms (Section II-B), we give a brief overview of feature selection in general (Section II-C), and then a brief overview of Shapley value feature selection (Section II-D). In Section III, we investigate the significance of the Shapley value axioms in the context of feature selection, introducing general "toy" examples to illustrate. Then, in Section IV we perform simulation studies on more concrete data sets, with various evaluation functions and game formulations, including the popular mean absolute SHAP and

[1]School of Mathematics and Physics, The University of Queensland, St Lucia, Australia
[2]SimulaMet, Simula Research Laboratory, Oslo, Norway
[3]Department of Mathematics and Statistics, La Trobe University, Melbourne, Australia

SAGE formulations.

> "*Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question...*" – John Tukey [50].

II. Background knowledge

A. The Shapley value

In Definition II.1, we define a Transferable Utility (TU) game. This definition captures the general scenario where a set of objects, denoted by F, has some associated evaluation, $C(F)$, and, for any subset S of F, the evaluation $C(S)$ is also well-defined. This captures a typical feature selection scenario, in which $F = \{1, \ldots, d\}$ represents the indices of all features in the full model of dimension d, and S represents the indices of features in some submodel of dimension $|S|$. In a TU game, we are guaranteed that the worth of every submodel can be evaluated, as $C(S)$.

In the following, we use 2^F to denote the set of all possible subsets of the objects F, which include F itself and the empty set, denoted by $\varnothing$.

Definition II.1 (TU game)**.** A TU game is a pair (F, C), where $F = \{1, \ldots, d\}$ is a set of indices called *players* and the *characteristic function* $C : 2^F \to \mathbb{R}$ assigns a non-negative real value $C(S)$ to every coalition $S \subseteq F$. Furthermore, C assigns the value zero to the empty coalition $\varnothing$, i.e. $C(\varnothing) = 0$.

For readability, we will often refer to C as an *evaluation function*, since it evaluates the worth of each coalition, and we will refer to the players as *features*.

The Shapley value (Definition II.3) is intuitively appealing for feature selection. At first glance, it extracts (or compresses) information from the evaluation function C, to assign a single value φ_i representing the worth of each feature in the modelling task. Here, the meaning of *worth* depends strongly on the choice of evaluation function, and (less obviously) on the manner in which features are understood to be removed from the model (see [36] and [24, Section 4]). Generalising the terminology in [36], we refer to these choices as the Shapley value *game formulation*.

An attractive notion is that, with a suitable game formulation, the Shapley value could guarantee a principled method of feature selection. However, while such a method of feature selection is *principled* via the axioms that are discussed in Section II-B, it is, as we will see, the principles (or axioms) themselves that are not *in general* suited to feature selection, and must be scrutinised in both the context of the specific game formulation, and the context of the outcome that is desired from the feature selection task.

Central to the definition of the Shapley value is the notion of a marginal contribution, which can be understood as the amount by which the evaluation of a given submodel increases, upon introducing a given feature to the submodel.

Definition II.2 (Marginal contribution)**.** The marginal contribution of feature i to submodel S is defined as the difference in evaluation when i is added to the submodel:

$$M_i(S) = C(S \cup \{i\}) - C(S).$$

Definition II.3 (The Shapley value)**.** The Shapley value of feature i is defined as a weighted average over all marginal contributions by feature i. That is, over $M_i(S)$ for every subset S of F that excludes i.

$$\varphi_i = \sum_{S \in 2^{F \setminus \{i\}}} \omega(S) M_i(S), \tag{1}$$

where $\omega(S) = |S|!(|F| - |S| - 1)!/|F|!$ are the specific weights that define the Shapley value.

The formula for the Shapley value, or at least that for its weights, may not immediately lend itself to intuition. Historically, much attention has been paid instead to the small set of axioms (in Section II-B) from which Definition II.3 can be *uniquely* derived. An *axiom* is a principle, usually taken to be self evident, from which other truths may be derived. The simple and intuitive nature of the Shapley value axioms encourages an assessment that the Shapley value is an *explainable* or *interpretable* approach to computing the importance of features. However, great care must be exercised in establishing the exact meaning of *importance*, both in the general Shapley value context *and* in the specific contexts in which the Shapley value is applied.

B. The Shapley value axioms

In keeping with the machine learning literature, and as a matter of preference, we present the following four axioms as a unique characterisation of the Shapley value. However, there are a number of alternative axiomatisations available that may provide varying levels of insight [28].

Axiom 1 (*Efficiency*)**.** In a TU game (C, F), the worth of the full model $C(F)$ is distributed in a lossless manner among the features:

$$\textstyle\sum_{i \in F} \varphi_i = C(F).$$

Axiom 2 (*Null player*)**.** In a TU game (C, F), if feature i contributes nothing to each submodel it enters, then its Shapley value is zero:

$$[(\forall\, S)\ C(S \cup \{i\}) = C(\{i\})] \implies \varphi_i = 0.$$

Axiom 3 (*Symmetry*)**.** In a TU game (C, F), any two features i, j that play equal roles have equal Shapley values:

$$[(\forall\, S \setminus \{i, j\})\ C(S \cup \{i\}) = C(S \cup \{j\})] \implies \varphi_i = \varphi_j.$$

Axiom 4 (*Additivity*)**.** Given two TU games $(C, F), (K, F)$, the Shapley value of feature i preserves addition of the evaluation functions:

$$\varphi_i(C + K) = \varphi_i(C) + \varphi_i(K).$$

In Axiom 4, the notation $\varphi_i(C)$ denotes the Shapley value of player i using the evaluation function C, and the addition of two evaluation functions is defined as the natural (pointwise) addition $(C + K)(S) = C(S) + K(S)$.

