Manifold Restricted Interventional Shapley Values



Muhammad Faaiz Taufiq, Patrick Bloebaum, Lenon Minorics

Motivation

Example

A bank uses a predictive model to predict the credit worthiness of loan applicants, based on their data.

Features for each applicant include race, gender, annual income, age, etc.

Suppose applicant A is denied a loan.

Motivation

Example

A bank uses a predictive model to predict the credit worthiness of loan applicants, based on their data.

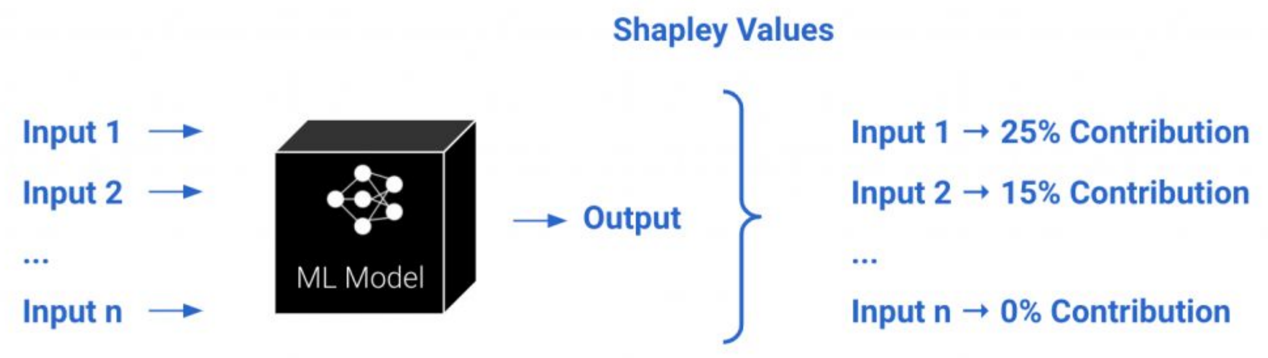
Features for each applicant include race, gender, annual income, age, etc.

Suppose applicant A is denied a loan.

Why did the model deny loan to applicant A?

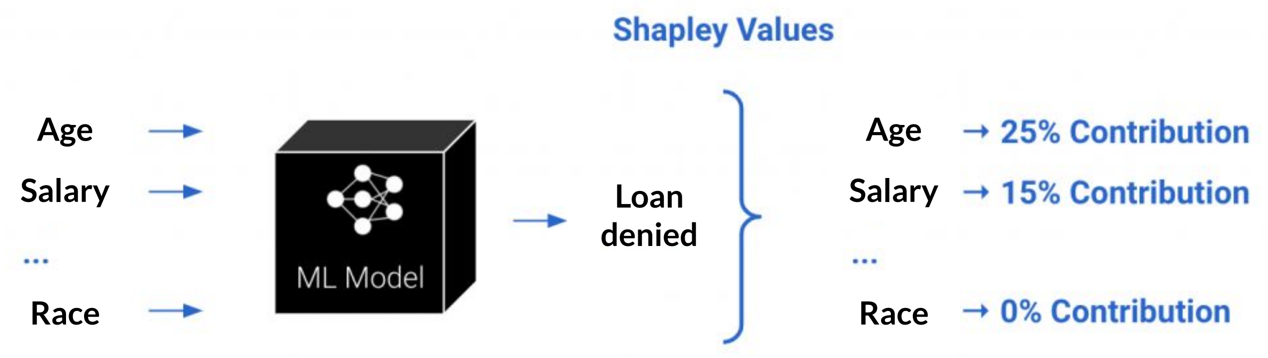
Shapley values provide such explanations.

Shapley values help us quantify how much each feature contributes towards the model output.



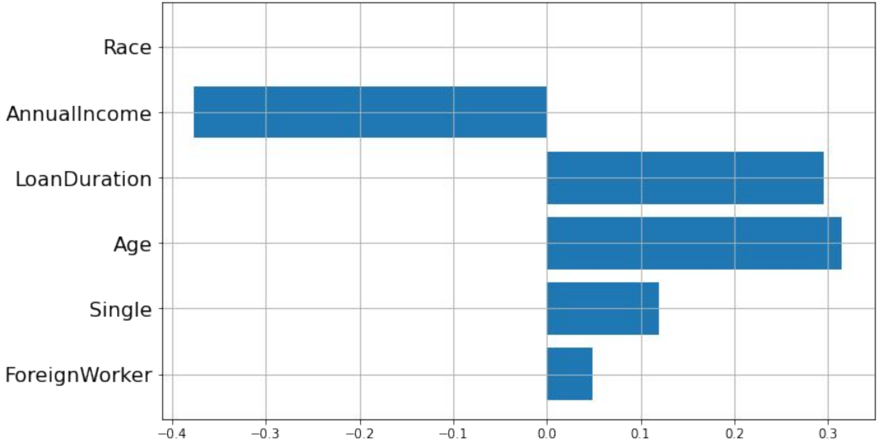
Shapley values provide such explanations.

Shapley values help us quantify how much each feature contributes towards the model output.



Shapley values provide such explanations.

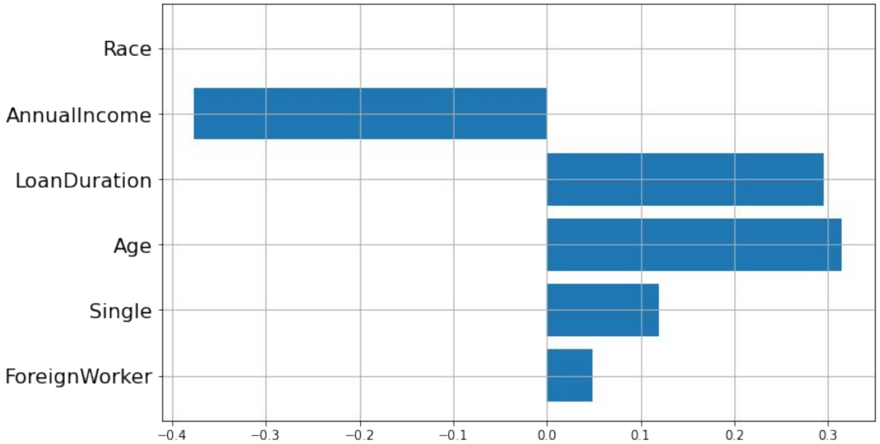
Shapley values help us quantify how much each feature contributes towards the model output.



