The Explanation Game: Explaining Machine Learning Models with Cooperative Game Theory

Luke Merrick[1] and **Ankur Taly**[1]

Abstract. A number of techniques have been proposed to explain a machine learning (ML) model's prediction by attributing it to the corresponding input features. Popular among these are techniques that apply the Shapley value method from cooperative game theory. While existing papers focus on the axiomatic motivation of Shapley values, and efficient techniques for computing them, they neither justify the game formulations used nor address the uncertainty implicit in their methods' outputs. For instance, the SHAP algorithm's formulation [14] may give substantial attributions to features that play no role in a model. Furthermore, without infinite data and computation, SHAP attributions are approximations subject to hitherto uncharacterized uncertainty. In this work, we illustrate how subtle differences in the underlying game formulations of existing methods can cause large differences in attribution for a prediction. We then present a general game formulation that unifies existing methods. Using the primitive of *single-reference games*, we decompose the Shapley values of the general game formulation into Shapley values of single-reference games. This decomposition enables us to introduce confidence intervals to quantify the uncertainty in estimated attributions. Additionally, this decomposition enables *contrastive explanations* of a prediction through comparisons with different groups of reference inputs. We tie this idea to classic work on Norm Theory [11] in cognitive psychology, and propose a general framework for generating explanations for ML models, called *formulate, approximate, and explain* (FAE).

1 INTRODUCTION

Complex machine learning (ML) models are rapidly spreading to high stakes tasks such as credit scoring, underwriting, medical diagnosis, and crime prediction. Consequently, it is becoming increasingly important to interpret and explain individual model predictions to decision-makers, end-users, and regulators. A common form of model explanations are based on *feature attributions*, wherein a score (*attribution*) is ascribed to each feature in proportion to the feature's contribution to the prediction. Over the last few years there has been a surge in feature attribution methods, with methods based on Shapley values from cooperative game theory being prominent among them.

Shapley values [17] provide a mathematically fair and unique method to attribute the payoff of a cooperative game to the players of the game. Due to its strong axiomatic guarantees, the Shapley values method is emerging as the de facto approach to feature attribution, and some researchers even speculate that it may be the only method compliant with legal regulation such as GDPR's "right to an explanation" [1].

We begin by studying several Shapley-value-based explanation algorithms: TreeSHAP [13], KernelSHAP [14], QII [6], and IME [19]. Paradoxically, while all of these techniques lay claim to the axiomatic uniqueness of Shapley values, we discover that they yield significantly different attributions even when evaluated exactly (without approximation). Although these techniques offer the axiomatic motivations of Shapley values and efficient algorithms for computing them, they do not clearly justify the various explanation game formulations (which we term *explanation games*) that they rely upon. We find that as a result, these techniques can yield counter-intuitive attributions even on simple toy models. For instance, in Section 3 we show a simple model for which the popular SHAP method gives substantial attribution to a feature that is irrelevant to the model function. We show mathematically that this is a shortcoming of the explanation game formulated by SHAP.[2]

We next pursue a fundamental understanding of how to formulate explanation games whose Shapley values admit meaningful and relevant explanations. To this end, we first unify KernelSHAP, QII, and IME with a general game formulation parameterized by a single probability distribution. We decompose the Shapley values of this general game formulation into the Shapley values of *single-reference games* that model a feature's absence by replacing its value with the corresponding value from a specific counterfactual *reference input*.

This decomposition is instructive in several ways. First, it allows us to efficiently compute confidence intervals and other supplementary information about attributions, a notable advancement over existing methods (which lack confidence intervals even though they approximate metrics of random variables using finite samples). Second, it offers conceptual clarity. We interpret the (decomposed) feature attributions using the lens of Norm Theory, a classic work in cognitive psychology. Leveraging insights from Norm Theory, we develop a general *formulate, approximate, and explain* (FAE) framework to create Shapley-value-based feature attributions that are not only axiomatically justified, but also relevant and meaningful to the humans who consume them. Notably, the FAE framework allows us to *contrastively* explain a model prediction relative to a chosen group of reference inputs.

To illustrate these ideas, we present case studies explaining the predictions of models trained on two UCI datasets (Bike Sharing and Adult Income) and a Lending Club dataset. We find that in these real-world situations, explanations generated using our FAE framework uncover important patterns that previous attribution methods cannot identify.

In summary, we make the following key contributions:

[1] Fiddler Labs, USA, email: {luke, ankur}@fiddler.ai

[2] We note that this shortcoming, and the multiplicity of game formulations has also been noted in parallel work [21].

- We highlight several shortcomings of existing Shapley-value-based feature attribution methods (Sections 3), and analyze the root cause of these issues (Section 4.2).
- We present a novel game formulation that unifies and illuminates existing methods (Section 4.3).
- We characterize attribution uncertainty with confidence intervals (Section 4.4).
- We combine our formulation with principles from Norm Theory [11] to establish the *formulate, approximate, and explain* (FAE) framework (Section 5), and demonstrate noticeable improvements over the prior art through case studies (Section 6).

2 PRELIMINARIES

2.1 Additive feature attributions

Additive feature attributions [14] are attributions that sum to the difference between the explained model output $f(\boldsymbol{x})$ and a reference output value ϕ_0. In practice, ϕ_0 is typically an average model output or model output for a domain-specific "baseline" input (e.g. an empty string for text sentiment classification).

Definition 1 (Additive feature attributions) *Suppose $f : \mathcal{X} \to \mathbb{R}$ is a model mapping an M-dimensional feature space $\mathcal{X}$ to real-valued predictions. Additive feature attributions for $f(\boldsymbol{x})$ at input $\boldsymbol{x} = (x_1, \dots, x_M) \in \mathcal{X}$ comprise of a reference (or baseline) attribution ϕ_0 and feature attributions $\boldsymbol{\phi} = (\phi_1, \phi_2, \dots, \phi_M)$ corresponding to the M features such that $f(\boldsymbol{x}) = \phi_0 + \sum_{i=1}^{M} \phi_i$.*

There currently exist a number of competing methodologies for computing these attributions (see [2]). Given the difficulty of empirically evaluating attributions, several methods offer an axiomatic justification, often through the Shapley values method.

2.2 Shapley values

The Shapley values method is a classic technique from game theory that fairly attributes the total payoff from a cooperative game to the game's players [17]. Recently, this method has found numerous applications in explaining ML models (e.g. [5, 14, 8]).

Formally, a cooperative game is played by a set of players $\mathcal{M} = \{1, \dots, M\}$ termed the *grand coalition*. The game is characterized by a set function $v : 2^{\mathcal{M}} \to \mathbb{R}$ such that $v(S)$ is the payoff for any coalition of players $S \subseteq \mathcal{M}$, and $v(\emptyset) = 0$. Shapley values are built by examining the marginal contribution of a player to an existing coalition S, i.e., $v(S \cup \{i\}) - v(S)$. The Shapley value of a player i, denoted $\phi_i(v)$, is a certain weighted aggregation of its marginal contribution to all possible coalitions of players.

$$\phi_i(v) = \frac{1}{M} \sum_{S \subseteq \mathcal{M} \setminus \{i\}} \binom{M-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S)) \quad (1)$$

The Shapley value method is the unique method satisfying four desirable axioms: *Dummy*, *Symmetry*, *Efficiency*, and *Linearity*. We informally describe the axioms in the Supplemental Material,[3] and refer the reader to [23] for formal definitions and proofs.

2.2.1 Approximating Shapley values

Computing Shapley values involves evaluating the game payoff for every possible coalition of players. This makes the computation exponential in the number of players. For games with few players, it is possible to exactly compute the Shapley values, but for games with many players, the Shapley values can only be approximated. Recently there has been much progress towards the efficient approximation of Shapley values. In this work we focus on a simple sampling approximation, presenting two more popular techniques in the Supplemental Material. We refer the reader to [15, 4, 1, 10, 3] for a fuller picture of recent advances in Shapley value approximation.

A simple sampling approximation (used by [8], among other works) relies on the fact that the Shapley value can be expressed as the expected marginal contribution a player has when players are added to a coalition in a random order. Let $\pi(M)$ be the ordered set of permutations of M, and $\boldsymbol{O}$ be an ordering randomly sampled from $\pi(M)$. Let $\mathrm{pre}_i(\boldsymbol{O})$ be the set of players that precede player i in $\boldsymbol{O}$. The Shapley value of player i is the expected marginal contribution of the player under all possible orderings of players.

$$\phi_i(v) = \mathop{\mathbb{E}}_{\boldsymbol{O} \sim \pi(M)} [v(\mathrm{pre}_i(\boldsymbol{O}) \cup \{i\}) - v(\mathrm{pre}_i(\boldsymbol{O}))] \quad (2)$$

By sampling a number of permutations and averaging the marginal contributions of each player, we can estimate this expected value for each player and approximate each player's Shapley value.

3 A motivating example

Table 1: Probability mass function and model outputs for the mover hiring system example.

x_{male}	x_{lift}	Pr $[\boldsymbol{X} = \boldsymbol{x}]$	$f_{male}(\boldsymbol{x})$	$f_{both}(\boldsymbol{x})$
0	0	0.1	0.0	0.0
0	1	0.0	0.0	0.0
1	0	0.4	1.0	0.0
1	1	0.5	1.0	1.0

To probe existing Shapley-value-based model explanation methods, we evaluate them on two toy models for which it is easy to intuit correct attributions. We leverage a modified version of the example provided in [6]: a system that recommends whether a moving company should hire a mover applicant. Models consider an input vector of the binary features "is male" and "is good lifter" (notated $\boldsymbol{x} = (x_{male}, x_{lift})$), and output a recommendation score between 0 ("no hire") and 1 ("hire"). We define $f_{male}(\boldsymbol{x}) ::= x_{male}$ (only hire males), and $f_{both}(\boldsymbol{x}) ::= x_{male} \wedge x_{lift}$ (only hire males who are good lifters). Table 1 specifies a probability distribution over the input space, along with the predictions from the two models.

