Supplemental Material: The Explanation Game

1 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. $\phi_i(v) = \alpha\phi_i(u) + \beta\phi_i(w)$).

2 Additional Shapley value approximations

Marginal contribution sampling An equivalent formulation to Equation **??** expresses 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.

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

Equation 1 can be approximated with a Monte Carlo estimate, i.e. by sampling from the random S and averaging the quantity within the expectation.

Weighted least squares Lastly we consider the weighted least squares approximation of the Shapley values. The Shapley values are the solution to a certain weighted least squares optimization problem originally presented in (**?**), later popularized through its use in the KernelSHAP algorithm of (**?**), and fully explained in (**?**).

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

The fraction in the left of Equation 2 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.

3 Proofs

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

3.1 Proof of Lemma 1

From the definitions of $v_{\boldsymbol{x},\mathcal{D}}$ (Equation 7) and $v_{\boldsymbol{x},\boldsymbol{r}}$ (Equation 8), it follows that $v_{\boldsymbol{x},\mathcal{D}}(S) = \mathbb{E}_{\boldsymbol{R}\sim\mathcal{D}}\left[v_{\boldsymbol{x},\boldsymbol{R}}(S)\right]$. Thus, the game $v_{\boldsymbol{x},\mathcal{D}}$ 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 $v_{\boldsymbol{x},\mathcal{D}}$ must be corresponding Shapley values of the games $v_{\boldsymbol{x},\boldsymbol{R}}$, and therefore, $\boldsymbol{\phi}(v_{\boldsymbol{x},\mathcal{D}}) = \mathbb{E}_{\boldsymbol{R}\sim\mathcal{D}}\left[\boldsymbol{\phi}(v_{\boldsymbol{x},\boldsymbol{R}})\right]$.

3.2 Proof of Lemma 2

From Lemma 1, we have $\phi_i(v_{\boldsymbol{x},\mathcal{D}}) = \mathbb{E}_{\boldsymbol{R}\sim\mathcal{D}}\left[\phi_i(v_{\boldsymbol{x},\boldsymbol{R}})\right]$. 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,

$$\forall \boldsymbol{r} \in \mathcal{X}\ \phi_i(v_{\boldsymbol{x},\boldsymbol{r}}) = 0 \tag{10}$$

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

$$\phi_i(v_{\boldsymbol{x},\boldsymbol{r}}) = \frac{1}{M}\sum_{S \subseteq \mathcal{M}\setminus\{i\}}\binom{M-1}{|S|}^{-1}\left(v_{\boldsymbol{x},\boldsymbol{r}}(S \cup \{i\}) - v_{\boldsymbol{x},\boldsymbol{r}}(S)\right) \tag{11}$$

Thus, to prove Equation 10 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:

$$v_{\boldsymbol{x},\boldsymbol{r}}(S \cup \{i\}) - v_{\boldsymbol{x},\boldsymbol{r}}(S) = f(\boldsymbol{z}(\boldsymbol{x},\boldsymbol{r}, S \cup \{i\})) - f(\boldsymbol{z}(\boldsymbol{x},\boldsymbol{r}, S)) \tag{12}$$

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 11, $v_{\boldsymbol{x},\boldsymbol{r}}(S \cup \{i\}) - v_{\boldsymbol{x},\boldsymbol{r}}(S) = 0$. 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 10.

4 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.

4.1 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%.

4.2 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.

4.3 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.

5 Case Study Supplemental Material

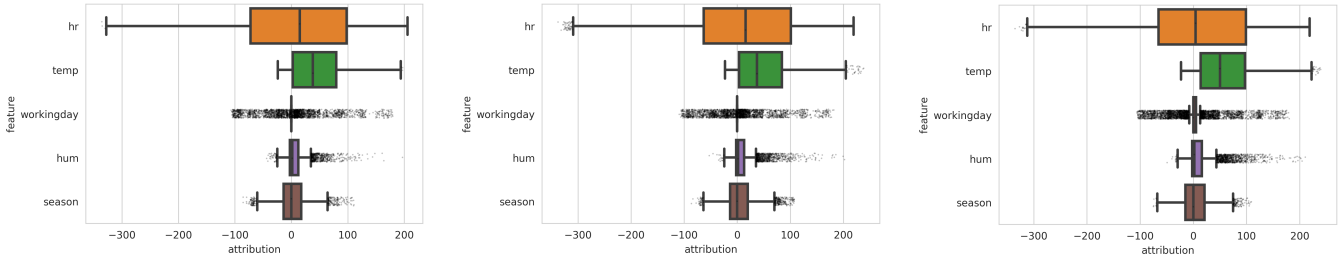
Here we present the full results of the case studies, including tables and boxplot visualizations of attribution distributions.