Shapley values provide such explanations.

A prediction can be explained by assuming that each feature value of the instance is a "player" in a game where the prediction is the payout.

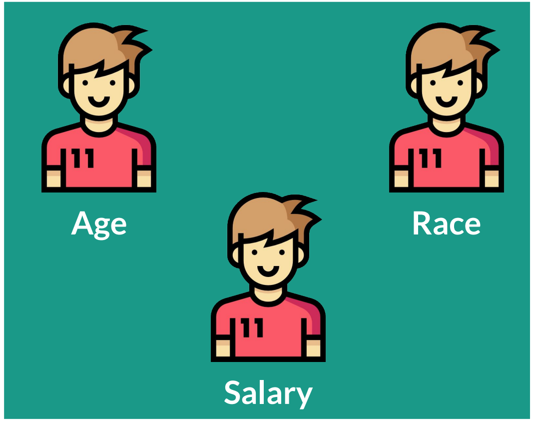
Shapley values – a method from coalitional game theory – tells us how to fairly distribute the "payout" among the features.



Background

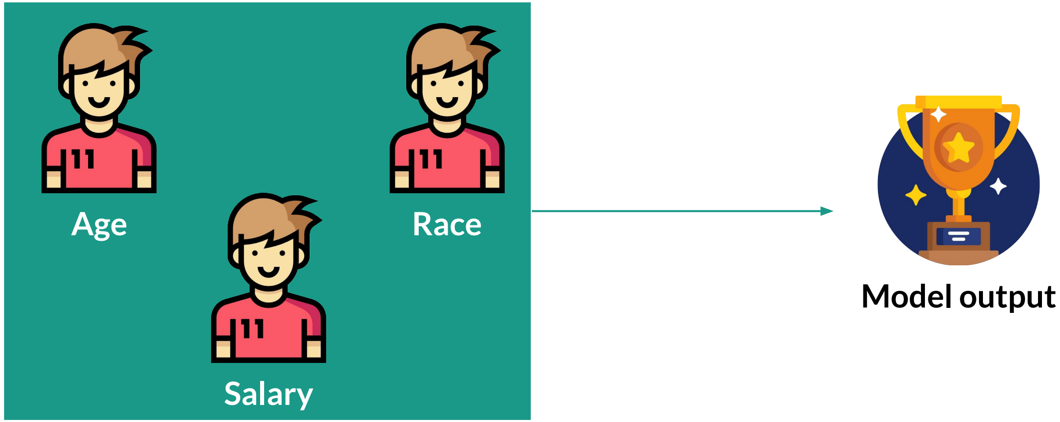
Shapley values: How do they work?

- Each feature value of the instance is a "player" in a game



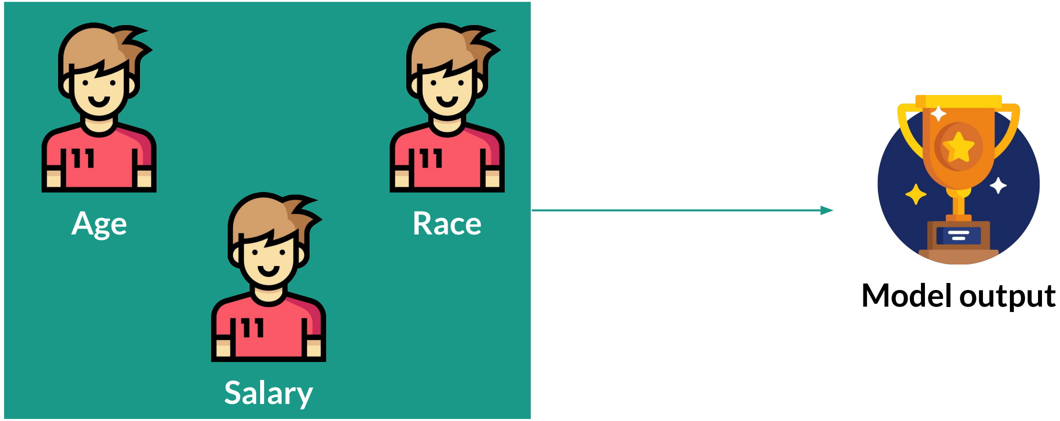
Shapley values: How do they work?

- Each feature value of the instance is a "player" in a game
- Where the total payout is the prediction made by the model



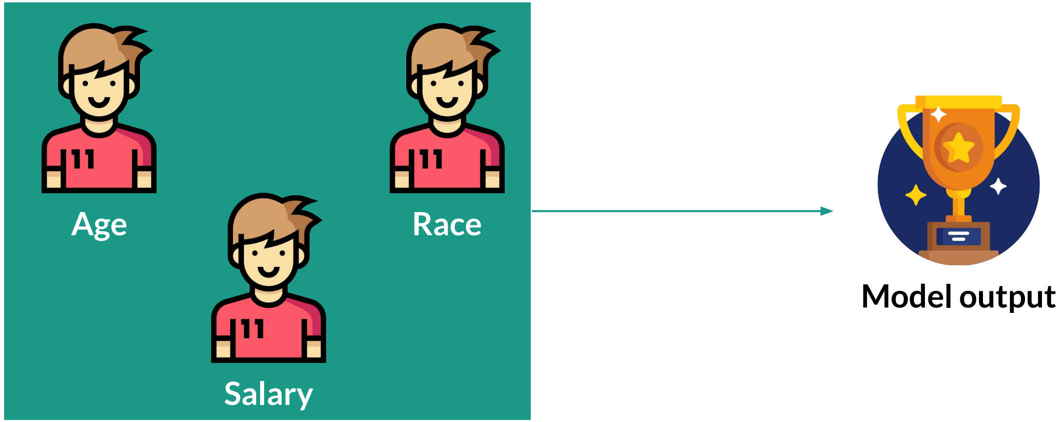
Shapley values: How do they work?