Axioms 2–4 can be replaced (as in [34]) by the following single axiom, which [33] named *balanced contributions*.

Axiom 5 (*Balanced contributions*)**.** Let C_i denote the game produced by restricting the feature set F to $F\backslash\{i\}$. Then,

$$\varphi_i(C) - \varphi_i(C_j) = \varphi_j(C) - \varphi_j(C_i)\,.$$

To paraphrase [33], balanced contributions is a principle of fairness in cooperation. It states that every pair of features should share equally the gain (or loss) received from their cooperation.

In a TU game, Axioms 1–4 are sufficient to uniquely define the *exact* Shapley value. However, this value is only unique up to the choice of characteristic function and game formulation. Between such choices, the Shapley value will vary greatly. Also, it should be noted that in many practical settings, the Shapley value is only approximated, since the complexity of (1) is exponential in the number of features (see, e.g., [15] and [24, Section 6.2]).

In algorithmic feature selection the search space generally involves 2^d submodels. Existing feature selection methods all take some approach to avoiding the exhaustive search of the model space.

C. Feature selection in general

If an evaluation function C is used for feature selection, it should correspond to a specific goal. Feature selection may target a number of typically exclusive goals, between which there is generally understood to be a trade-off in performance. These goals include, but may not be limited to, succinctly describing a data generating process; improving the predictive performance of the model; maximising the overall significance or power of the model, or of its parameters, with respect to a hypothesis; and producing a more cost-effective or computationally efficient predictor. There are two prominent categories of feature selection techniques identified by the surveys of [51], [52]:

- *Wrapper methods* evaluate a number of trained submodels selected via sequential (e.g., stepwise forward/backward) elimination, or via a heuristic search algorithm. While these model-driven evaluations are extrinsic to the data, they are native to the specific modelling task.
- *Filter methods* are general model-independent frameworks that avoid the computational burden of model training, present in wrapper methods. These methods rank features via empirical estimates of intrinsic properties of the data, such as covariance or mutual information.

Both of the above methods can be applied in the context of Shapley values, given appropriate choices for the evaluation function and game formulation (see Section II-D). Regardless of the specific goal, the feature selection task is motivated by the notion that a submodel exists that is of higher value than the full model, in the sense of Definition II.4.

Definition II.4 (Monotonicity)**.** In a TU game (F, C), the evaluation function $C : 2^F \to \mathbb{R}$ is called monotonic when $C(S)$ does not decrease whenever new features are added to S.

$$(\forall\, S, T \subseteq F)\; S \subset T \implies C(S) \leq C(T).$$

Popular examples of non-monotonic evaluation functions in feature selection are the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), often used in conjunction with a stepwise procedure [1].

D. Shapley values for feature selection

The following is the simplest general Shapley value feature selection procedure.

Algorithm 1 Attribution selection

1: Choose an objective function C.
2: Compute the Shapley value φ_i for all features $i \in F$.
3: Select the k highest ranking features, for $k < d = |F|$.

An alternative to the final step in Algorithm 1 is to select features i for which φ_i lies above some threshold. Upon review of the literature, we found a number of articles suggesting to use a variant of Algorithm 1 [44], [45], [48], [49], two completely applied uses of Algorithm 1 [46], [47], and one paper critical of the algorithm [53].

Alternatives to Algorithm 1 are found in [40]–[43]. In [40], [41] the considered coalition sizes are restricted, and a stepwise selection procedure is performed. A genetic algorithm is described in [42]. In [43], the problem of model averaging (see Section III-A) is discussed, and it is claimed that the method avoids the influence of unselected features via a decomposition of the Shapley value into high-order interaction components. We do not investigate these methods, but we expect that controlling the coalition sizes may have a positive impact, at least on the problem considered in Example 2 and Section IV-C.

III. The Meaning of the Axioms

In this section we use "toy" examples to interrogate some general consequences of the axioms, and to gain insight into how feature selection may be impacted by them. The insights discussed in this section are too general for drawing conclusions about the value of any specific Shapley value feature selection procedure, in practice.

A. The meaning of model averaging

As a consequence of Axioms 1 and 4 (efficiency and additivity), Shapley values are a model averaging procedure, being the weighted average (1) of marginal contributions. In statistics, model averaging has been used to combine the strengths of several candidate selected models [1]. These candidate models may arise from perturbations, e.g., when the result of a model selection procedure is recognised to be sensitive to sample effects or other conditions, or in circumstances where there is no clear optimum of the evaluation function. In any case, model averaging procedures are traditionally not used for feature selection, but to give a weighted average of estimates or predictions associated with a number of different models.