Table 1: 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
$v_{\boldsymbol{x}}^{\text{J.M.}}$	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

Table 2: 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
$v_{\boldsymbol{x}}^{\text{J.M.}}$	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

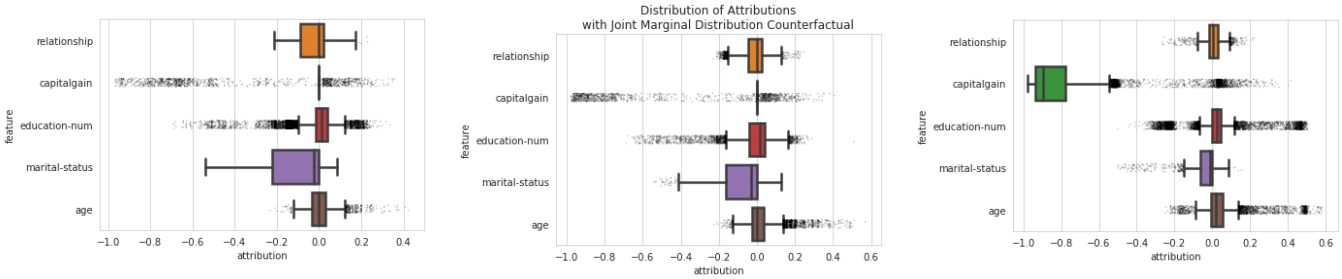


(a) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{D}^{inp}$. (b) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{D}^{J.M.}$. (c) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{U}$.

Figure 1: Bike Sharing attributions for decompositions of $v_{\boldsymbol{x}}^{inp}$, $v_{\boldsymbol{x}}^{\text{J.M.}}$, and $v_{\boldsymbol{x}}^{unif}$.

Table 3: 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
$v_{\boldsymbol{x},\mathcal{D}}$	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
$v_{\boldsymbol{x}}^{\text{J.M.}}$	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

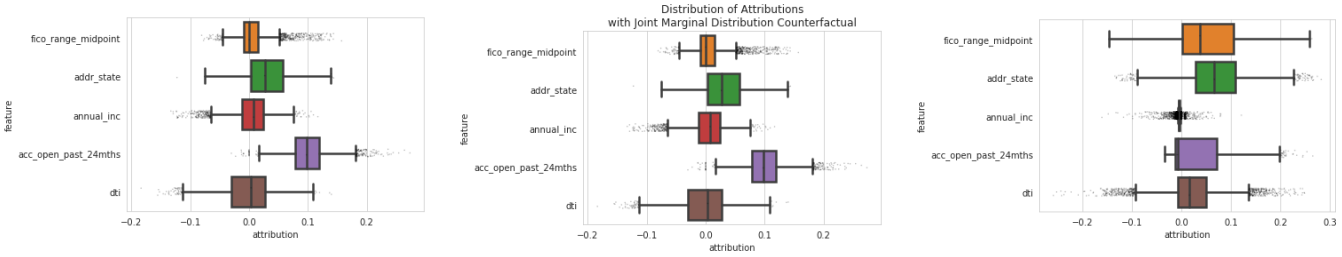


(a) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{D}^{inp}$.

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

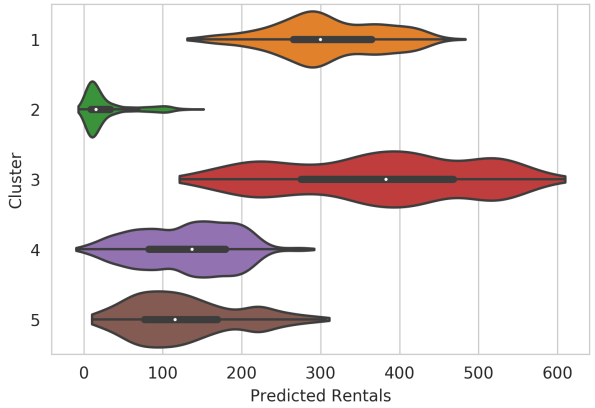


(a) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{D}^{inp}$.