Consider the input $\boldsymbol{x} = (1, 1)$ (i.e. a male who is a good lifter), for which both models output a recommendation score of 1. Table 2 lists these predictions' attributions from several existing methods. Focusing on the relative attribution between x_{male} and x_{lift} (ignoring for now the magnitude of the attributions and the reference term ϕ_0), we make the following surprising observations. First, even though x_{lift} is irrelevant to f_{male}, the popular SHAP algorithm[4] results in equal attribution to both features, contradicting our intuition around the Dummy axiom of Shapley values. Additionally, these SHAP attributions are misleading from a fairness perspective: f_{male} relies solely on x_{male}, yet the attributions downplay this gender bias by claiming

[3] Supplemental material available at `https://github.com/fiddler-labs/the-explanation-game-supplemental`

[4] As defined by Equation 9 in [14].

Table 2: Attributions for the input $x_{male} = 1, x_{lift} = 1$.

		f_{male}			f_{both}	
Payoff formulation	ϕ_0 (baseline)	ϕ_1 (*male*)	ϕ_2 (*lifting*)	ϕ_0 (baseline)	ϕ_1 (*male*)	ϕ_2 (*lifting*)
SHAP (Conditional Distribution)	0.9	0.05	0.05	0.50	0.028	0.472
KernelSHAP (Input Distribution)	0.9	0.10	0.00	0.50	0.050	0.450
QII (Joint Marginal Distribution)	0.9	0.10	0.00	0.45	0.075	0.475
IME (Uniform Distribution)	0.5	0.50	0.00	0.25	0.375	0.375

the model uses both features equally. Second, although f_{both} treats its features symmetrically and $\boldsymbol{x}$ has identical values in both its features, many of the methods considered do not provide symmetrical attributions. This again is intuitively at odds with the Shapley value axioms, as Symmetry appears to be violated. These unintuitive behaviors surfaced by the above observations demand an in-depth study of these methods' internal design choices. We carry out this study in the next section.

4 EXPLANATION GAMES

In order to explain a model prediction with the Shapley values method, it is necessary to formulate a cooperative game with players that correspond to the features and a payoff that corresponds to the prediction. In this section, we analyze the methods examined in Section 3, and show that their surprising attributions are an artifact of their game formulations. We then discuss a unified game formulation and its decomposition to *single-reference games*, enabling conceptual clarity about the meanings of existing methods' attributions.

4.1 Notation

Let $\mathcal{D}^{inp}$ be the input distribution, which characterizes the process that generates model inputs. We denote the input of an explained prediction as $\boldsymbol{x} = (x_1, \ldots, x_M)$ and use $\boldsymbol{r}$ to denote another "reference" input. We use boldface to indicate when a variable or function is vector-valued, and we use capital letters for random variable inputs (although S continues to represent the set of contributing players/features). Thus, x_i is a scalar input, $\boldsymbol{x}$ is an input vector, and $\boldsymbol{X}$ is a *random* input vector. We use $\boldsymbol{x}_S = \{x_i, i \in S\}$ to represent a sub-vector of features indexed by S.

Lastly, we introduce the *composite input* $\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{r}, S)$, which agrees with the input $\boldsymbol{x}$ on all features in S and with $\boldsymbol{r}$ on all features not in S. Note that $\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{r}, \emptyset) = \boldsymbol{r}$, and $\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{r}, \mathcal{M}) = \boldsymbol{x}$.

$$\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{r}, S) = (z_1, z_2, ..., z_M), \text{ where } z_i = \begin{cases} x_i & i \in S \\ r_i & i \notin S \end{cases} \tag{3}$$

4.2 Existing game formulations

The explanation game payoff function $v_{\boldsymbol{x}}$ must be defined for every feature subset S such that $v_{\boldsymbol{x}}(S)$ captures the contribution of $\boldsymbol{x}_S$ to the model's prediction. This allows us to compute each feature's possible marginal contributions to the prediction and derive its Shapley value (see Section 2.2).

By the definition of *additive feature attributions* (Definition 1) and the Shapley values' Efficiency axiom, we must define $v_{\boldsymbol{x}}(\mathcal{M}) ::= f(\boldsymbol{x}) - \phi_0$ (i.e. the payoff of the full coalition must be the difference between the explained model prediction and a baseline prediction). Although this definition is fixed, it leaves us the challenge of coming up with the payoff when some features do not contribute (that is, when they are *absent*).

We find that all existing approaches handle this feature-absent payoff by randomly sampling absent features according to a particular counterfactual *reference* distribution and then computing the expected value of the prediction. The resulting game formulations differ from one another only in the counterfactual distribution they use. Additionally, we note that in practice small samples are used to approximate the expected value present in these payoff functions. This introduces a significant source of attribution uncertainty not clearly acknowledged or quantified by existing work.

4.2.1 Conditional distribution

The game formulation of SHAP [14], TreeSHAP [13], and [1] simulates feature absence by sampling absent features from the input distribution conditional on knowing the values of the present (or contributing) features:

$$v_{\boldsymbol{x}}^{cond}(S) = \mathop{\mathbb{E}}_{\boldsymbol{R} \sim \mathcal{D}^{inp}} [f(\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{R}, S)) \mid \boldsymbol{R}_S = \boldsymbol{x}_S] - \mathop{\mathbb{E}}_{\boldsymbol{R} \sim \mathcal{D}^{inp}} [f(\boldsymbol{R})] \tag{4}$$

Unfortunately, this formulation does not properly simulate the absence of a feature. This flaw explains why the irrelevant feature x_{lift} receives a nonzero attribution in the f_{male} example from Section 3. Specifically, since the event $x_{male} = 1$ is correlated[5] with $x_{lift} = 1$, once $x_{lift} = 1$ is given, the expected prediction becomes 1. This causes the x_{lift} feature to have a non-zero marginal contribution (relative to when both features are absent), and therefore a nonzero Shapley value. More generally, whenever a feature is correlated with a model's prediction on inputs drawn from $\mathcal{D}^{inp}$, this game formulation results in non-zero attribution to the feature regardless of whether the feature *causally* impacts the prediction.

4.2.2 Input distribution

Another option for simulating feature absence, which is used by KernelSHAP, is to sample absent features from the corresponding marginal distribution in $\mathcal{D}^{inp}$:

$$v_{\boldsymbol{x}}^{inp}(S) = \mathop{\mathbb{E}}_{\boldsymbol{R} \sim \mathcal{D}^{inp}} [f(\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{R}, S))] - \mathop{\mathbb{E}}_{\boldsymbol{R} \sim \mathcal{D}^{inp}} [f(\boldsymbol{R})] \tag{5}$$

Since this formulation breaks correlation with the contributing features, it ensures irrelevant features receive no attribution (e.g. no attribution to x_{lift} when explaining $f_{male}(1, 1) = 1$). We formally describe this property via the *Insentivity* axiom in Section 4.3.

Unfortunately, this formulation is still subject to artifacts of the input distribution, as evident from the asymmetrical attributions when explaining the prediction $f_{both}(1, 1) = 1$ (see Table 2). The features receive different attributions because they have different marginal distributions in $\mathcal{D}^{inp}$, not because they impact the model differently.

[5] In this context, *correlation* refers to general statistical dependence, not just a nonzero Pearson correlation coefficient.

4.2.3 Joint-marginal distribution

QII [6] simulates feature absence by sampling absent features one at a time from their own univariate marginal distributions. In addition to breaking correlation with the contributing features, this breaks correlation between absent features as well. Formally, the QII formulation uses a distribution we term the "joint-marginal" distribution ($\mathcal{D}^{J.M.}$), where:

$$\Pr_{X \sim \mathcal{D}^{J.M.}}[X = (x_1, \ldots, x_M)] = \prod_{i=1}^{M} \Pr_{X \sim \mathcal{D}^{inp}}[X_i = x_i]$$

The joint-marginal formulation $v_{\boldsymbol{x}}^{J.M.}$ is similar to $v_{\boldsymbol{x}}^{inp}$, except that the reference distribution is $\mathcal{D}^{J.M.}$ instead of $\mathcal{D}^{inp}$:

$$v_{\boldsymbol{x}}^{J.M.}(S) = \mathop{\mathbb{E}}_{\boldsymbol{R} \sim \mathcal{D}^{J.M.}}[f(\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{R}, S))] - \mathop{\mathbb{E}}_{\boldsymbol{R} \sim \mathcal{D}^{J.M.}}[f(\boldsymbol{R})] \qquad (6)$$

Unfortunately, like $v_{\boldsymbol{x}}^{inp}$, this game formulation is also tied to the input distribution and under-attributes features that take on common values in the background data. This is evident from the attributions for the f_{both} model shown in Table 2.

4.2.4 Uniform distribution

The last formulation we study from the prior art simulates feature absence by drawing values from a uniform distribution $\mathcal{U}$ over the entire input space, as in IME [19].[6] Completely ignoring the input distribution, this payoff $v_{\boldsymbol{x}}^{unif}$ considers all possible feature values (edge-cases and common cases) with equal weighting.

$$v_{\boldsymbol{x}}^{unif}(S) = \mathop{\mathbb{E}}_{\boldsymbol{R} \sim \mathcal{U}}[f(\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{R}, S))] - \mathop{\mathbb{E}}_{\boldsymbol{R} \sim \mathcal{U}}[f(\boldsymbol{R})] \qquad (7)$$

In Table 2, we see that this formulation yields intuitively correct attributions for f_{male} and f_{both}. However, the uniform distribution can sample so heavily from irrelevant outlier regions of $\mathcal{X}$ that relevant patterns of model behavior become masked (we study the importance of *relevant references* both theoretically in Section 5.1 and empirically in Section 6).