There is a compelling reason for caution around the direct use of model averaging for feature selection: The average

performance of a feature across *all* submodels may not be indicative of the particular performance of that feature in the set of optimal submodels. Ideally, one would select all features explicitly on the weight of their contribution to submodels that are optimal. We illustrate this with the following example.

Example 1 (*Taxicab payoff*)**.** Consider a game with $d = 3$ players and "taxicab" style payoff with characteristic function given by

$$C(S) = \begin{cases} 10, & \text{if } 3 \in S, \\ 7, & \text{if } 3 \notin S \text{ and } 2 \in S, \\ 3, & \text{if } 2, 3 \notin S \text{ and } 1 \in S, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This game can be pictured as one taxicab ride, where the homes of players 1 and 2 lie on route to the home of player 3. From the driver's perspective, the maximum profit is obtained from player 3, regardless of any absence of players 1 or 2. If this is a feature selection task, where all features have equal cost, then the optimal model does not contain features 1 or 2. From a model selection point of view, players 1 and 2 are useless features. However, since the Shapley values are

$$\varphi = (\varphi_1, \varphi_2, \varphi_3) = (1, 3, 6)\,,$$

from a *fairness* point of view, in the sense of the Shapley value, the players 1 and 2 are not worthless, since they add value to at least one other set of players (at least to the empty set, in the case of player 1).

In Example 1, players 1 and 2 are not "null players", in the sense of the Axiom 2, but they contribute nothing to the optimal model. As the example demonstrates, performance of features may increase within submodels due to an interaction with a dominant feature. Averaging across all submodels takes these superfluous performances into account, when ideally the presence of the dominant feature should be fixed. Indeed, a possible solution to this problem may be to identify and fix the presence of such dominating features, prior to computing the Shapley value – though at this stage feature selection may no longer be required. A concrete manifestation of Example 1 is explored in Section IV-D.

B. The meaning of efficiency

Axiom 1 (efficiency), states that the evaluation of the full model, $C(F)$, should be distributed losslessly amongst the features. This axiom narrows the scope of possible model averaging procedures to those that treat the full model, not the selected model, as the final outcome. When the objective function is non-monotonic, this becomes especially distinguished from the problems discussed in Section III-A. Since the full model is not generally the target, non-monotonic evaluation functions imply that *efficient payoffs may waste value*.

Non-monotonic evaluation functions (Definition II.4) are those that may decrease as the number of features increases, i.e., there exist submodels T, S such that $C(T) < C(S)$ for some $T \supset S$. This means, for example, that we may have

$$\sum_{i \in F} \varphi_i = C(F) < C(S),$$

for some $S \subset F$. In this case, the Shapley values do not sum to the maximum possible value of the feature set, over all subsets of features. In other words, the Shapley values sum to the payoff for the full model, but they don't sum to a payoff that is optimal. Note that, at least for the exact Shapley value, evaluation of all 2^F submodels is an intermediate step in the calculation. From these evaluations, the optimal model can be computed. In practice, for some approximation to *optimal efficiency*, it may be preferable to estimate the optimal model prior to approximating the Shapley value – at which stage feature selection is no longer required.

C. The meaning of balanced contributions

Axiom 5 (balanced contributions), can be substituted for Axioms 2–4 (null player, symmetry and additivity). Axiom 5 captures a notion of fairness in the rewards of cooperation between players in a TU game. The symmetry axiom alone may be undesirable in certain contexts. In the case of two strongly correlated features, the symmetry axiom dictates that both should receive approximately the same attribution. However, in feature selection, the high correlation implies that one of the two features is redundant. Here, we regard a feature X_i as *redundant* in the presence other features $\mathcal{X} = \{X_j,\, j \in S\}$, if feature X_i is conditionally independent of Y, conditional on $\mathcal{X}$, for which we write $Y \perp\!\!\!\perp X_i | \mathcal{X}$. Redundancy is investigated in more precise contexts in Sections IV-A and IV-B.

A second consequence that may be attributed to Axiom 5, is that the earnings received by a coalition, after discounting the earnings of its subcoalitions, are shared equally amongst the players in that coalition. To better understand this, see the formulation of the Shapley value in terms of Harsanyi dividends [28], which we do not enter into here. In particular, from (1), a player's contribution to teams of size k is averaged across all $\binom{d}{k}$ teams of size k. Since the binomial coefficient decreases in $|k - d/2|$ for fixed d, a player's single contribution to a team at the extremes of the spectrum of team sizes will be weighted higher than to team sizes close to $d/2$. As a consequence, poor performance of a feature on particularly small submodels, such as the singleton models, may be weighted highly even if such small submodels are not attractive for the model selection task. We illustrate this in Example 2

Example 2. One should be wary of deciding that the features with high Shapley values are also the strongest contributors to model performance. Consider the scenario with $d = 3$ and a "secret holder" style payoff where player 1 is alone worthless, but has the "secret" that endows any team with the maximum possible payoff. Specifically, suppose we have the payoff lattice in Figure 1.