- Each feature value of the instance is a "player" in a game
- Where the total payout is the prediction made by the model
- This requires a value function $\mathcal{V}$ which quantifies the value of a group of players.



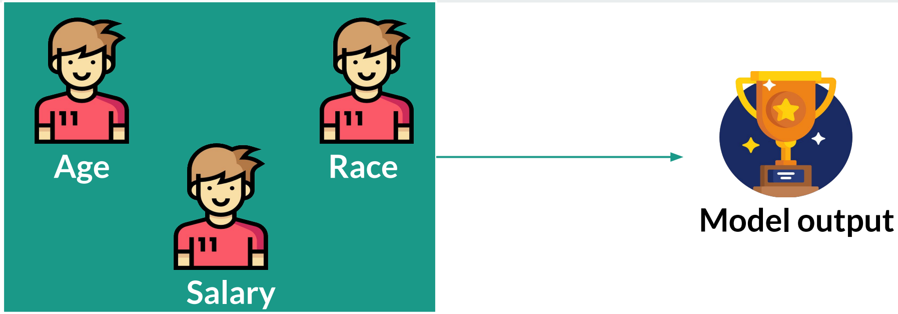
Shapley values: How do they work?

- This requires a value function $\mathcal{V}$ which quantifies the value of a group of players.



Value functions

$$v : 2^{[d]} \to \mathbb{R}$$



Value functions

- The Shapley value is defined via a value function v of players in S

$$v : 2^{[d]} \to \mathbb{R}$$

Value functions

- The Shapley value is defined via a value function v of players in S

$$v : 2^{[d]} \to \mathbb{R}$$

The value of the feature subset *S* = *{Age, Salary}*.

Value functions

- The Shapley value is defined via a value function v of players in S

$$v : 2^{[d]} \to \mathbb{R}$$

The value of the feature subset *S* = *{Age, Salary}*.

- $v(S \cup \{j\}) - v(S)$ can be intuitively interpreted as the contribution of feature j w.r.t. the set S.

Shapley values

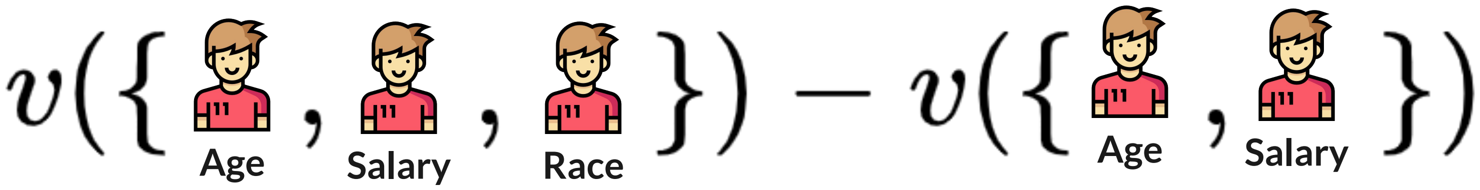
$$v(\{\text{Age}, \text{Salary}, \text{Race}\}) - v(\{\text{Age}, \text{Salary}\})$$

Shapley values

$$v(\{\text{Age}, \text{Salary}, \text{Race}\}) - v(\{\text{Age}, \text{Salary}\})$$

The value of the feature ***race*** w.r.t. the set ***{Age, Salary}***

Shapley values



The value of the feature ***race*** w.r.t. the set ***{Age, Salary}***

- The Shapley value of feature ***i*** is defined as a weighted sum over all possible subsets *S*:

$$\phi_i := \sum_{S \subseteq [d] \setminus \{i\}} \frac{|S|!(d - |S| - 1)!}{d!} (v(S \cup \{i\}) - v(S))$$

Shapley values

- The Shapley value is defined via a value function v of players in S

$$v : 2^{[d]} \rightarrow \mathbb{R}$$

- $v(S \cup \{j\}) - v(S)$ can be intuitively interpreted as the contribution of feature j w.r.t. the set S.

Shapley values

- The Shapley value is defined via a value function v of players in S

$$v : 2^{[d]} \to \mathbb{R}$$

- $v(S \cup \{j\}) - v(S)$ can be intuitively interpreted as the contribution of feature j w.r.t. the set S.
- The Shapley value of feature *j* is defined as a weighted sum over all possible subsets *S*:

$$\phi_i \coloneqq \sum_{S \subseteq [d] \setminus \{i\}} \frac{|S|!(d - |S| - 1)!}{d!} (v(S \cup \{i\}) - v(S))$$

Shapley values

- The Shapley value is defined via a value function v of players in S

$$v : 2^{[d]} \to \mathbb{R}$$

- $v(S \cup \{j\}) - v(S)$ can be intuitively interpreted as the contribution of feature j w.r.t. the set S.
- The Shapley value of feature *j* is defined as a weighted sum over all possible subsets *S*:

$$\phi_i \coloneqq \sum_{S \subseteq [d] \setminus \{i\}} \frac{|S|!(d - |S| - 1)!}{d!} (v(S \cup \{i\}) - v(S))$$

- The Shapley value can be thought of as the average contribution of a feature value to the prediction among different subsets.