4.3 A unified formulation

We observe that the existing game formulations $v_{\boldsymbol{x}}^{inp}$, $v_{\boldsymbol{x}}^{J.M.}$, and $v_{\boldsymbol{x}}^{unif}$ can be unified as a single game formulation $v_{\boldsymbol{x},\mathcal{D}^{ref}}$ that is parameterized by a reference distribution $\mathcal{D}^{ref}$.

$$v_{\boldsymbol{x},\mathcal{D}^{ref}}(S) = \mathop{\mathbb{E}}_{\boldsymbol{R} \sim \mathcal{D}^{ref}}[f(\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{R}, S))] - \mathop{\mathbb{E}}_{\boldsymbol{R} \sim \mathcal{D}^{ref}}[f(\boldsymbol{R})] \qquad (8)$$

For instance, the formulation for KernelSHAP is recovered when $\mathcal{D}^{ref} = \mathcal{D}^{inp}$, and QII is recovered when $\mathcal{D}^{ref} = \mathcal{D}^{J.M.}$. In the rest of this section, we discuss several properties of this general formulation that help us better understand its attributions. Notably, the formulation $v_{\boldsymbol{x}}^{cond}$ cannot be expressed in this framework; we discuss the reason for this later in Section 4.3.2.

4.3.1 A decomposition in terms of single-reference games

We now introduce *single-reference games*, a conceptual building block that helps us interpret the Shapley values of the $v_{\boldsymbol{x},\mathcal{D}^{ref}}$ game. A single-reference game $v_{\boldsymbol{x},\boldsymbol{r}}$ simulates feature absence by replacing the feature value with a counterfactual value from a specific reference input $\boldsymbol{r}$:

$$v_{\boldsymbol{x},\boldsymbol{r}}(S) = f(\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{r}, S)) - f(\boldsymbol{r}) \qquad (9)$$

The attributions from a single-reference game explain the difference between the prediction for the input and the prediction for the reference (i.e. $\sum_i \phi_i(v_{\boldsymbol{x},\boldsymbol{r}}) = v_{\boldsymbol{x},\boldsymbol{r}}(\mathcal{M}) = f(\boldsymbol{x}) - f(\boldsymbol{r})$, and $\phi_0 = f(\boldsymbol{r})$). Computing attributions relative to a single reference point (also referred to as a "baseline") is common to several others methods [22, 18, 7, 3]. However, while those works seek a neutral "informationless" reference (e.g. an all-black image for image models), we find it beneficial to consider arbitrary references and interpret the resulting attributions relative to the reference. We develop this idea further in our FAE framework (see Section 5).

We now state Lemma 1, which shows how the Shapley values of $v_{\boldsymbol{x},\mathcal{D}^{ref}}$ can be expressed as the expected Shapley values of a (randomized) single-reference game $v_{\boldsymbol{x},\boldsymbol{R}}$, where $\boldsymbol{R} \sim \mathcal{D}$. The proof (given in full in the Supplemental Material) follows from the Shapley values' Linearity axiom and the linearity of expectation.

Lemma 1 $\boldsymbol{\phi}(v_{\boldsymbol{x},\mathcal{D}^{ref}}) = \mathbb{E}_{\boldsymbol{R} \sim \mathcal{D}^{ref}}[\boldsymbol{\phi}(v_{\boldsymbol{x},\boldsymbol{R}})]$

Lemma 1 brings conceptual clarity and practical improvements (confidence intervals and supplementary metrics) to existing methods. It shows that the attributions from existing games ($v_{\boldsymbol{x}}^{inp}$, $v_{\boldsymbol{x}}^{J.M.}$, and $v_{\boldsymbol{x}}^{unif}$) are in fact differently weighted aggregations of attributions from a space of single-reference games. For instance, $v_{\boldsymbol{x}}^{unif}$ weighs attributions relative to all reference points equally, while $v_{\boldsymbol{x}}^{inp}$ weighs them using the input distribution $\mathcal{D}^{inp}$.

4.3.2 Insensitivity axiom

We show that attributions from the game $v_{\boldsymbol{x},\mathcal{D}^{ref}}$ satisfy the *Insensitivity* axiom from [22], which states that a feature that is mathematically irrelevant to the model must receive zero attribution. Formally, a feature i is irrelevant to a model f if for any input, changing the feature does not change the model output. That is, $\forall \boldsymbol{x}, \boldsymbol{r} \in \mathcal{X}$: $\boldsymbol{x}_{\mathcal{M}\setminus\{i\}} = \boldsymbol{r}_{\mathcal{M}\setminus\{i\}} \implies f(\boldsymbol{x}) = f(\boldsymbol{r})$.

Lemma 2 *If a feature i is irrelevant to a model f then $\phi_i(v_{\boldsymbol{x},\mathcal{D}^{ref}}) = 0$ for all distributions $\mathcal{D}^{ref}$.*

The proof (given in the Supplemental Material) is based on showing that the axiom is obeyed by all single-reference games, and therefore by Lemma 1 is also obeyed by $v_{\boldsymbol{x},\mathcal{D}^{ref}}$ games. Notably, the $v_{\boldsymbol{x}}^{cond}$ formulation does not obey the Insensitivity axiom (a counterexample being the f_{male} attributions from Section 3). Accordingly, our general formulation (Equation 7) cannot express this formulation. In the rest of the paper, we focus on game formulations that satisfy the Insensitivity axiom. We refer to [21] for a comprehensive analysis of the axiomatic guarantees of various game formulations.

4.4 Confidence intervals on attributions

Existing game formulations involve computing an expected value (over a reference distribution) in every invocation of the payoff function. In practice, this expectation is approximated via sampling, which introduces uncertainty. The original formulations of these games do not lend themselves well to quantify such uncertainty. We show that by leveraging our unified game formulation, one can efficiently quantify the uncertainty using confidence intervals.

[6] It is somewhat unclear whether IME proposes $\mathcal{U}$ or $\mathcal{D}^{inp}$, as [19] assumes $\mathcal{D}^{inp} = \mathcal{U}$, while [20] calls for values to be sampled from $\mathcal{X}$ "at random."

Our decomposition in Lemma 1 shows that the attributions themselves can be expressed as an expectation over (deterministic) Shapley value attributions from a distribution of single-reference games. Consequently, we can quantify attribution uncertainty by estimating the standard error of the mean (SEM) across a sample of Shapley values from single-reference games. In terms of the sample standard deviation (SSD), 95% CIs on the mean attribution ($\bar{\boldsymbol{\phi}}$) from a sample of size N are given by

$$\bar{\boldsymbol{\phi}} \pm \frac{1.96 \times \mathsf{SSD}(\{\boldsymbol{\phi}(v_{\boldsymbol{x},\boldsymbol{r}_i})\}_{i=1}^{N})}{\sqrt{N}} \tag{10}$$

We note that while one could use bootstrap to obtain confidence interval (CIs), the SEM approach is more efficient as it requires no additional Shapley value computations.

4.4.1 A unified CI

As discussed in Section 2.2, often the large number of features (players) in an explanation game necessitates the approximation of Shapley values. The approximation may involve random sampling, which incurs its own uncertainty. In what follows, we derive a general SEM-based CI that quantifies the combined uncertainty from sampling-based approximations of Shapley values and the sampling of references.

Let us consider a generic estimator $\hat{\phi}_i^{(\boldsymbol{G})}(v_{\boldsymbol{x},\boldsymbol{r}})$ parameterized by some random sample $\boldsymbol{G}$. An example of such an approach is the feature ordering based approximation of Equation 2, for which $\boldsymbol{G} = (\boldsymbol{O}_j)_{j=1}^{k}$ represents a random sample of feature orderings, and:

$$\hat{\phi}_i^{(\boldsymbol{G})}(v_{\boldsymbol{x},\boldsymbol{r}}) = \frac{1}{k}\sum_{j=1}^{k} v(\mathrm{pre}_i(\boldsymbol{O}_j) \cup \{i\}) - v(\mathrm{pre}_i(\boldsymbol{O}_j))$$

As long as the generic $\hat{\phi}_i^{(\boldsymbol{G})}$ is an unbiased estimator (like the feature ordering estimator of Equation 2), and $\boldsymbol{G}$ and $\boldsymbol{R} \sim \mathcal{D}^{ref}$ are sampled independently from one another, we can derive a unified CI using the SEM. By the estimator's unbiasedness and Lemma 1, the Shapley value attributions can be expressed as:

$$\phi_i(v_{\boldsymbol{x},\mathcal{D}^{ref}}) = \mathop{\mathbb{E}}_{\boldsymbol{R}}\mathop{\mathbb{E}}_{\boldsymbol{G}}\left[\hat{\phi}_i^{(\boldsymbol{G})}(v_{\boldsymbol{x},\boldsymbol{R}})\right] \tag{11}$$

Since $\boldsymbol{G}$ is independent of $\boldsymbol{R}$, this expectation can be Monte Carlo estimated using the sample mean of the sequence $\left(\hat{\phi}_i^{(\boldsymbol{g}_j)}(v_{\boldsymbol{x},\boldsymbol{r}_j})\right)_{j=1}^{k}$ (where $(\boldsymbol{g}_j, \boldsymbol{r}_j)_{j=1}^{k}$ is a joint sample of $(\boldsymbol{G}, \boldsymbol{R})$). As the attribution recovered by this estimation is simply the mean of a sample from a random variable, its uncertainty can be quantified by estimating the SEM. In terms of the sample standard deviation (SSD), 95% CIs on the mean attribution ($\bar{\boldsymbol{\phi}}$) from a sample of size N are given by:

$$\bar{\boldsymbol{\phi}} \pm \frac{1.96 \times \mathsf{SSD}\left(\left(\hat{\phi}_i^{(\boldsymbol{g}_j)}(v_{\boldsymbol{x},\boldsymbol{r}_j})\right)_{j=1}^{k}\right)}{\sqrt{N}} \tag{12}$$

5 FORMULATE, APPROXIMATE, EXPLAIN

So far we have shown that existing Shapley value feature attribution methods can yield misleading and unintuitive attributions due to the game formulations that these methods use. We also have noted and quantified the approximation uncertainty incurred by these formulations. We now present the*formulate, approximate, and explain* (FAE) framework as a first principles approach to explaining machine learning models using the primitive of Shapley values.