The Shapley value vector is

$$\varphi = (\varphi_1, \varphi_2, \varphi_3) = (2, 4, 4).$$

The players 2 and 3 are attributed twice the value of player 1, but if the submodel $\{2, 3\}$ is selected, then performance will be suboptimal. Furthermore, from the full model $\{1, 2, 3\}$,

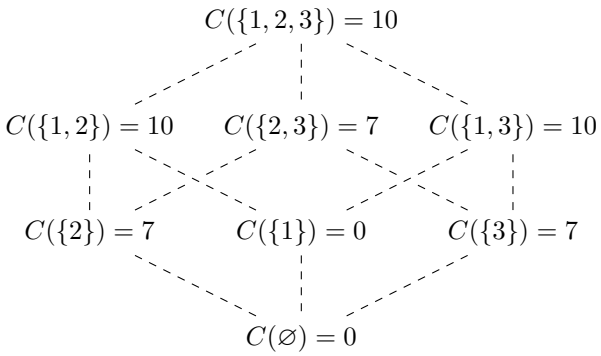

Fig. 1: A diagram of the evaluation function with "secret holder" style payoff described in Example 2.

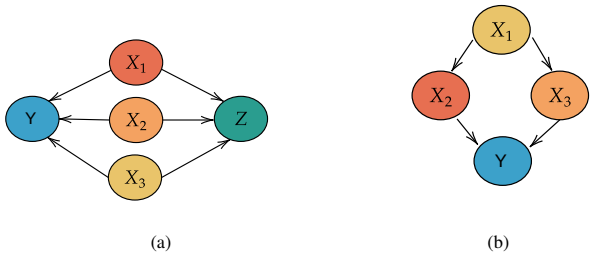

Fig. 2: Causal graphs, where (a) corresponds to the DGP (3) and (b) corresponds to (4).

we would do well to discard player 2 or 3. This example is investigated further in Section IV-C.

IV. Experimentation

In this section, where the Data Generating Process (DGP) is known to us, we investigate, through simulation, a number of situations in which the results produced by Algorithm 1 may be undesirable. In Sections IV-A and IV-B, we reproduce the results of [53, Theorem 8 and 9] and extend them via simulation to include the `SHAP FSelection` formulation [48] (i.e., feature selection by ranking mean absolute SHAP values of model predictions), and the SAGE formulation. Accurate definitions of SHAP and SAGE methods are very detailed, so rather than reproducing them here, we encourage the reader to access [17] (SHAP) and [49] (SAGE).

A. Markov boundary experiment 1

In this experiment we consider a scenario where we wish to predict a response variable Y from four features X_1, X_2, X_3, Z, with the following DGP suggested in [53, Theorem 8],

$$\begin{aligned} X_1,\, X_2,\, X_3 &\sim \mathcal{N}(0,4), \\ Y &= X_1 + X_2 + X_3 + \varepsilon, \\ Z &= X_1 + X_2 + X_3 + \gamma, \end{aligned} \tag{3}$$

where $\varepsilon, \gamma \sim \mathcal{N}(0,4)$ introduce irreducible uncertainty into the relationships defining Y and Z, respectively, via normal perturbations with means zero and variances 4. The causal graph is shown in fig. 2a, where $X_1, X_2, X_3 \to Y$ and $X_1, X_2, X_3 \to Z$. Regardless of any implied causal relationships, we can interpret it as a case where Z is separated from Y via two terms of uncertainty (ε and γ), while the set $\{X_1, X_2, X_3\}$ is separated from Y by only one term of uncertainty (ε). It follows that Y is conditionally independent of Z, given the values $\{X_1, X_2, X_3\}$, but not vice versa. Furthermore, $\{X_1, X_2, X_3\}$ is the minimal set with the property that the remaining features are conditionally independent of Y. Thus, X_1, X_2, X_3 are referred to as *Markov boundary members* of the DGP. However, as also shown by [53], when we employ the R^2 evaluation function, the Shapley values are

$$(\phi_Z, \phi_{X_1}, \phi_{X_2}, \phi_{X_3}) = (0.26, 0.16, 0.16, 0.16)\,,$$

The three features X_1, X_2, X_3, the Markov boundary members, all have smaller Shapley values than Z, the non-Markov boundary member. This is not a peculiarity of that particular game formulation. Simulating a data set with sample size $n = 10^6$ from DGP (3), and training an `XGBoost` regression model, the `SHAP FSelection` method sorts the features as (Z, X_1, X_2, X_3) with corresponding mean absolute SHAP values $(1.4, 1.1, 1.1, 1.1)$. In both cases, a top-3 feature selection procedure will select (Z, X_1, X_2), rather than the preferred triple (X_1, X_2, X_3), thus introducing an unnecessary term of uncertainty (γ) into the model.

The SAGE values, on the other hand, which compute a global Shapley value of model loss, representing the predictive power associated with each feature in a model, produce feature importance scores of $(1, 4, 4, 4)$ for the features (Z, X_1, X_2, X_3). Thus, while the `SHAP FSelection` and R^2 formulations produce poor results for model selection, the SAGE values successfully highlight the appropriate features in this scenario.