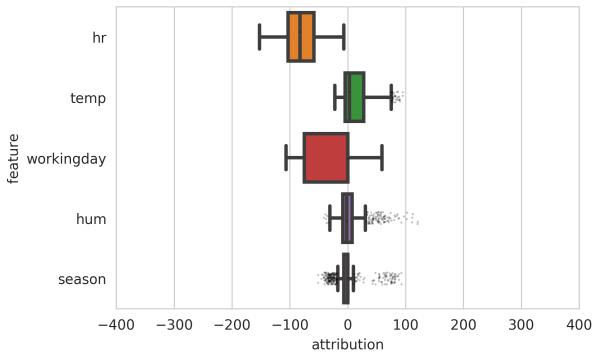
(b) Attributions for $v_{\boldsymbol{x},\boldsymbol{R}}$ when $\boldsymbol{R} \sim \mathcal{D}^{J.M.}$.

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

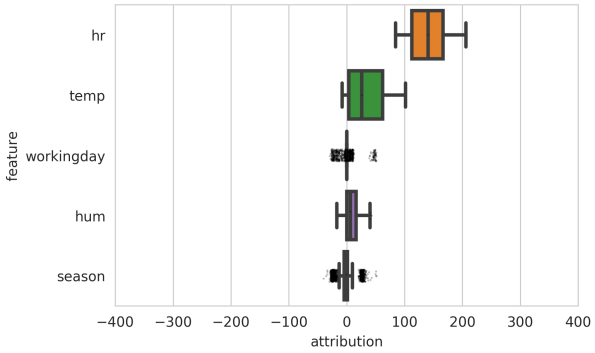
Figure 3: Lending Club attributions for decompositions of $v_{\boldsymbol{x}}^{inp}$, $v_{\boldsymbol{x}}^{\text{J.M.}}$, and $v_{\boldsymbol{x}}^{unif}$.



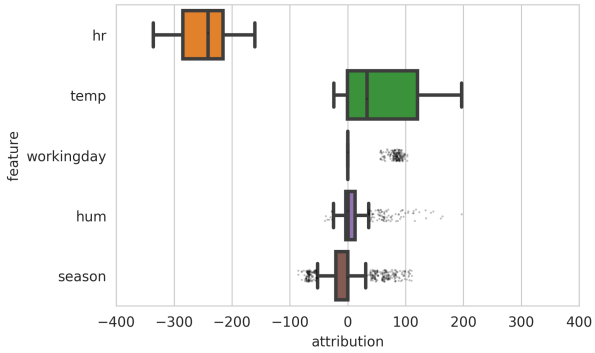
(a) Model predictions by cluster



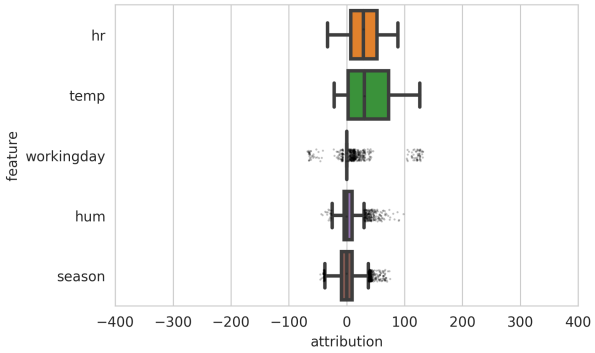
(b) Attributions contrasting against cluster 1



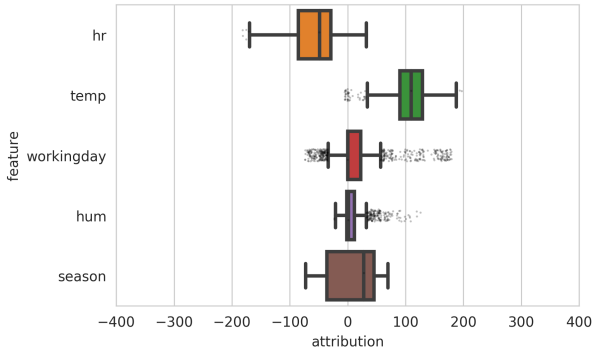
(c) Attributions contrasting against cluster 2