5.1 Formulate

We take inspiration from Norm Theory [11], a classic work from cognitive psychology, to understand how to formulate explanation games in ways that match how humans frame explanations. Norm Theory describes the psychological norms that shape the emotional responses, social judgments, and explanations of humans. We focus on three pertinent findings from Norm Theory: that "why" questions evoke counterfactual norms, that norms vary depending on their context, and that norms tend to be relevant to to the question at hand.

Norm Theory posits that the interpretation of a "why" question implicitly refers to one or more counterfactual norms. As stated in [11], *"A why question indicates that a particular event is surprising and requests the explanation of an effect, denned as a contrast between an observation and a more normal alternative."* This suggests that explanations are *contrastive*, in the sense that they explain an event in contrast to one or more norms. To respect this, we should formulate explanation games that capture a contrastive question against a clear counterfactual. We note that this prescribes a departure from the typical practice of using a broad distribution like $\mathcal{D}^{inp}$ in the game formulation. Such broad distributions may not represent a concrete and relevant counterfactual scenario.

Norm Theory recognizes that these "normal alternatives" may vary across explainers, explainees, and the context of the explanation. Therefore, explanations must be interpreted relative to the chosen[7] norm. The choice of norm(s) or reference(s) is an important knob for obtaining different explanations, and using several explanation games characterized by different norms may provide a richer explanation than what a single explanation game can.

We also note that norms (references) must be *relevant* to the situation at hand. As noted in [9]: *"Our capacity for counterfactual reasoning seems to show a strong resistance to any consideration of irrelevant counterfactuals."* This principle may require applying task-specific domain knowledge, as it can be challenging to automatically identify relevant counterfactuals. For instance, if we are explaining why an auto-grading software assigns a B+ to a student's submission $\boldsymbol{x}$, it would be proper to contrast with the submissions that were graded as A- (next higher grade after B+), instead of contrasting with the entire pool of submissions.

The mandate of the Formulate step is to *generate one or more contrastive questions that each specify relevant counterfactual references*. Each question pins down the distribution $\mathcal{D}^{ref}$ of the chosen references (norms).

5.2 Approximate

Once a meaningful contrastive question and its corresponding reference distribution $\mathcal{D}^{ref}$ has been formulated, we bring to bear the axiomatic power of Shapley values in the Approximate step. We consider the distribution of single-reference games whose references are drawn from $\mathcal{D}^{ref}$, and approximate the Shapley values of these games. Formally, for each explanation game, we approximate the distribution of the random-valued attribution vector $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{R}} = \boldsymbol{\phi}(v_{\boldsymbol{x},\boldsymbol{R}})$, where $\boldsymbol{R} \sim \mathcal{D}^{ref}$. This involves two steps: (1) sampling a sequence of references $(\boldsymbol{r}_i)_{i=1}^{N}$ from $\boldsymbol{R} \sim \mathcal{D}^{ref}$, and (2) approximating the Shapley values for each of the single-reference games defined relative to a reference in $(\boldsymbol{r}_i)_{i=1}^{N}$. This yields a sequence of approximated Shapley values. It is important to be mindful of the uncertainty result-

[7] For humans, norms are said to be "evoked," in other words, chosen subconsciously.

ing from the sampling in steps (1) and (2), and appropriate quantify it in the Explain step.

5.3 Explain

In the final step, we summarize $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{R}}$ using our sampled Shapley value vectors, and present this summary in the context of the formulation. One simple summarization would be the presentation of a few representative examples of Shapley values of the sampled single-reference games, in the style of the SP-LIME algorithm [16]. Another simple summarization is the sample mean, which approximates $\mathbb{E}\left[\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{R}}\right]$, and is equivalent to the attributions from the unified explanation game $v_{\boldsymbol{x},\mathcal{D}^{ref}}$. When using the sample mean, the framework of Section 4.4 can be used to quantify the uncertainty from sampling.

A key limitation of the mean is that it may hide important information. For instance, a feature's attributions may have opposite signs relative to different references. Averaging these attributions will cause them to cancel each other out, yielding a small mean that incorrectly suggests the feature is unimportant. We discuss a concrete example of this in Section 6.1.1. At the very least, we recommend confirming through visualization and summary statistics like variance and interquartile range that the mean is a good summarization, before relying upon it.

5.3.1 Summarizing attributions using clustering

In cases where the mean does not offer a good summarization, we turn to unsupervised learning. We propose clustering the attribution distribution $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{R}}$ into clusters with lower intra-cluster variation as a method to obtain a more representative summary. As we show in Section 6.2.3, the clusters help segregate different attribution patterns along with their corresponding references.

The clustering approach can be seen as a middle ground between representative examples of single-reference games, and a global average. Using the maximum number of non-empty clusters (one per single-reference game) reduces to looking at representative examples, while using the minimum number of clusters (one) yields a global average across attributions.

Finally, clustering the attribution distribution into distinct groups can also be seen as reformulating an explanation game into a number of finer-grained games whose references partition the original game's reference distribution. In this way unsupervised learning "closes the loop" of FAE, returning to the formulate step and allowing for more refined formulations that yield better explanations.

6 CASE STUDIES

In this section we apply the formulate, approximate and explain framework (FAE) to LightGBM [12] Gradient Boosted Decision Trees (GBDT) models trained on real data: the UCI Bike Sharing and Adult Income datasets, and a Lending Club dataset.[8] For parsimony, we analyze models that use only five features.[9] For the Bike Sharing model, we explain a randomly selected prediction of 210 rentals for a certain hour. For the Adult Income model, we explain a counter-intuitively low prediction for an individual with high *education-num*. For the Lending Club model, we explain a counter-intuitive rejection (assuming a threshold that accepts 15% of loan applications) for a high-income borrower.

6.1 Shortcomings of existing methods

Recall from Section 4.3 that the attributions from existing methods amount to computing the mean attribution for the single-reference game $v_{\boldsymbol{x},\boldsymbol{R}}$, where the reference $\boldsymbol{R}$ is sampled from a certain distribution.

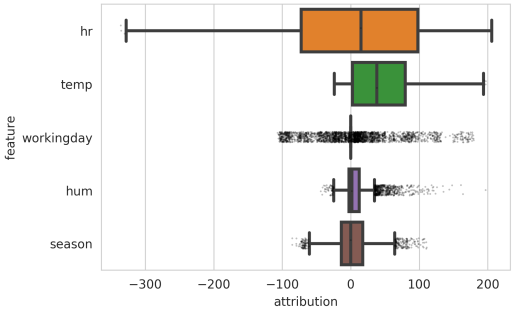


Figure 1: Distribution of attributions from the $v_{\boldsymbol{x},\boldsymbol{R}}$ formulation with $\boldsymbol{R} \sim \mathcal{D}^{inp}$ for the Bike Sharing example.

6.1.1 Misleading means

In Section 5, we discussed that the mean attribution can be a misleading summarization. Here we illustrate this using the attributions from the KernelSHAP game $v_{\boldsymbol{x}}^{inp}$ for the Bike Sharing example (see Table 6). The mean attribution to the feature *hr* is tiny, suggesting that the feature has little impact. However, the distribution of single-reference game attributions (Figure 1) reveals a large spread centered close to zero. In fact, we find that by absolute value *hr* receives the largest attribution in over 60% of the single-reference games. Consequently, only examining the mean of the distribution may be misleading.

Table 3: Bike Sharing comparison of mean attributions. 95% CIs ranged from ±0.4 (*hum* in $\mathcal{D}^{inp}$ and $\mathcal{D}^{J.M.}$) to ±2.5 (*hr* in $\mathcal{D}^{inp}$ and $\mathcal{D}^{J.M.}$).

Game Formulation	Avg. Prediction (ϕ_0)	hr	temp	work.	hum	season
$v_{\boldsymbol{x}}^{inp}$	151	3	47	1	7	2
$v_{\boldsymbol{x}}^{J.M.}$	141	6	50	1	9	3
$v_{\boldsymbol{x}}^{unif}$	128	3	60	3	12	3

Table 4: Adult Income comparison of mean attributions. 95% CIs ranged from ±0.0004 (Cluster 2, *relationship*) to ±0.0115 (Cluster 5, *marital-status* and *age*).

Game Formulation	Size	Avg. Prediction (ϕ_0)	rel.	cap.	edu.	mar.	age
$v_{\boldsymbol{x}}^{inp}$	-	0.24	-0.04	-0.03	-0.01	-0.10	-0.00
$v_{\boldsymbol{x}}^{J.M.}$	-	0.19	-0.02	-0.03	-0.01	-0.08	0.01
$v_{\boldsymbol{x}}^{unif}$	-	0.82	0.01	-0.79	0.02	-0.03	0.04
Cluster 1	10.2%	0.67	-0.15	-0.01	-0.15	-0.28	-0.02
Cluster 2	55.3%	0.04	0.01	0.00	0.00	-0.01	0.02
Cluster 3	4.4%	0.99	-0.04	-0.70	-0.06	-0.12	-0.01
Cluster 4	28.0%	0.31	-0.09	0.00	0.08	-0.21	-0.03
Cluster 5	2.1%	0.67	-0.04	0.01	-0.47	-0.14	0.03

[8] In Bike Sharing we model hourly bike rentals from temporal and weather features, in Adult Income we model whether an adult earns more than $50,000 annually, and in Lending Club we model whether a borrower will default on a loan.

[9] See Supplemental Material for model details (`https://github.com/fiddler-labs/the-explanation-game-supplemental`).

Table 5: Lending Club comparison of mean attributions. 95% CIs ranged from ±0.0004 to ±0.0007 for both games.