B. Markov boundary experiment 2

We study a generalisation of a DGP suggested by [53, Theorem 9], who consider the special case $\ell = 0.05$. First, we sample feature X_1 uniformly from $\{1, 2, 3, 4\}$. Then, X_1, X_2, X_3 are sampled as follows.

$$\begin{aligned} &P(X_2 = 1|X_1 = 1) = \ell,\, P(X_2 = 1|X_1 = 3) = \ell - 1, \\ &P(X_2 = 1|X_1 = 2) = \ell,\, P(X_3 = 1|X_1 = 2) = \ell - 1, \\ &P(X_3 = 1|X_1 = 1) = \ell,\, P(X_2 = 1|X_1 = 4) = \ell - 1, \\ &P(X_3 = 1|X_1 = 3) = \ell,\, P(X_3 = 1|X_1 = 4) = \ell - 1, \\ &P(Y = 1|X_2 = 0, X_3 = 0) = 0.9, \\ &P(Y = 1|X_2 = 0, X_3 = 1) = 0.05, \\ &P(Y = 1|X_2 = 1, X_3 = 0) = 0.15, \\ &P(Y = 1|X_2 = 1, X_3 = 1) = 0.9\,. \end{aligned} \tag{4}$$

Here, $\ell \in (0, 1)$. The causal graph is depicted in fig. 2b. In [53], the evaluation function m is used, defined as,

$$m(S) = \sum_{\mathbf{x}_S} P(\mathbf{X}_S = \mathbf{x}_S) V(S), \tag{5}$$

$$V(S) = \max\{P(Y = 1|\mathbf{X}_S = \mathbf{x}_s), P(Y = 0|\mathbf{X}_S = \mathbf{x}_s)\},$$

where $\mathbf{X}_S$ is the vector of features indexed by S. The resulting Shapley values, as given in [53], with $\ell = 0.05$ are

$$(\phi_{X_1}, \phi_{X_2}, \phi_{X_3}) = (0.22, 0.09, 0.09)\,.$$

Here, despite introducing unnecessary irreducible error into the model (see Figure 2b), the non-Markov boundary variable X_1 is given the highest preference. Simulating a data set of sample size $n = 10^6$, and training a simple `XGBoost` classification model, for $\ell = 0.05$, the `SHAP FSelection` method also sorts the features as (X_1, X_2, X_3), with respective mean absolute SHAP values $(1.43, 0.50, 0.40)$. On the other hand, SAGE sorts the features as (X_2, X_3, X_1) with values $(0.0797, 0.0787, 0.0003)$. Thus, as in Section IV-A, the SAGE values sidestep the issues of the `SHAP FSelection` and m formulations, instead producing a correct ranking for feature selection.

To determine the relationship of the parameter $\ell \in (0, 1)$ to the pathology, we simulate 20 more data sets, equally spaced on the grid $0.05 \leq \ell \leq 0.95$, and calculate `SHAP FSelection` and SAGE values for each value of ℓ. The variation of the SAGE values with ℓ is shown in fig. 3a. Calculating the differences $\varphi_1 - \varphi_2$ in Shapley values of the variables X_1 and X_2, for the SHAP, SAGE and m formulations, yields fig. 3b. Similar behaviour is realised in the differences $\varphi_1 - \varphi_3$. From this, we see that SAGE performs admirably over the investigated parameter space, while the SHAP and m formulations perform poorly for approximately $|\ell - 1/2| > 0.3$.

C. A secret holder experiment

In this experiment we consider the DGP

$$Y = \sum_{i=1}^{3} \beta_i X_i + \sum_{i=1}^{3}\sum_{j=1}^{3} \beta_{ij} X_i X_j + \varepsilon, \tag{6}$$

with $\varepsilon \sim \mathcal{N}(0, 1)$ and, given parameters $t_1, t_2 \in \mathbb{R}$, having $\beta_1 = 0, \beta_2 = \beta_3 = t_1 \neq 0, \beta_{23} = 0, \beta_{12} = \beta_{13} = t_2 \neq 0$. The features $X_i, i \in \{1, 2, 3\}$ are generated as independent standard normal random variables.

From (6), we simulate a total of 6561 data sets of sample size $n = 1000$, producing one data set for each position on the 81×81 parameter grid (t_1, t_2) with $-2 \leq t_i \leq 2$ and increments of length 0.05. On each data set we compute Shapley values for the conditional log likelihood evaluation function

$$L(S) = \mathcal{L}(\theta; \mathbf{x}_\varnothing) - \mathcal{L}(\theta; \mathbf{x}_S)\,, \tag{7}$$

where θ is estimated via least squares, using the closest submodel of the true model (6), given the vector $\mathbf{x}_S$ of features indexed by S. In Figure 4 we present the results of highlighting all (t_1, t_2) such that the characteristic function is pathological in the sense of matching Example 2. From this we see that pathologies occur at approximately $t_2 = \pm(|t_1| + \alpha)$ for $0 < \alpha < 0.4$.