Types of value functions

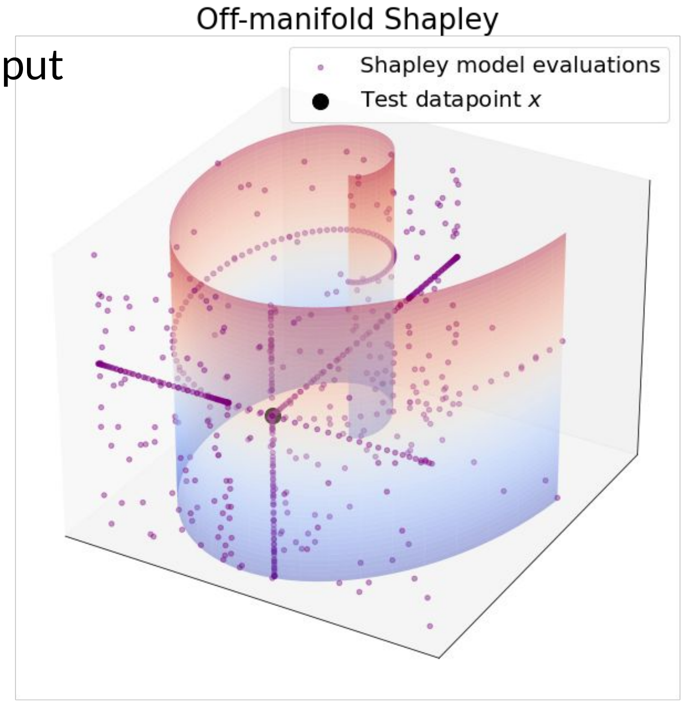
Types of value functions

Value functions can be broadly classified into:

1. Off-Manifold value functions
2. On-Manifold value functions

Off-manifold Shapley values

Relies on function evaluations on out-of-distribution input samples when computing Shapley explanations.



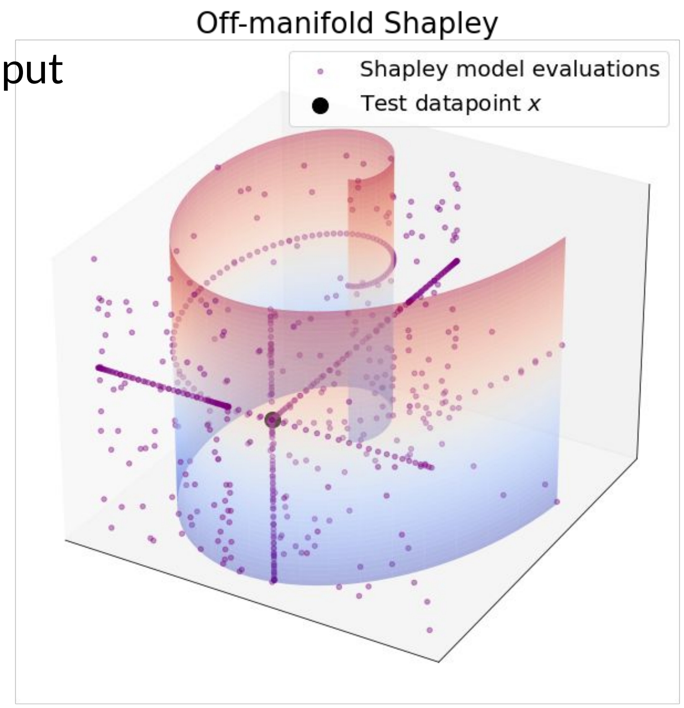
Off-manifold Shapley values

Relies on function evaluations on out-of-distribution input samples when computing Shapley explanations.

Examples:

- Marginal Shapley:

$$v^{\mathrm{MS}}_{\mathbf{x},f}(S) := \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}})]$$



Off-manifold Shapley values

Relies on function evaluations on out-of-distribution input samples when computing Shapley explanations.

Examples:

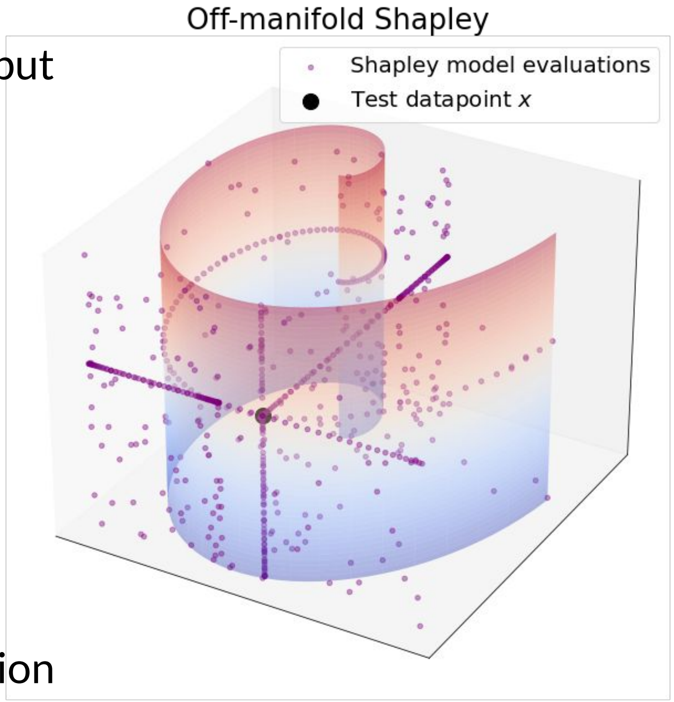
- Marginal Shapley:

$$v^{\mathrm{MS}}_{\mathbf{x},f}(S) := \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}})]$$

- Interventional Shapley:

$$v^{\mathrm{IS}}_{\mathbf{x},f}(S) := \mathbb{E}[f(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S)]$$

Interventional Shapley estimates the "causal" contribution of features towards the overall prediction.