Game Formulation	Avg. Prediction (ϕ_0)	fico.	addr.	inc.	acc.	dti
$v_{\boldsymbol{x}}^{inp}$	0.14	0.00	0.03	0.00	0.10	0.00
$v_{\boldsymbol{x},\mathcal{D}^{ref}}$	0.05	0.02	0.04	0.02	0.11	0.03

6.1.2 *Unquantified uncertainty*

Lack of uncertainty quantification in existing techniques can result in misleading attributions. For instance, taking the mean attribution of 100 randomly-sampled Bike Sharing single-reference games[10] gives *hr* an attribution of -10 and *workingday* an attribution of 8. Without any sense of uncertainty, we do not know how how accurate these (estimated) attributions are. The 95% confidence intervals we compute for these estimates in Section 6.2.2 show that they are uncertain indeed: the CIs span both positive and negative values.

6.1.3 *Irrelevant references*

In Section 5.1, we noted the importance of relevant references (or norms), and how the IME game $v_{\boldsymbol{x}}^{unif}$ based on the uniform distribution $\mathcal{U}$ can focus on irrelevant references. We illustrate this on the Adult Income example (see the third row of Table 7), where almost all attribution from the $v_{\boldsymbol{x}}^{unif}$ game falls on the *capitalgain* feature. This is surprising as *capitalgain* is zero for the example being explained, and for over 90% of individuals in the Adult Income dataset. The attributions are an artifact of uniformly sampling reference feature values, which causes nearly all references to have non-zero capital gain (as the probability of sampling an exactly zero capital gain is infinitesimal).

6.2 Applying the FAE framework

We now consider how the FAE framework enables meaningful explanations for the three models we study.

6.2.1 *Formulating contrastive questions*

A key benefit of FAE is that it enables explaining predictions relative to a selected group of references. For instance, in the Lending Club model, rather than asking "Why did our rejected example received a score of 0.28?" we ask the contrastive question "Why did our rejected example receive a score of 0.28 *relative to the examples that were accepted*?" This is a more apt question, as it explicitly discards irrelevant comparisons to other rejected applications. In terms of game formulation, the contrastive approach amounts to considering single-reference games where the reference is drawn from the distribution of accepted applications, rather than all applications. The attributions for each of these questions (shown in Table 8) turn out to be quite different. For instance, although number of recently-opened accounts (*acc*) is still the highest-attributed feature, we find that credit score (*fico*), income (*inc*), and debt-to-income ratio (*dti*) receive significantly higher attribution in the contrastive formulation. Without formulating the contrastive question, we would be misled into believing that these features are unimportant for the rejection.

6.2.2 *Quantifying uncertainty*

When summarizing the attribution distribution with the mean, confidence intervals (CIs) can be computed using the standard error of the mean (see Section 4.4). Returning to our Bike Sharing example, with 100 samples, the 95% confidence intervals for *hr* and *workingday* are -36 to 15, and -1 to 12, respectively. The large CIs caution us that 100 samples are perhaps too few. When using the full test set, the 95% CIs drop to 0.0 to 5.1 for *hr*, and 0.6 to 2.0 for *workingday*. It should be noted that even when exactly computing Shapley values and using thousands of points, the resulting attributions were still noticeably uncertain.

6.2.3 *Summarizing attribution distributions*

To demonstrate the power of unsupervised learning in explanation, we apply k-means clustering to the attributions of the Adult case study (with $k = 5$). This clustering identifies a large group of irrelevant references (cluster 2) which are similar to the explained point, demonstrating low attributions and predictions. Cluster 3 discovers the same pattern that the $v_{\boldsymbol{x}}^{unif}$ formulation did: high *capitalgain* causes extremely high scores. Since over 90% of points in the dataset have zero *capitalgain*, this pattern is "washed out in the average" of attributions for single-reference games sampled according to $\mathcal{D}^{inp}$ (as in KernelSHAP); see the first row of Table 7. On the other hand, the IME formulation identifies nothing but this pattern. Our clustering also helps identify other patterns. Clusters 1 and 5 show that when compared to references that obtain a high-but-not-extreme score, *marital-status*, *relationship*, and *education-num* are the primary factors accounting for the lower prediction score for the example at hand.

7 CONCLUSION

Our first contribution is an in-depth study of various Shapley-value-based model explanation methods. We find cases where existing methods yield counter-intuitive attributions, and we trace these misleading attributions to the cooperative games formulated by these methods. We propose a generalizing formulation that unifies attribution methods, offers conceptual clarity for interpreting each method's attributions, and admits straightforward confidence intervals for attributions.

Our second contribution is a framework for model explanations, called *formulate, approximate, and explain* (FAE), which is built on principles from Norm Theory [11]. We advise practitioners to *formulate* contrastive explanation questions that specify the references relative to which a prediction should be explained, for example "Why did this rejected loan application receive a score of 0.28 *compared to the applications that were accepted?*" By *approximating* the Shapley values of games formulated relative to the chosen references, and *explaining* the distribution of approximated Shapley values, we provide a meaningful and relevant answer to the explanation question at hand.

We conclude that axiomatic guarantees do not inherently guarantee relevant explanations, and that game formulations must be constructed carefully. In summarizing attribution distributions, we caution practitioners to avoid coarse-grained summaries that hide information, and to appropriately quantify any uncertainty resulting from the approximations used.

[10] The official implementation of KernelSHAP [14] raises a warning if over 100 references are used.

REFERENCES

[1] Kjersti Aas, Martin Jullum, and Anders Løland, 'Explaining individual predictions when features are dependent: More accurate approximations to shapley values', *arXiv preprint arXiv:1903.10464*, (2019).

[2] Marco Ancona, Enea Ceolini, Cengiz ztireli, and Markus Gross, 'Towards better understanding of gradient-based attribution methods for deep neural networks', in *International Conference on Learning Representations*, (2018).

[3] Marco Ancona, Cengiz Oztireli, and Markus Gross, 'Explaining deep neural networks with a polynomial time algorithm for shapley value approximation', in *Proceedings of the 36th International Conference on Machine Learning*, (2019).

[4] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan, 'L-shapley and c-shapley: Efficient model interpretation for structured data', *arXiv preprint arXiv:1808.02610*, (2018).

[5] Shay Cohen, Eytan Ruppin, and Gideon Dror, 'Feature selection based on the shapley value', *In other words*, **1**, 98Eqr, (2005).

[6] 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)*, pp. 598–617. IEEE, (2016).

[7] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Pai-Shun Ting, Karthikeyan Shanmugam, and Payel Das, 'Explanations based on the missing: Towards contrastive explanations with pertinent negatives', *CoRR*, (2018).

[8] Amirata Ghorbani and James Zou, 'Data shapley: Equitable valuation of data for machine learning', in *Proceedings of the 36th International Conference on Machine Learning*, (2019).

[9] Christopher Hitchcock and Joshua Knobecaus, 'Cause and norm', *Journal of Philosophy*, **106**(11), 587–612, (2009).

[10] Xin J Hunt, Ralph Abbey, Ricky Tharrington, Joost Huiskens, and Nina Wesdorp, 'An ai-augmented lesion detection framework for liver metastases with model interpretability', *arXiv preprint arXiv:1907.07713*, (2019).

[11] Daniel Kahneman and Dale T Miller, 'Norm theory: Comparing reality to its alternatives.', *Psychological review*, **93**(2), 136, (1986).

[12] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, 'Lightgbm: A highly efficient gradient boosting decision tree', in *Advances in Neural Information Processing Systems*, pp. 3146–3154, (2017).

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

[14] Scott M Lundberg and Su-In Lee, 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems*, pp. 4765–4774, (2017).

[15] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers, 'Bounding the estimation error of sampling-based shapley value approximation', *arXiv preprint arXiv:1306.4265*, (2013).

[16] 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*, pp. 1135–1144. ACM, (2016).

[17] Lloyd S Shapley, 'A value for n-person games', *Contributions to the Theory of Games*, **2**(28), 307–317, (1953).

[18] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje, 'Learning important features through propagating activation differences', in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3145–3153. JMLR. org, (2017).

[19] Erik Štrumbelj and Igor Kononenko, 'An efficient explanation of individual classifications using game theory', *Journal of Machine Learning Research*, **11**, 1–18, (2010).

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

[21] Mukund Sundararajan and Amir Najmi, 'The many shapley values for model explanation', *arXiv preprint arXiv:1908.08474*, (2019).

[22] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, 'Axiomatic attribution for deep networks', in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3319–3328. JMLR. org, (2017).

[23] H. P. Young, 'Monotonic solutions of cooperative games', *International Journal of Game Theory*, **14**, 65–72, (1985).

Supplemental Material: The Explanation Game

8 Shapley Value Axioms

We briefly summarize the four Shapley value axioms.

- The *Dummy* axiom requires that if player i has no possible contribution (i.e. $v(S \cup \{i\}) = v(S)$ for all $S \subseteq \mathcal{M}$), then that player receives zero attribution.
- The *Symmetry* axiom requires that two players that always have the same contribution receive equal attribution, Formally, if $v(S \cup \{i\}) = v(S \cup \{j\})$ for all S not containing i or j then $\phi_i(v) = \phi_j(v)$.
- The *Efficiency* axiom requires that the attributions to all players sum to the total payoff of all players. Formally, $\sum_i \phi_i(v) = v(\mathcal{M})$).
- The *Linearity* axiom states that for any payoff function v that is a linear combination of two other payoff functions u and w (i.e. $v(S) = \alpha u(S) + \beta w(S)$), the Shapley values of v equal the corresponding linear combination of the Shapley values of u and w (i.e. $\phi_i(v) = \alpha\phi_i(u) + \beta\phi_i(w)$).