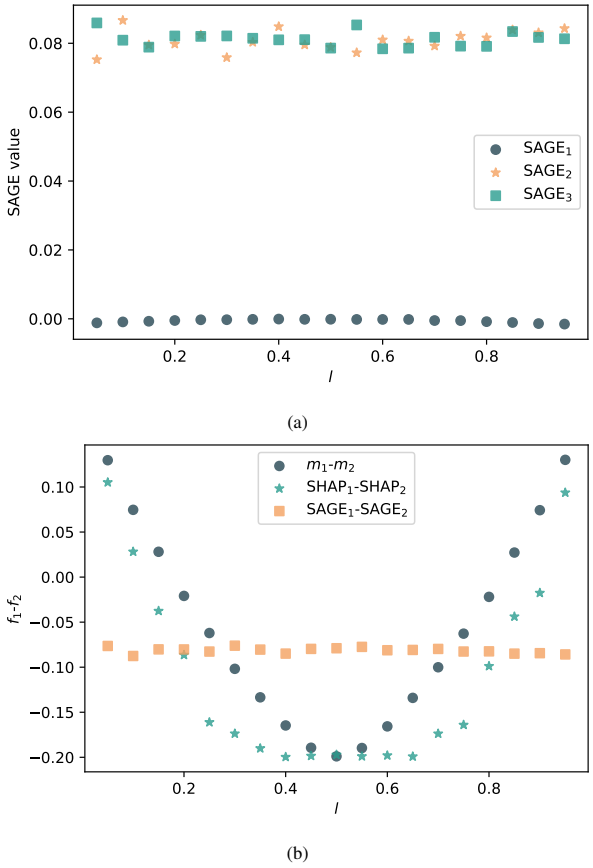

Fig. 3: The (a) SAGE values for the DGP in (4) across a grid of $0.05 \leq \ell \leq 0.95$, (b) Shapley values for the mean absolute SHAP, SAGE and m function formulations, for the difference $\varphi_1 - \varphi_2$ between attribution to X_1 and X_2, across the same range of ℓ values. Similar results were observed for the difference $\varphi_1 - \varphi_3$. Values $\varphi_1 - \varphi_2 < 0$ indicate regions of the parameter space for which model selection results are inadequate.

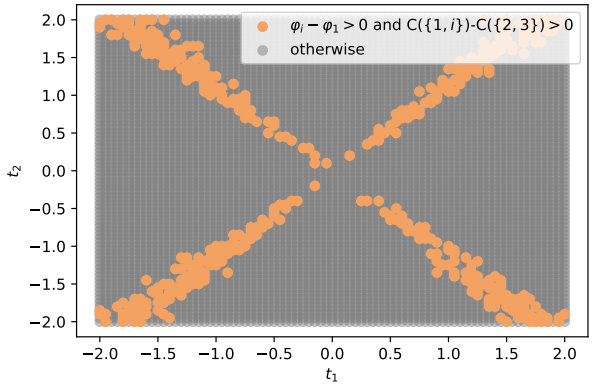

Fig. 4: Grid points (t_1, t_2) for $-2 \leq t_i \leq 2$ at 0.05 increments, in the parameter space for the DGP (6) described in Section IV-C. Points in orange are where $\varphi_i - \varphi_1 > 0$, as well as $C(\{1, i\}) - C(\{2, 3\}) > 0$, for both $i = 2$ and $i = 3$.

Choosing $t_1 = 2, t_2 = 2.2$, we compute the characteristic function for (7):

$$\begin{aligned} C(\{1,2,3\}) &= 1.52, \\ C(\{1,2\}) = 0.34,\ C(\{1\}) &= 0.00, \\ C(\{2,3\}) = 0.27,\ C(\{2\}) &= 0.13, \\ C(\{1,3\}) = 0.34,\ C(\{3\}) &= 0.11, \end{aligned}$$

which we compare to Figure 1. The Shapley values are

$$(\varphi_1, \varphi_2, \varphi_3) = (0.5, 0.52, 0.51).$$

Thus, although the results are quite close, we see that X_2 and X_3 are favoured in feature selection, despite X_1 being the "secret holder" and a member of both optimal submodels of size 2.

The corresponding mean absolute SHAP values are $(0.97, 1.52, 1.48)$ and the SAGE values are $(4.6, 5.8, 5.9)$. SHAP and SAGE disagree regarding whether feature X_2 or X_3 is the most important, but neither method gives precedence to X_1, which is in fact most important.