(d) Attributions contrasting against cluster 3

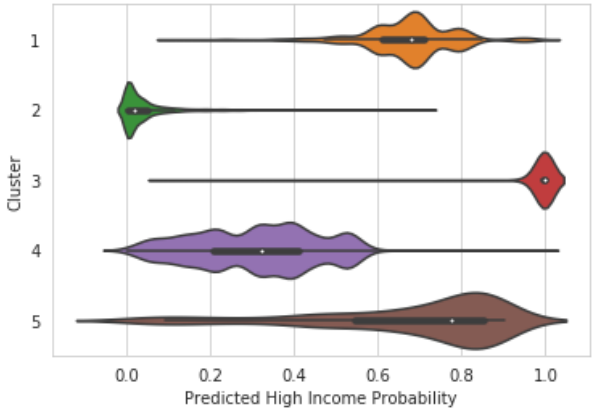


(e) Attributions contrasting against cluster 4

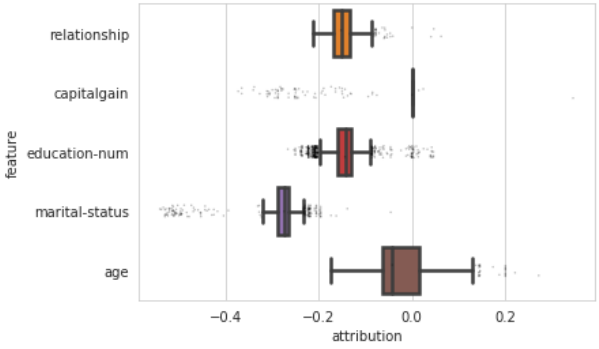


(f) Attributions contrasting against cluster 5

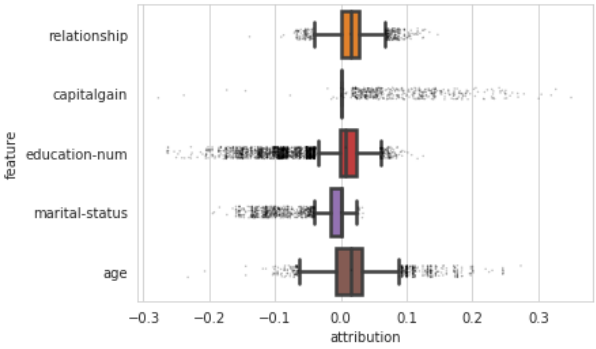
Figure 4: Bike Sharing predictions by cluster and attributions from contrastive games against counterfactual clusters.



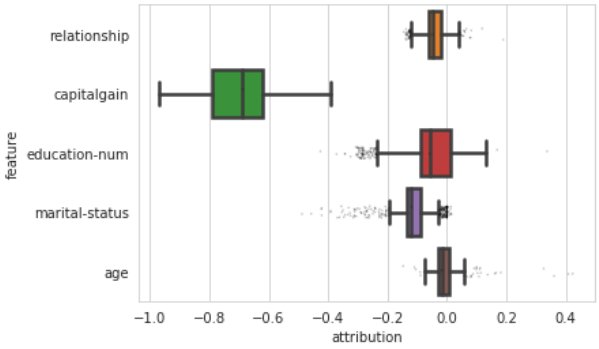
(a) Model predictions by cluster



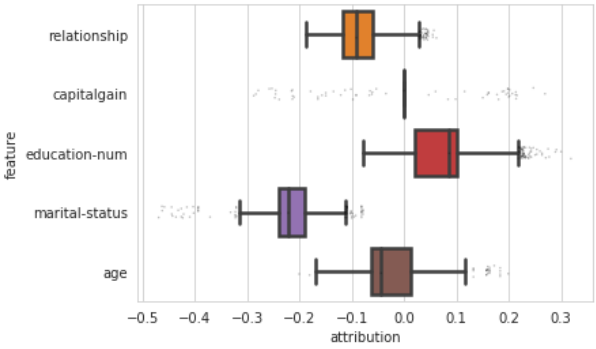
(b) Attributions contrasting against cluster 1



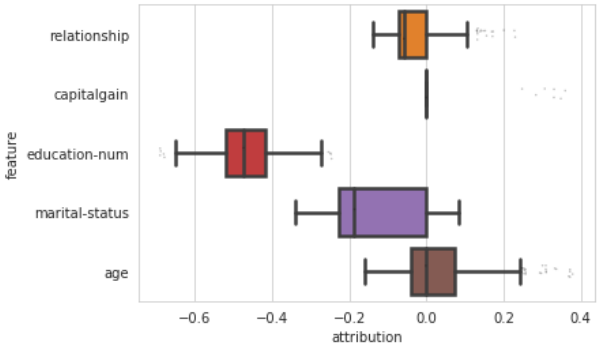
(c) Attributions contrasting against cluster 2