9 Additional Shapley value approximations

9.0.1 Marginal contribution sampling

We can express the Shapley value of a player as the expected value of the weighted marginal contribution to a random coalition S sampled uniformly from all possible coalitions excluding that player, rather than an exhaustive weighted sum. A sampling estimator of this expectation is by nature unbiased, so this can be used as an alternative to the permutation estimator in approximating attributions with confidence intervals.

$$\phi_i(v) = \mathop{\mathbb{E}}_{S}\left[\frac{2^{M-1}}{M}\binom{M-1}{|S|}^{-1}\left(v(S \cup \{i\}) - v(S)\right)\right] \tag{13}$$

Equation 13 can be approximated with a Monte Carlo estimate, i.e. by sampling from the random S and averaging the quantity within the expectation.

9.0.2 Weighted least squares

The Shapley values are the solution to a certain weighted least squares optimization problem which was popularized through its use in the KernelSHAP algorithm. For a full explanation, see `https://arxiv.org/abs/1903.10464`.

$$\boldsymbol{\phi} = \arg\min_{\boldsymbol{\phi}} \sum_{S \subseteq \mathcal{M}} \frac{M-1}{\binom{M}{|S|}|S|(M-|S|)}\left(v(S) - \sum_{i=1}^{M}\phi_i\right)^2 \tag{14}$$

The fraction in the left of Equation 14 is often referred to as the Shapley kernel . In practice, an approximate objective function is minimized. The approximate objective is defined as a summation over squared error on a sample of coalitions rather than over squared error on all possible coalitions. Additionally, the "KernelSHAP trick" may be employed, wherein sampling is performed according to the Shapley kernel (rather than uniformly), and the least-squares optimization is solved with uniform weights (rather than Shapley kernel weights) to account for the adjusted sampling.

To the best of our knowledge, there exists no proof that the solution to a subsampled objective function of the form in Equation 14 is an estimator (unbiased or otherwise) of the Shapley values. In practice, it does appear that subsampling down to even a small fraction of the total number of possible coalitions (weighted by the Shapley kernel or uniformly) does a good job of estimating the Shapley values for explanation games. Furthermore, approximation errors in such experiments do not yield signs of bias. However, we do note that using the weighted least squares approximation with our confidence interval equation does inherently imply an unproved assumption that it is an ubiased estimator.

10 Proofs

In what follows, we prove the lemmas from the main paper. The proofs refer to equations and definition from the main paper.

10.1 Proof of Lemma 1

From the definitions of $v_{\boldsymbol{x},\mathcal{D}^{ref}}$ (Equation 7) and $v_{\boldsymbol{x},\boldsymbol{r}}$ (Equation 8), it follows that $v_{\boldsymbol{x},\mathcal{D}^{ref}}(S) = \mathbb{E}_{\boldsymbol{R}\sim\mathcal{D}}\left[v_{\boldsymbol{x},\boldsymbol{R}}(S)\right]$. Thus, the game $v_{\boldsymbol{x},\mathcal{D}^{ref}}$ is a linear combination of games $\{v_{\boldsymbol{x},\boldsymbol{r}} \mid \boldsymbol{r} \in \mathcal{X}\}$ (with weights defined by the distribution $\mathcal{D}$). From the Linearity axiom of Shapley values, it follows that the Shapley values of the game $v_{\boldsymbol{x},\mathcal{D}^{ref}}$ must be corresponding Shapley values of the games $v_{\boldsymbol{x},\boldsymbol{R}}$, and therefore, $\boldsymbol{\phi}(v_{\boldsymbol{x},\mathcal{D}^{ref}}) = \mathbb{E}_{\boldsymbol{R}\sim\mathcal{D}}\left[\boldsymbol{\phi}(v_{\boldsymbol{x},\boldsymbol{R}})\right]$.

10.2 Proof of Lemma 2

From Lemma 1, we have $\phi_i(v_{\boldsymbol{x},\mathcal{D}^{ref}}) = \mathbb{E}_{\boldsymbol{R}\sim\mathcal{D}}\left[\phi_i(v_{\boldsymbol{x},\boldsymbol{R}})\right]$. Thus, to prove this lemma, it suffices to show that for any irrelevant feature i, the Shapley value from the game $v_{\boldsymbol{x},\boldsymbol{r}}$ is zero for all references $r \in \mathcal{X}$. That is,

$$\forall \boldsymbol{r} \in \mathcal{X} \ \phi_i(v_{\boldsymbol{x},\boldsymbol{r}}) = 0 \tag{15}$$

From the definition of Shapley values (Equation 1), we have:

$$\phi_i(v_{\boldsymbol{x},\boldsymbol{r}}) = \frac{1}{M} \sum_{S\subseteq\mathcal{M}\setminus\{i\}} \binom{M-1}{|S|}^{-1} (v_{\boldsymbol{x},\boldsymbol{r}}(S\cup\{i\}) - v_{\boldsymbol{x},\boldsymbol{r}}(S)) \tag{16}$$

Thus, to prove Equation 15 it suffices to show the marginal contribution $(v_{\boldsymbol{x},\boldsymbol{r}}(S\cup\{i\}) - v_{\boldsymbol{x},\boldsymbol{r}}(S))$ of an irrelevant feature i to any subset of features $S \subseteq \mathcal{M}\setminus\{i\}$ is always zero. From the definition of the game $v_{\boldsymbol{x},\boldsymbol{r}}$, we have:

$$v_{\boldsymbol{x},\boldsymbol{r}}(S\cup\{i\}) - v_{\boldsymbol{x},\boldsymbol{r}}(S) = f(\boldsymbol{z}(\boldsymbol{x},\boldsymbol{r},S\cup\{i\})) - f(\boldsymbol{z}(\boldsymbol{x},\boldsymbol{r},S)) \tag{17}$$

From the definition of composite inputs $\boldsymbol{z}$ (Equation 2), it follows that the inputs $\boldsymbol{z}(\boldsymbol{x},\boldsymbol{r},S\cup\{i\})$ and $\boldsymbol{z}(\boldsymbol{x},\boldsymbol{r},S)$ agree on all features except i. Thus, if feature i is irrelevant, $f(\boldsymbol{z}(\boldsymbol{x},\boldsymbol{r},S\cup\{i\})) = f(\boldsymbol{z}(\boldsymbol{x},\boldsymbol{r},S))$, and consequently by Equation 16, $v_{\boldsymbol{x},\boldsymbol{r}}(S\cup\{i\}) - v_{\boldsymbol{x},\boldsymbol{r}}(S) = 0$. Thus feature i has zero marginal contribution to all subsets $S \subseteq \mathcal{M}\setminus\{i\}$ in the game $v_{\boldsymbol{x},\boldsymbol{r}}$. Combining this with the definition of Shapley values (Equation 1) proves Equation 15.

11 Reproducibility

For brevity, we omitted from the main paper many of the mundane choices in the design of our toy examples and case studies. To further transparency and reproducibility, we include them here.

11.1 Fitting models

For both case studies, we used the LightGBM package configured with default parameters to fit a Gradient Boosted Decision Trees (GBDT) model.

For the Bike Sharing dataset, we fit on all examples from 2011 while holding out the 2012 examples for testing. We omitted the *atemp* feature, as it is highly correlated to *temp* ($r = 0.98$), and the *instant* feature because the tree-based GBDT model cannot capture its time-series trend. For parsimony, we refitted the model to the top five most important features by cumulative gain (*hr*, *temp*, *workingday*, *hum*, and *season*). This lowered test-set r^2 from 0.64 to 0.63.

For the Adult Income dataset, we used the pre-defined train/test split. Again, we refitted the model to the top five features by cumulative gain feature importance (*relationship*, *capitalgain*, *education-num*, *marital-status*, and *age*). This increased test-set misclassification error from 14.73% to 10.97%.

11.2 Selection of points to explain

For the Bike Share case study, we sampled ten points at random from the test set. We selected one whose prediction was close to the middle of the range observed over the entire test set (predictions ranged approximately from 0 to 600). Specifically, we selected instant 11729 (2012-05-08, 9pm). We examined other points from the same sample of ten to suggest a random but meaningful comparitive question. We found another point with comparable *workingday*, *hum*, and *season*: instant 11362. This point caught our eye because it differed only in *hr* (2pm rather than 9pm), and *temp* (0.36 rather than 0.64) but had a much lower prediction.

For the Adult Income case study, we wanted to explain why a point was scored as likely to have low income, a task roughly analogous to that of explaining why an application for credit is rejected by a creditworthiness model in a lending setting. We sampled points at random with scores between 0.01 and 0.1, and chose the 9880th point in the test set due to its strikingly high *education-num* (most of the low-scoring points sampled had lower *education-num*).

For the Lending Club data, we chose an open-source subset of the dataset that has been pre-cleaned to a predictive task on 3-year loans. For the five-feature model, we selected the top five features by cumulative gain feature importance from a model fit to the full set of features.

11.3 K-means clustering

We choose $k = 5$ arbitrarily, having observed a general tradeoff of conciseness for precision as k increases. In the extremes, $k = 1$ maintains the overall attribution distribution, while $k = N$ examines each single-reference game separately.

12 Case Study Supplemental Material

Here we present the full results of the case studies, including tables and boxplot visualizations of attribution distributions.

[h]

Table 6: Bike Sharing comparison of mean attributions. 95% CIs ranged from ±0.4 (*hum* in $\mathcal{D}^{inp}$ and $\mathcal{D}^{J.M.}$) to ±2.5 (*hr* in $\mathcal{D}^{inp}$ and $\mathcal{D}^{J.M.}$).