D. A taxicab experiment

The following scenario is a concrete realisation of Example 1. Consider d variables generated as

$$X_i = \mathcal{N}(0,1) + a_i\,, \tag{8}$$

where $a_1 < a_2 < \cdots < a_d$. The generative model is

$$Y = \max\{X_1, X_2, \ldots, X_d\} + \varepsilon\,, \tag{9}$$

where $\varepsilon \sim \mathcal{N}(0,1)$. The predictive model that we apply to this data is the correct model, which is simply

$$\hat{Y} = \max\{X_1, X_2, \ldots, X_d\}\,.$$

We evaluate this model using as evaluation function the following difference between the mean squared errors

$$\text{MSE}(y, 0) - \text{MSE}(y, \hat{y})\,,$$

and choosing $d = 3$ with $(a_1, a_2, a_3) = (5, 10, 20)$. The characteristic function evaluates to

$$\begin{aligned} C(\{1,2,3\}) &= 546, \\ C(\{1,2\}) = 371,\ C(\{1\}) &= 210, \\ C(\{2,3\}) = 546,\ C(\{2\}) &= 371, \\ C(\{1,3\}) = 546,\ C(\{3\}) &= 546. \end{aligned}$$

The resulting Shapley values are

$$(\varphi_1, \varphi_2, \varphi_3) = (70, 151, 325)\,.$$

We see that the scenario in Example 1 has been generated in this more concrete setting.

V. Discussion

Our investigations of the axioms in Section III prompted a number of experiments on potentially pathological DGPs, presented in Section IV. In each experiment, we were able to define a reasonable evaluation function and game formulation for which the Shapley value behaved in an undesirable way for feature selection. When these experiments were applied to the mean absolute SHAP (i.e., `SHAP FSelect`) and SAGE formulations, the former performed poorly in all of the three experiments (Sections IV-A to IV-C), while the latter performed favourably in two out of those three experiments.

Our results confirm that the axioms do not *in general* provide any guarantee that the Shapley value is suited to feature selection, and may in some cases imply the opposite. However, more importantly, the relevance of the Shapley value to the feature selection task (indeed, to any ML task) is governed by the specific game formulation, and the justification of this relevance from the axioms is non-trivial. Crucially, it is the authors' perception that abstract general axioms presented as "favourable and fair" may introduce a significant potential for *magical thinking* [54] in XAI. It is our intention to caution against this, and to emphasize the nuance in any application of Shapley values.

There are a large variety of alternatives to the Shapley value provided in the literature on game theory [28]. As an example, [28, p.55] suggests (under Games with Hierarchies), an *inessential player* axiom, dictating that players that are not essential should receive zero value.

Future work is needed to thoroughly explore the application of game theoretic solution concepts to feature selection and attribution. Extensive empirical studies are needed to understand the game formulations and axioms that are suited to the large variety of practical and frequently occurring feature selection tasks in XAI and ML.

References

[1] G. Claeskens, N. L. Hjort et al., Model selection and model averaging. Cambridge: Cambridge University Press, 2008.

[2] F. Huettner and M. Sunder, "Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values," Electronic Journal of Statistics, vol. 6, no. none, pp. 1239 – 1250, 2012. [Online]. Available: https://doi.org/10.1214/12-EJS710

[3] A. B. Owen and C. Prieur, "On Shapley value for measuring importance of dependent inputs," 2017.

[4] A. B. Owen, "Sobol' indices and Shapley value," SIAM/ASA Journal on Uncertainty Quantification, vol. 2, no. 1, pp. 245–251, 2014. [Online]. Available: https://doi.org/10.1137/130936233

[5] O. Israeli, "A Shapley-based decomposition of the r-square of a linear regression," Journal of Economic Inequality, vol. 5, pp. 199–212, 02 2007.

[6] E. Song, B. Nelson, and J. Staum, "Shapley effects for global sensitivity analysis: Theory and computation," SIAM/ASA Journal on Uncertainty Quantification, vol. 4, pp. 1060–1083, 01 2016.

[7] D. V. Fryer, I. Strümke, and H. D. Nguyen, "Shapley value confidence intervals for attributing variance explained," Frontiers in Applied Mathematics and Statistics, vol. 6, p. 58, 2020.

[8] U. Grömping and S. Landau, "Do not adjust coefficients in shapley value regression," Applied Stochastic Models in Business and Industry, vol. 26, pp. 194 – 202, 03 2010.

[9] R. H. Lindeman, "Introduction to bivariate and multivariate analysis," Tech. Rep., 1980.

[10] W. Kruskal, "Relative importance by averaging over orderings," The American Statistician, vol. 41, no. 1, pp. 6–10, 1987.

[11] ——, "Correction to "relative importance by averaging over orderings,"," The American Statistician, vol. 41, p. 341, 1987.

[12] S. Lipovetsky and M. Conklin, "Analysis of regression in game theory approach," Applied Stochastic Models in Business and Industry, vol. 17, pp. 319 – 330, 10 2001.

[13] D. V. Budescu, "Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression." Psychological bulletin, vol. 114, no. 3, p. 542, 1993.

[14] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values," 2020.

[15] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "Explainable ai for trees: From local explanations to global understanding," 2019.

[16] N. Sellereite and M. Jullum, "shapr: An R-package for explaining machine learning models with dependence-aware Shapley values," Journal of Open Source Software, vol. 5, no. 46, p. 2027, 2019. [Online]. Available: https://doi.org/10.21105/joss.02027

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[18] E. Strumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," Knowledge and Information Systems, vol. 41, pp. 647–665, 12 2013.

[19] ——, "An efficient explanation of individual classifications using game theory," Journal of Machine Learning Research, vol. 11, pp. 1–18, 01 2010.