(d) Attributions contrasting against cluster 3

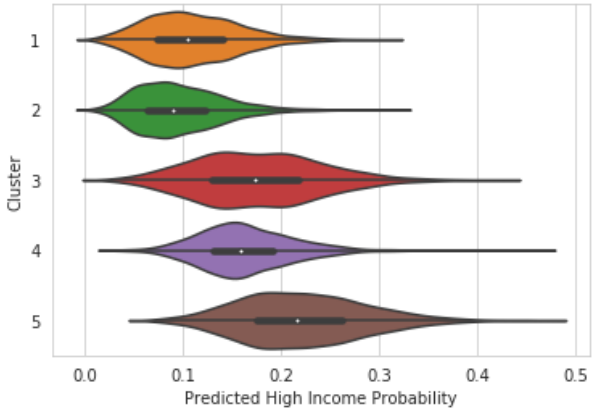


(e) Attributions contrasting against cluster 4

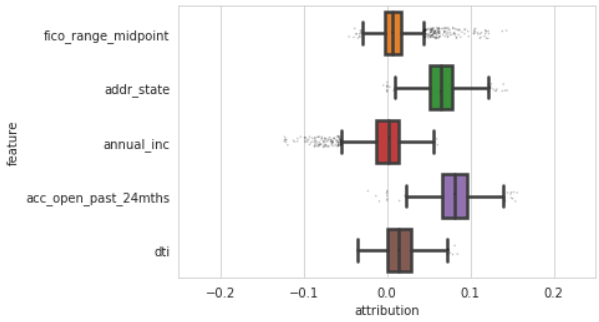


(f) Attributions contrasting against cluster 5

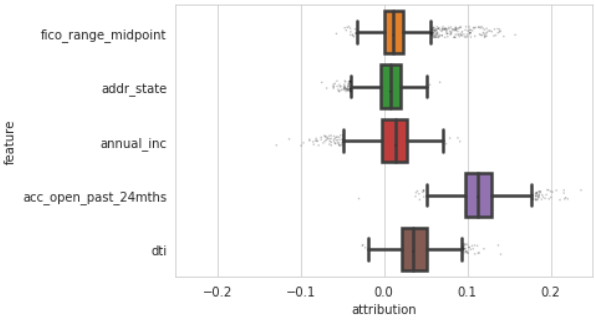
Figure 5: Adult Income predictions by cluster and attributions from contrastive games against counterfactual clusters.



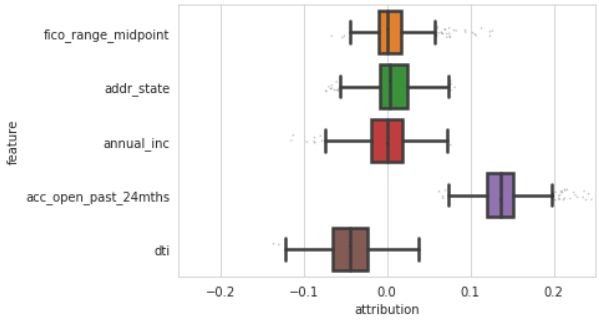
(a) Model predictions by cluster



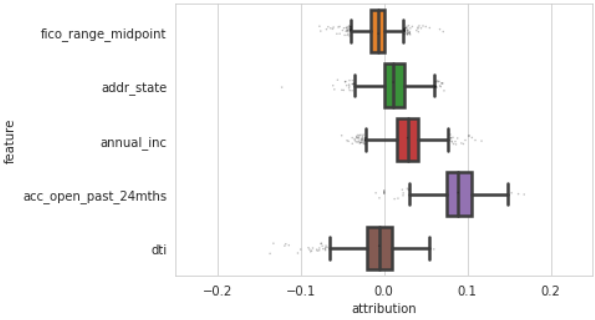
(b) Attributions contrasting against cluster 1



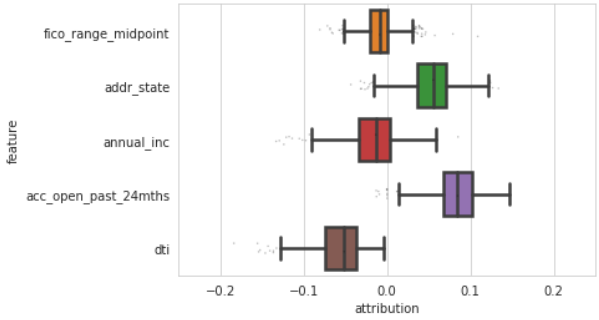
(c) Attributions contrasting against cluster 2



(d) Attributions contrasting against cluster 3



(e) Attributions contrasting against cluster 4



(f) Attributions contrasting against cluster 5

Figure 6: Lending Club predictions by cluster and attributions from contrastive games against counterfactual clusters.