Game Formulation	Size	Avg. Prediction (ϕ_0)	hr	temp	work.	hum	season
$v_{\boldsymbol{x}}^{inp}$	100%	151	3	47	1	7	2
$v_{\boldsymbol{x}}^{J.M.}$	100%	141	6	50	1	9	3
$v_{\boldsymbol{x}}^{unif}$	100%	128	3	60	3	12	3
Cluster 1	12.9%	309	-86	14	-28	3	-1
Cluster 2	27.6%	28	140	32	0	9	0
Cluster 3	10.5%	375	-247	58	16	9	-1
Cluster 4	32.5%	131	31	38	3	4	2
Cluster 5	16.5%	128	-57	107	13	9	9

[h]

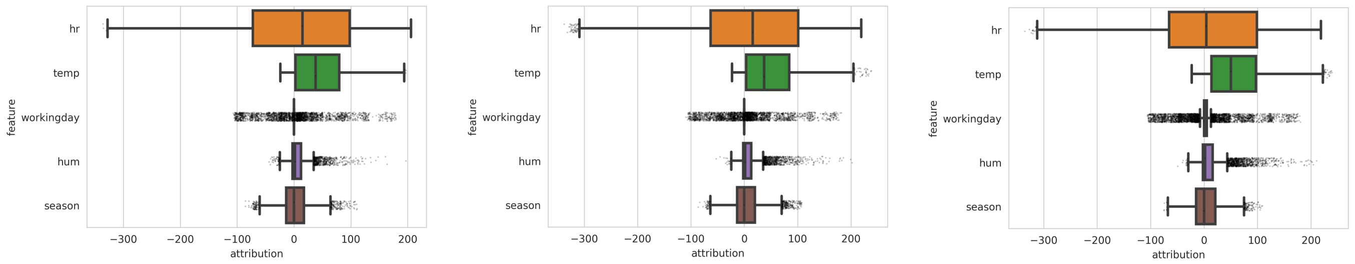
Table 7: Adult Income comparison of mean attributions. 95% CIs ranged from ±0.0004 (Cluster 2, *relationship*) to ±0.0115 (Cluster 5, *marital-status* and *age*).

Game Formulation	Size	Avg. Prediction (ϕ_0)	rel.	cap.	edu.	mar.	age
$v_{\boldsymbol{x}}^{inp}$	100%	0.24	-0.04	-0.03	-0.01	-0.10	-0.00
$v_{\boldsymbol{x}}^{J.M.}$	100%	0.19	-0.02	-0.03	-0.01	-0.08	0.01
$v_{\boldsymbol{x}}^{unif}$	100%	0.82	0.01	-0.79	0.02	-0.03	0.04
Cluster 1	10.2%	0.67	-0.15	-0.01	-0.15	-0.28	-0.02
Cluster 2	55.3%	0.04	0.01	0.00	0.00	-0.01	0.02
Cluster 3	4.4%	0.99	-0.04	-0.70	-0.06	-0.12	-0.01
Cluster 4	28.0%	0.31	-0.09	0.00	0.08	-0.21	-0.03
Cluster 5	2.1%	0.67	-0.04	0.01	-0.47	-0.14	0.03

[h]

Table 8: Lending Club comparison of mean attributions. 95% CIs ranged from ±0.0004 to ±0.0007 for both games.

Game Formulation	Size	Avg. Prediction (ϕ_0)	fico.	addr.	inc.	acc.	dti
$v_{\boldsymbol{x},\mathcal{D}^{ref}}$	20%	0.05	0.02	0.04	0.02	0.11	0.03
$v_{\boldsymbol{x}}^{inp}$	100%	0.14	0.00	0.03	0.00	0.10	0.00
$v_{\boldsymbol{x}}^{J.M.}$	100%	0.14	0.01	0.03	0.01	0.10	0.00
$v_{\boldsymbol{x}}^{unif}$	100%	0.11	0.05	0.07	-0.01	0.03	0.02
Cluster 1	28.5%	0.11	0.01	0.06	0.00	0.08	0.01
Cluster 2	24.4%	0.10	0.01	0.00	0.01	0.11	0.04
Cluster 3	15.4%	0.18	0.00	0.01	0.00	0.14	-0.05
Cluster 4	17.6%	0.16	-0.01	0.01	0.03	0.09	-0.01
Cluster 5	14.0%	0.22	-0.01	0.05	-0.02	0.08	-0.06

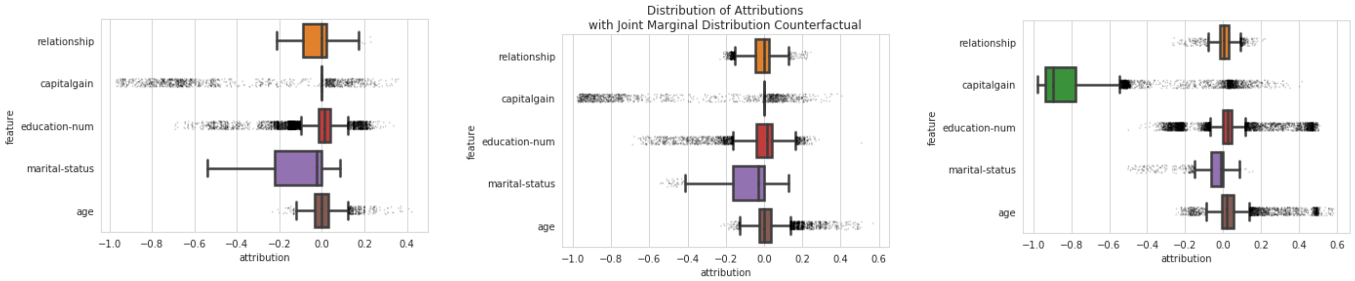


(a) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{D}^{inp}$.

(b) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{D}^{J.M.}$.

(c) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{U}$.

Figure 2: Bike Sharing attributions for decompositions of $v_{\boldsymbol{x}}^{inp}$, $v_{\boldsymbol{x}}^{J.M.}$, and $v_{\boldsymbol{x}}^{unif}$.

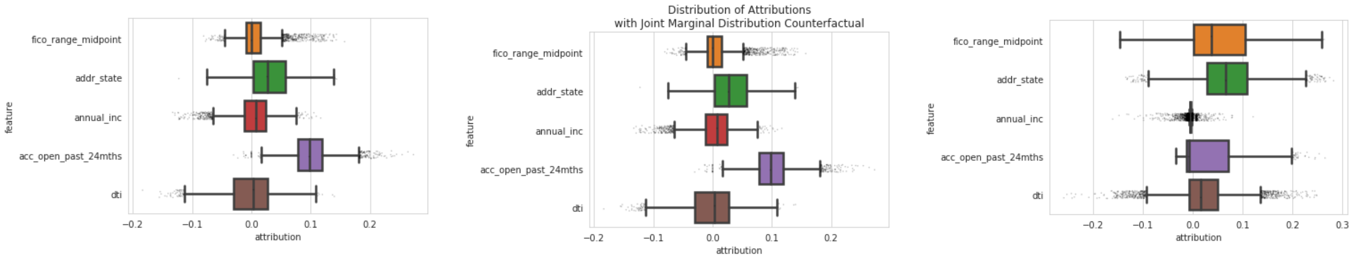


(a) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{D}^{inp}$.

(b) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{D}^{J.M.}$.

(c) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{U}$.

Figure 3: Adult Income attributions for decompositions of $v_{\boldsymbol{x}}^{inp}$, $v_{\boldsymbol{x}}^{J.M.}$, and $v_{\boldsymbol{x}}^{unif}$.

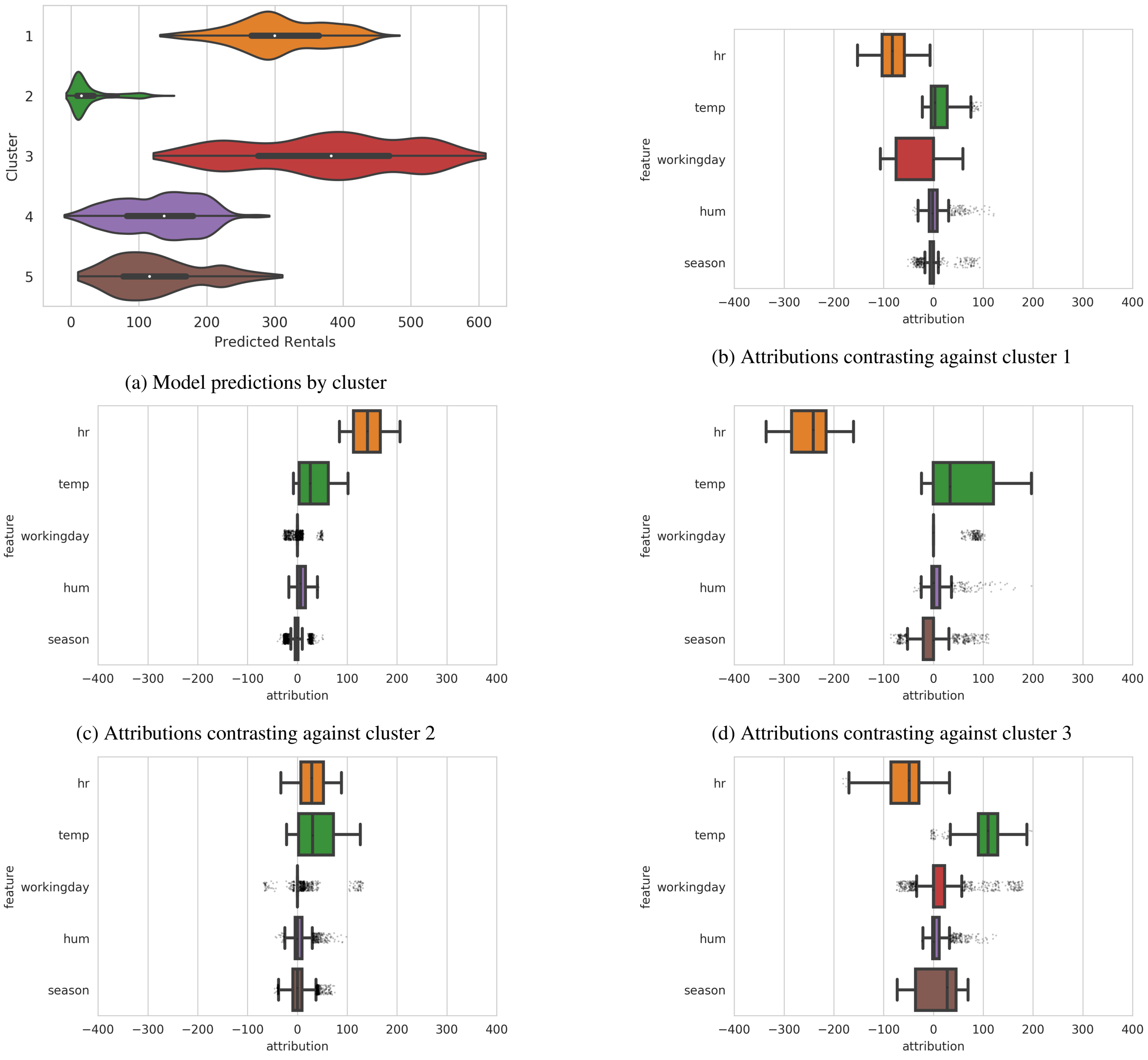


(a) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{D}^{inp}$.