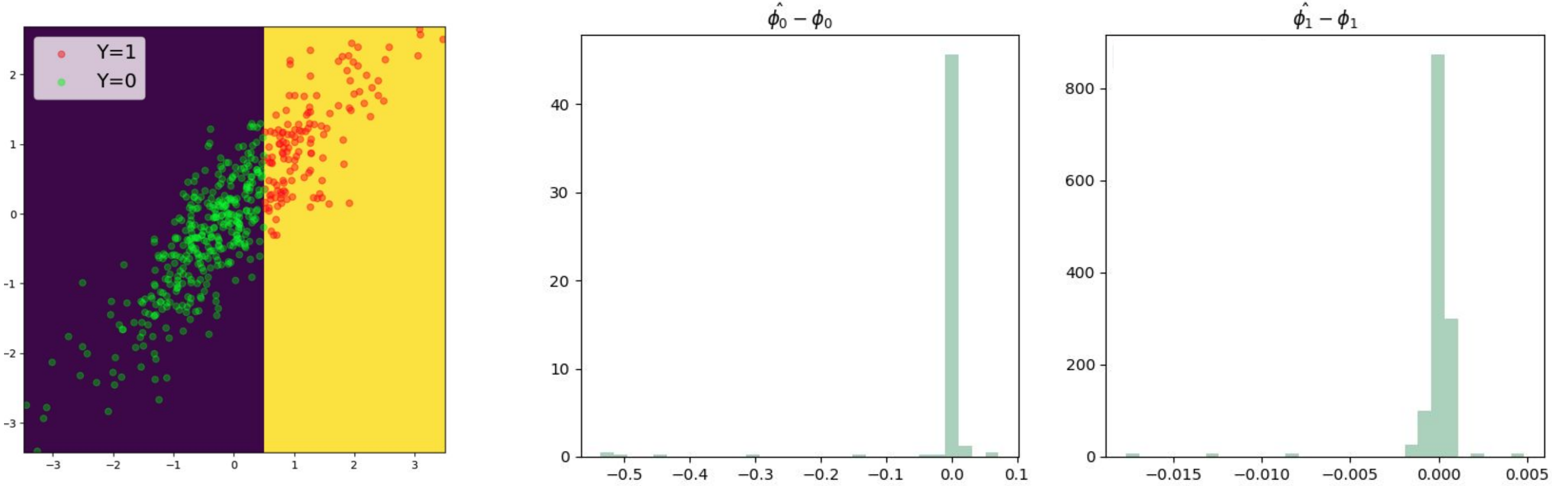
Off-manifold Shapley values: Limitation

Interventional Shapley evaluate the function outside it’s domain of validity, where it hasn’t been trained.

Off-manifold Shapley values: Limitation

Interventional Shapley evaluate the function outside it's domain of validity, where it hasn't been trained.

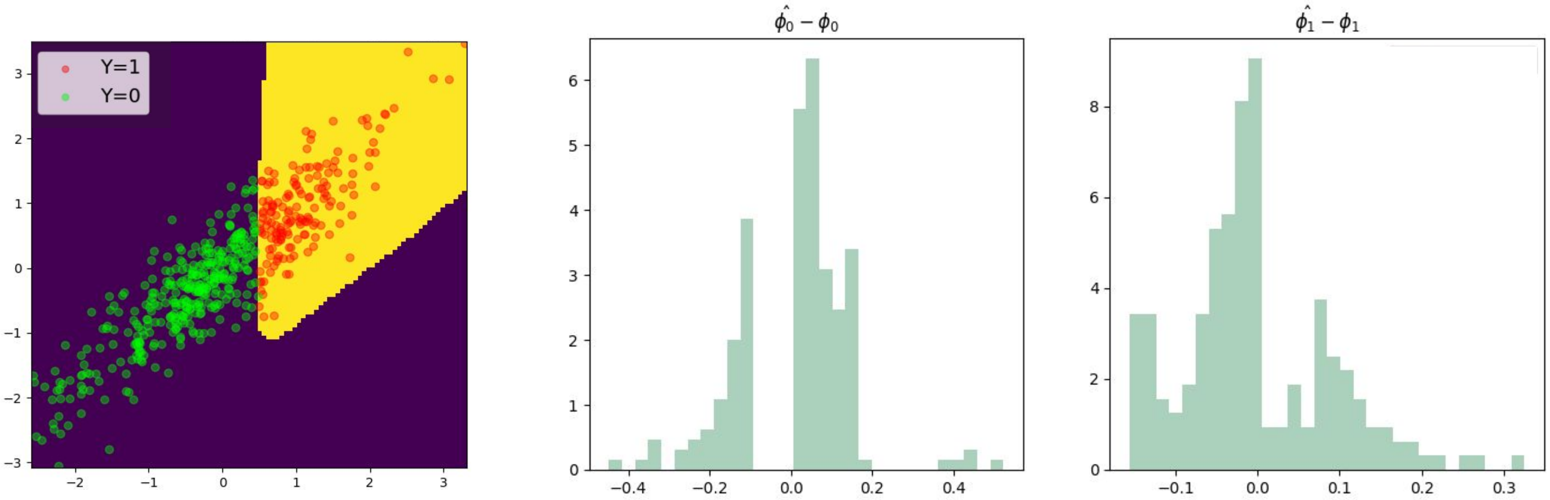
Ground Truth Classifier: $1(X_1 \geq 1/2)$. Accuracy of trained classifier: 100%



Off-manifold Shapley values: Limitation

Perturbing the model outside data manifold can drastically change the Shapley values, even though it remains constant on-manifold.