[20] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," 2019.

[21] A. Redelmeier, M. Jullum, and K. Aas, "Explaining predictive models with mixed features using Shapley values and conditional inference trees," 2020.

[22] Y. Kwon, M. A. Rivas, and J. Zou, "Efficient computation and analysis of distributional Shapley values," 2021.

[23] N. Moehle, S. Boyd, and A. Ang, "Portfolio performance attribution via Shapley value," 2021.

[24] I. Covert, S. Lundberg, and S.-I. Lee, "Explaining by removing: A unified framework for model explanation," 2020.

[25] A. Keinan, C. C. Hilgetag, I. Meilijson, and E. Ruppin, "Fair attribution of functional contribution in artificial and biological networks," Neural Computation, vol. 16, pp. 1887–1915, 2003.

[26] D. V. Fryer, I. Strümke, and H. Nguyen, "Explaining the data or explaining a model? Shapley values that uncover non-linear dependencies," arXiv preprint arXiv:2007.06011, 2020.

[27] L. S. Shapley, "A value for n-person games," in Contributions to the Theory of Games (AM-28), Volume II, 1953.

[28] E. Algaba, V. Fragnelli, and J. Sanchez-Soriano, Handbook of the Shapley Value, Berlin, Heidelberg, 12 2019.

[29] G. Chalkiadakis, E. Elkind, and M. Wooldridge, "Computational aspects of cooperative game theory," in Lecture Notes in Computer Science, vol 6682, vol. 5, 10 2011.

[30] A. Roth and L. S. Shapley, "The Shapley value : essays in honor of lloyd s. Shapley," Economica, vol. 101, p. 123, 1991.

[31] L. Á. Kóczy, Partition Function Form Games - Coalitional Games with Externalities. Springer International Publishing, 2018.

[32] R. Gilles. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 01 2010, vol. 44.

[33] R. B. Myerson, "Conference structures and fair allocation rules," International Journal of Game Theory, vol. 9, no. 3, pp. 169–182, Sep 1980. [Online]. Available: https://doi.org/10.1007/BF01781371

[34] S. Hart and A. Mas-Colell, "Potential, Value, and Consistency," Econometrica, vol. 57, no. 3, pp. 589–614, May 1989.

[35] C. Frye, C. Rowat, and I. Feige, "Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability," 2020.

[36] L. Merrick and A. Taly, "The explanation game: Explaining machine learning models with cooperative game theory," CoRR, vol. abs/1909.08128, 2019. [Online]. Available: http://arxiv.org/abs/1909.08128

[37] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with shapley-value-based explanations as feature importance measures," 2020.

[38] M. Sundararajan and A. Najmi, "The many Shapley values for model explanation," 2020.

[39] U. Grömping, "Variable importance assessment in regression: Linear regression versus random forest," The American Statistician, vol. 63, no. 4, pp. 308–319, 2009. [Online]. Available: https://doi.org/10.1198/tast.2009.08199

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

[41] ——, "Feature selection based on the Shapley value," in Proceedings of IJCAI, 2005, pp. 1–6.

[42] 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," in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 319–323.

[43] C. Chu and D. Chan, "Feature selection using approximated high-order interaction components of the Shapley value for boosted tree classifier," IEEE Access, vol. PP, pp. 1–1, 06 2020.

[44] S. Tripathi, N. Hemachandra, and P. Trivedi, "On feature interactions identified by Shapley values of binary classification games," 2020.

[45] ——, "Interpretable feature subset selection: A Shapley value based approach," Proceedings of 2020 IEEE International Conference on Big Data, Special Session on Explainable Artificial Intelligence in Safety Critical Systems, 2020.

[46] R. Guha, A. Kha, P. Singh, and R. Sarkar, "Cga: A new feature selection model for visual human action recognition," Neural Computing and Applications, 05 2020.

[47] N. Jothi, W. Husain, and N. A. Rashid, "Predicting generalized anxiety disorder among women using Shapley value," Journal of Infection and Public Health, vol. 14, no. 1, pp. 103–108, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876034120304019

[48] W. Estécio Marcílio Júnior and D. Eler, "From explanations to feature selection: assessing SHAP values as feature selection mechanism," in 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 11 2020.

[49] I. Covert, S. Lundberg, and S.-I. Lee, "Understanding global feature contributions with additive importance measures," 2020.

[50] J. W. Tukey, "The future of data analysis," The annals of mathematical statistics, vol. 33, no. 1, pp. 1–67, 1962.

[51] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, no. 1, pp. 16–28, 2014, 40th-year commemorative issue. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790613003066

[52] H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao, "Intelligent data engineering and automated learning - ideal 2007," 8th International Conference, Birmingham, UK, December 16-19, 2007, Proceedings, 01 2007.

[53] S. Ma and R. Tourani, "Predictive and causal implications of using shapley value for model interpretation," 2020.

[54] P. Diaconis, "Theories of data analysis: From magical thinking through classical statistics," Exploring data tables, trends, and shapes, pp. 1–36, 2006.