(b) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{D}^{J.M.}$.

(c) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{U}$.

Figure 4: Lending Club attributions for decompositions of $v_{\boldsymbol{x}}^{inp}$, $v_{\boldsymbol{x}}^{J.M.}$, and $v_{\boldsymbol{x}}^{unif}$.



(a) Model predictions by cluster

(b) Attributions contrasting against cluster 1

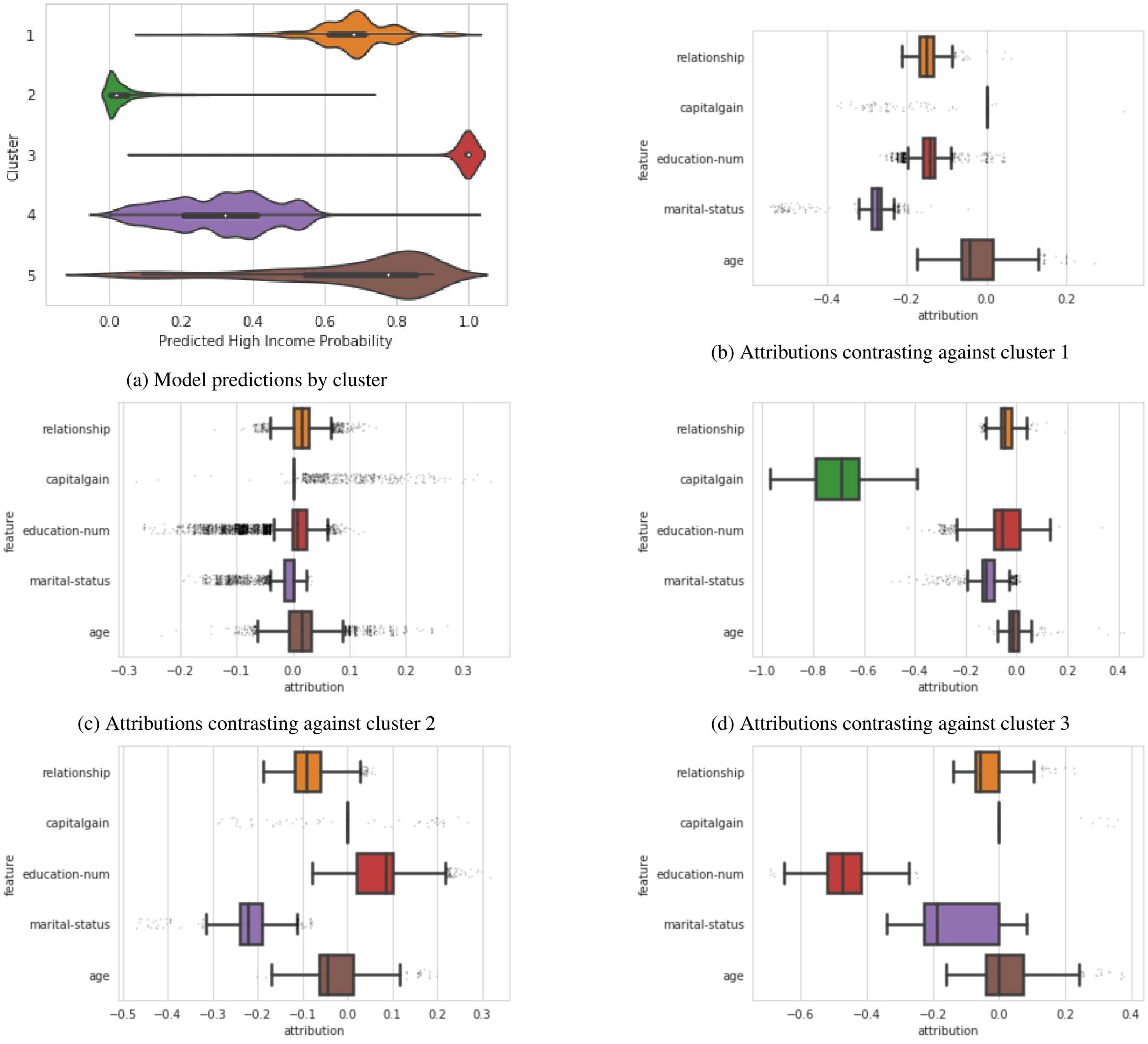
(c) Attributions contrasting against cluster 2

(d) Attributions contrasting against cluster 3

(e) Attributions contrasting against cluster 4

(f) Attributions contrasting against cluster 5

Figure 5: Bike Sharing predictions by cluster and attributions from contrastive games against counterfactual clusters.



(a) Model predictions by cluster

(b) Attributions contrasting against cluster 1

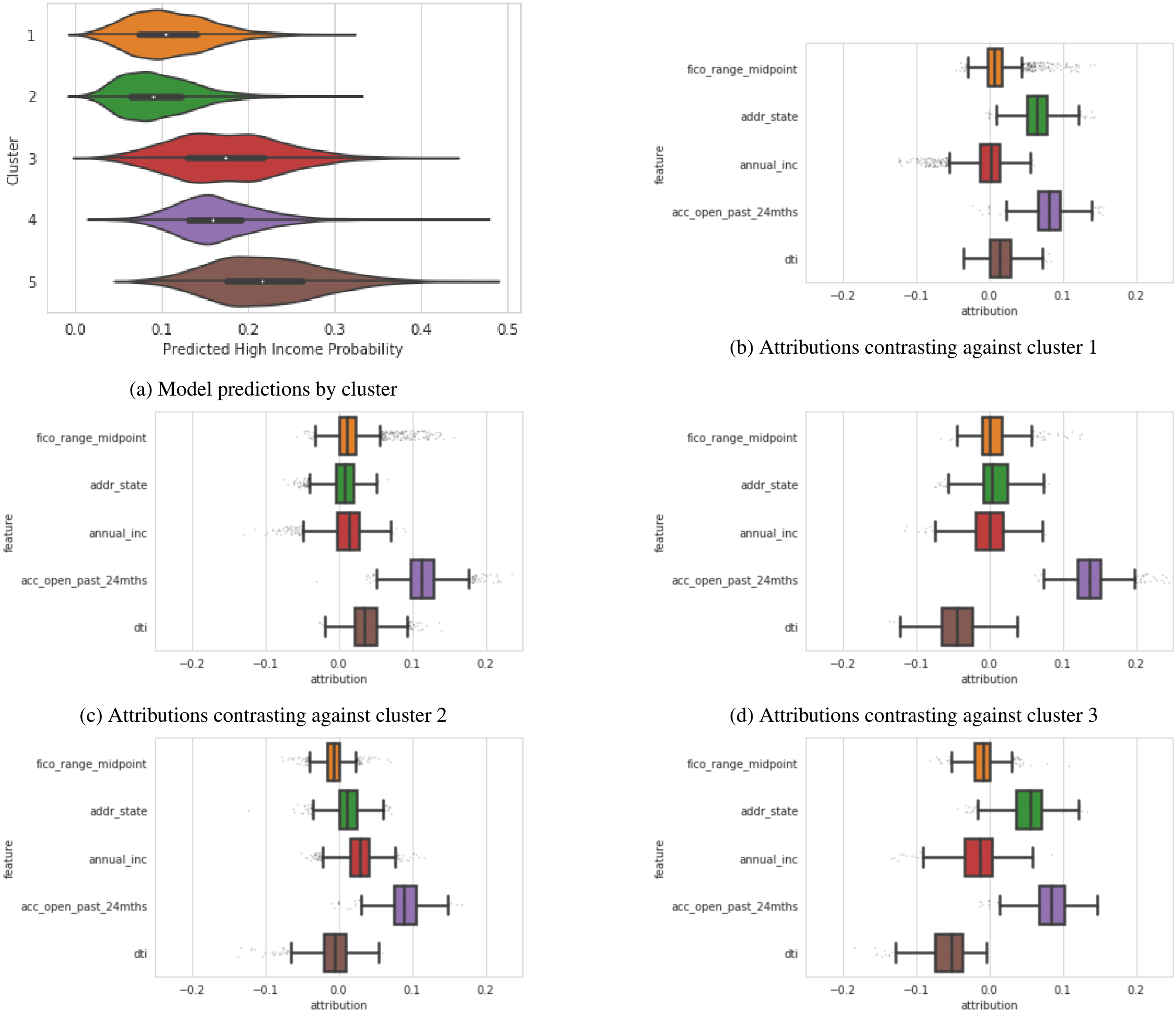
(c) Attributions contrasting against cluster 2

(d) Attributions contrasting against cluster 3

(e) Attributions contrasting against cluster 4

(f) Attributions contrasting against cluster 5

Figure 6: Adult Income predictions by cluster and attributions from contrastive games against counterfactual clusters.



(a) Model predictions by cluster

(b) Attributions contrasting against cluster 1

(c) Attributions contrasting against cluster 2

(d) Attributions contrasting against cluster 3

(e) Attributions contrasting against cluster 4

(f) Attributions contrasting against cluster 5

Figure 7: Lending Club predictions by cluster and attributions from contrastive games against counterfactual clusters.