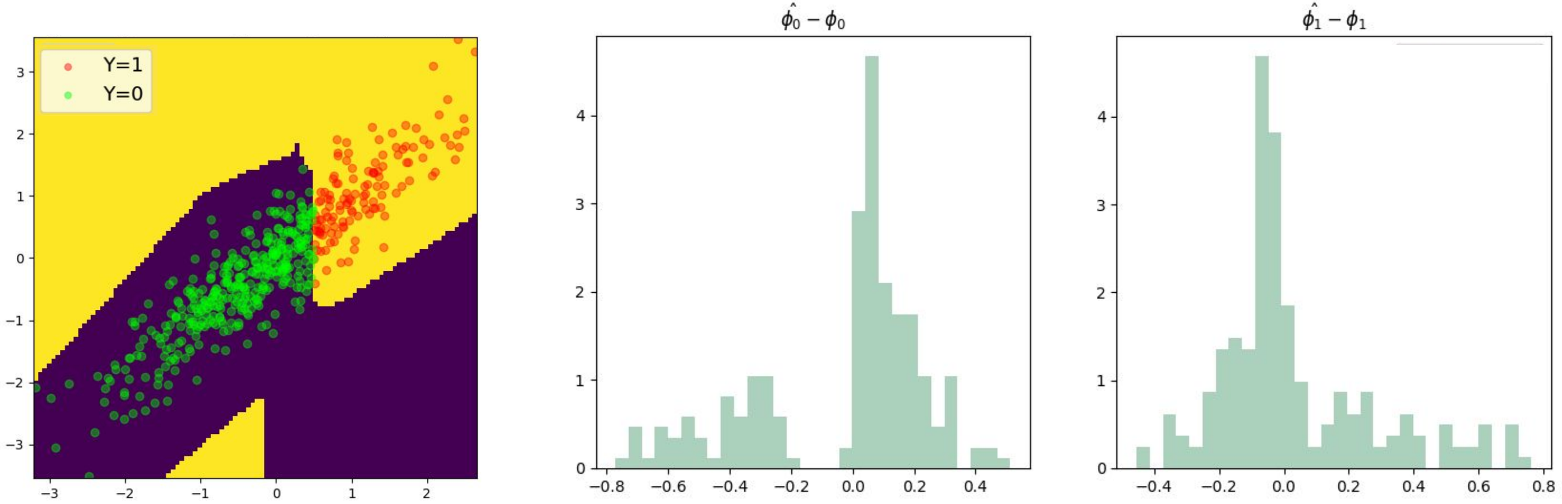
Accuracy of trained classifier: 100%



Off-manifold Shapley values: Limitation

Perturbing the model outside data manifold can drastically change the Shapley values, even though it remains constant on-manifold.

Accuracy of trained classifier: 100%



Off-manifold Shapley values: Limitation

The Shapley computations are heavily influenced by function behaviour outside the data manifold.

This can lead to misleading Shapley values.

On-manifold Shapley values

Does not rely on function behaviour outside the data distribution when computing Shapley explanations.

On-manifold Shapley values

Does not rely on function behaviour outside the data distribution when computing Shapley explanations.

Examples:

- Conditional Expectation Shapley:

$$v_{\mathbf{x},f}^{\text{CES}}(S) := \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$$

On-manifold Shapley values

Does not rely on function behaviour outside the data distribution when computing Shapley explanations.

Examples:

- Conditional Expectation Shapley:

$$v^{\text{CES}}_{\mathbf{x},f}(S) := \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$$

- Random Joint Baseline Shapley:

$$v^{\text{RJ}}_{\mathbf{x},f,p}(S) := \mathbb{E}_{p_b(\mathbf{X}_{\bar{S}})}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}})p(\mathbf{x}_S, \mathbf{X}_{\bar{S}})]$$

On-manifold Shapley values: Limitation

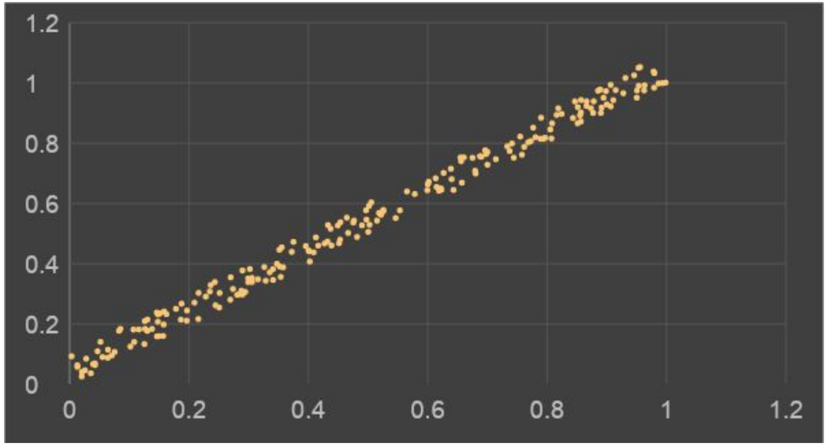
On-Manifold Shapley values are often highly dependent on feature correlations.

On-manifold Shapley values: Limitation

On-Manifold Shapley values are often highly dependent on feature correlations.

Example: $f(x_1, x_2) = x_1$

X_2



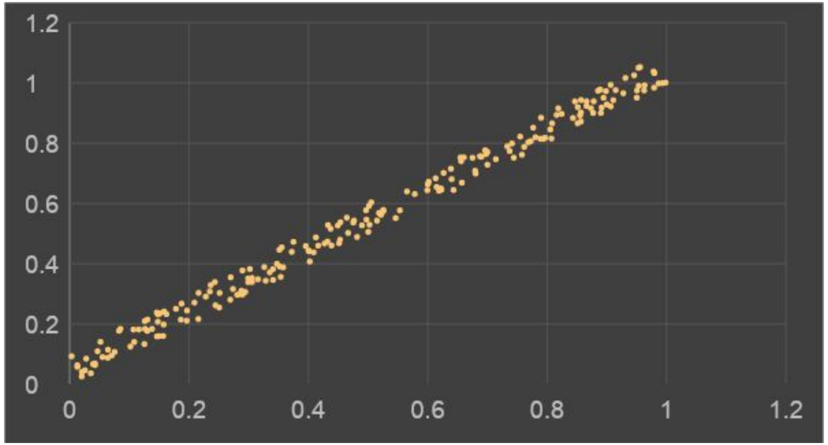
X_1

On-manifold Shapley values: Limitation

On-Manifold Shapley values are often highly dependent on feature correlations.

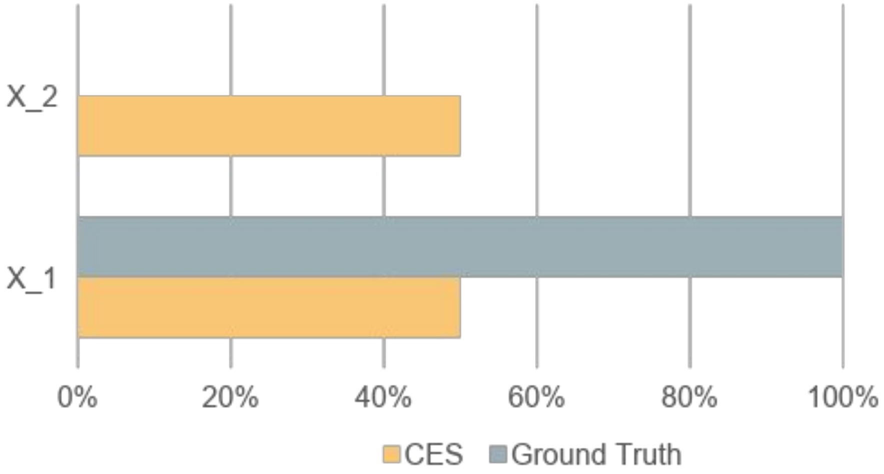
Example: $f(x_1, x_2) = x_1$

X_2



X_1

Feature importance (Conditional Shapley values)



Problem Statement

We seek to propose a methodology which:

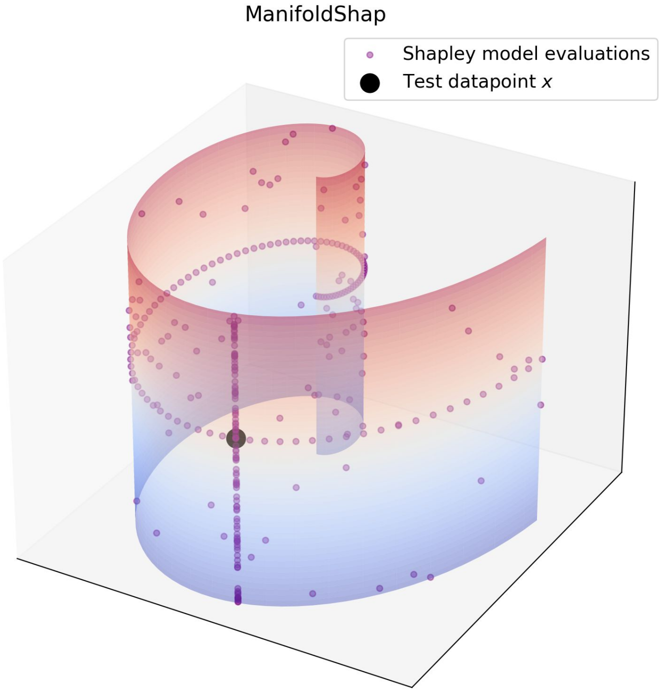
1. Is robust to off-manifold perturbations in the model.
2. Provides causally accurate model explanations.

Proposed Method: ManifoldShap

ManifoldShap

We propose ManifoldShap, a value function, which restricts function evaluations to the regions of interest Z.

We call these regions of interest Z, the 'manifold'.



ManifoldShap

We propose ManifoldShap, a value function, which restricts function evaluations to the regions of interest Z.

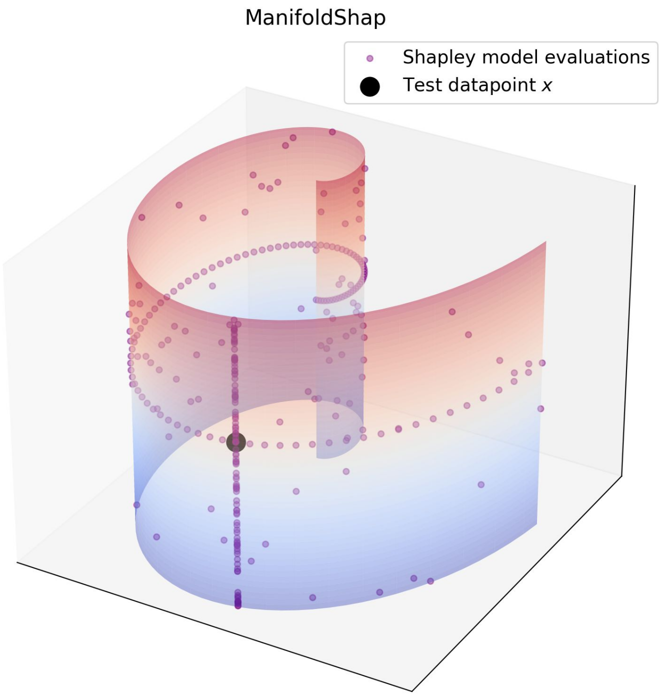
We call these regions of interest Z, the ‘manifold’.

Definition:
Let Z be an open open set with

$$P(X \in Z \mid do(X_S = x_S)) > 0$$

and $x \in Z$. Then, we define the ManifoldShap value function on Z as follows:

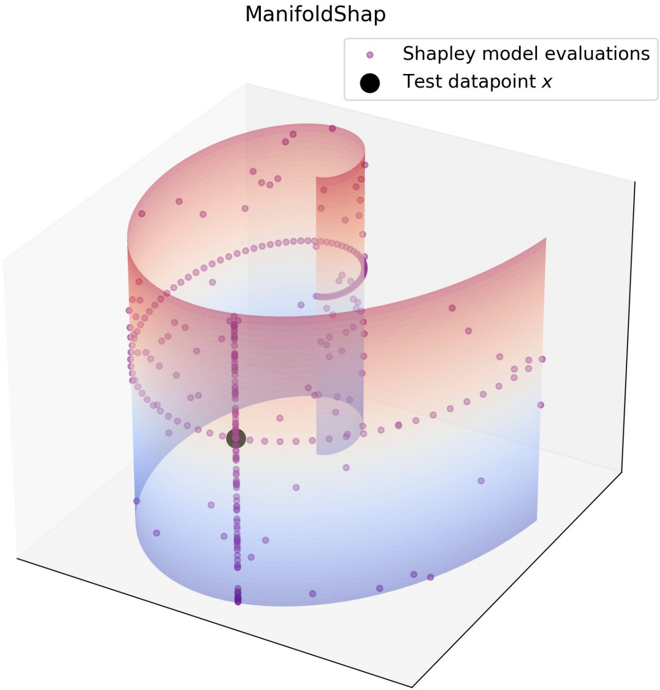
$$v^{\mathrm{MAN}}_{\boldsymbol{x},f,\mathcal{Z}}(S) := \mathbb{E}[f(\boldsymbol{X}) \mid do(\boldsymbol{X}_S = \boldsymbol{x}_S), \boldsymbol{X} \in \mathcal{Z}]$$



ManifoldShap

- In practice, Z can be chosen to be the data manifold, or any other region of interest, where model behaviour is relevant to explanations sought.
- One way of choosing Z explored in our work is based on density values

$$Z = \mathcal{D}_\epsilon := \{x \in \mathbb{R}^d : p(x) > \epsilon\}$$



Off-manifold robustness of ManifoldShap

Idea: Changing the function in regions of small mass should not result in drastic changes in the Shapley values

Off-manifold robustness of ManifoldShap

Idea: Changing the function in regions of small mass should not result in drastic changes in the Shapley values

Definition (Subspace T-robustness).
Let Z be such that $P(X \in Z) > 0$

Suppose two models $f_1(x), f_2(x)$ are such that $\sup_{x \in Z} |f_1(x) - f_2(x)| \leq \delta$

Then, we say that a value function, $v_{x,f}$, is strong T-robust on subspace Z, if it satisfies the following condition:

$$|v_{x,f_1}(S) - v_{x,f_2}(S)| \leq T\delta \quad \text{for any} \quad S \subseteq [d]$$

Off-manifold robustness of ManifoldShap

Idea: Changing the function in regions of small mass should not result in drastic changes in the Shapley values

Definition (Subspace T-robustness).
Let Z be such that $P(X \in Z) > 0$

Suppose two models $f_1(x), f_2(x)$ are such that $\sup_{x \in Z} |f_1(x) - f_2(x)| \leq \delta$

Then, we say that a value function, $v_{x,f}$, is strong T-robust on subspace Z, if it satisfies the following condition:

$$|v_{x,f_1}(S) - v_{x,f_2}(S)| \leq T\delta \quad \text{for any} \quad S \subseteq [d]$$

ManifoldShap satisfies subspace robustness, whereas all other value functions (both on and off-manifold value functions) do not.

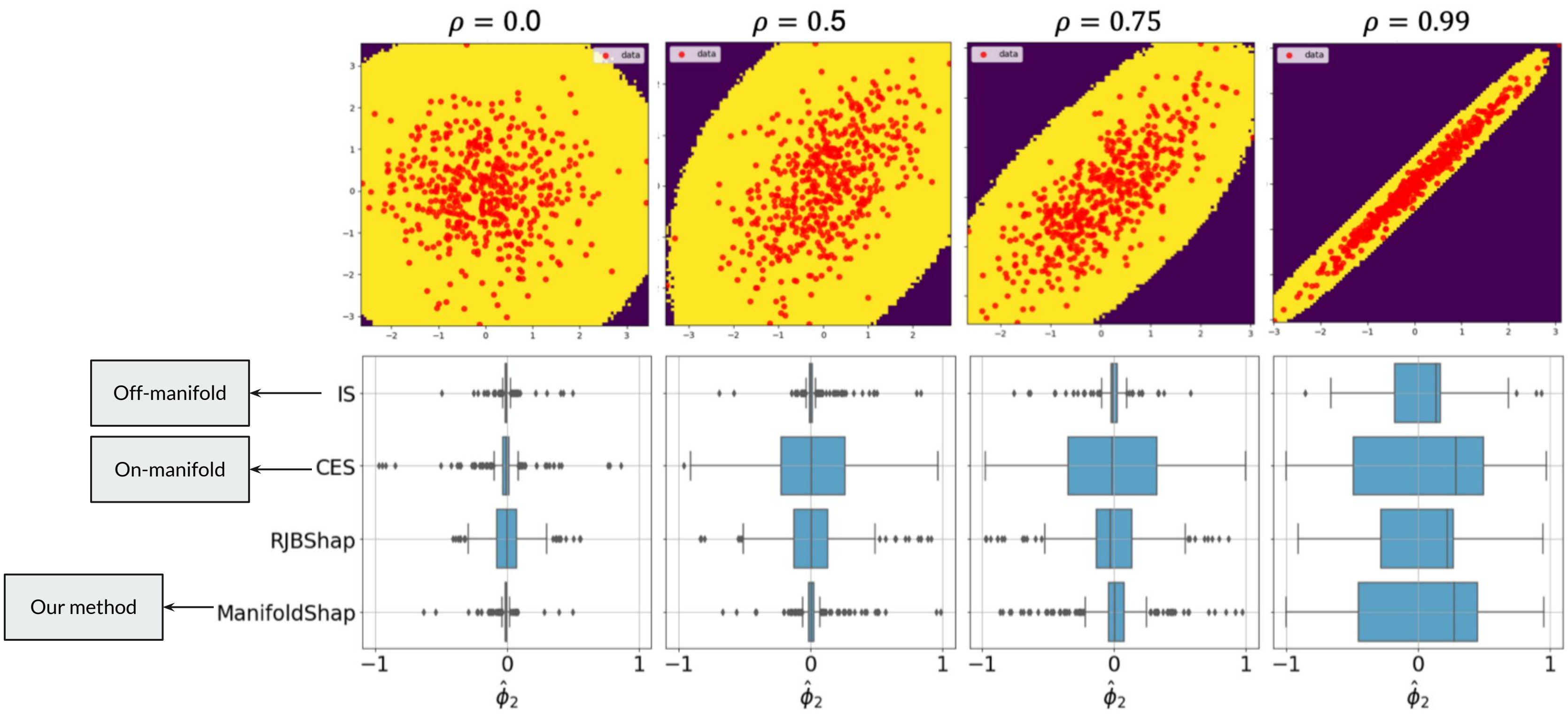
Off-manifold robustness of ManifoldShap

ManifoldShap remains highly causally accurate (by minimising the sensitivity to feature correlations), while satisfying the **Robustness property**.

Proposition 8. *The measure $p_{\mathcal{Z},\boldsymbol{x}_S}$ satisfies*

$$p_{\mathcal{Z},\boldsymbol{x}_S} \in \arg\min_{p_S}\{\mathrm{TV}(p_S, p^{\mathrm{do}}_{\boldsymbol{x}_S}) : v_{f,p_S} \text{ is strong T-robust on subspace } \mathcal{Z}\}$$

Sensitivity to correlation



Robustness, causal accuracy trade-off

$$Z = \mathcal{D}_\epsilon := \{x \in \mathbb{R}^d : p(x) > \epsilon\}$$

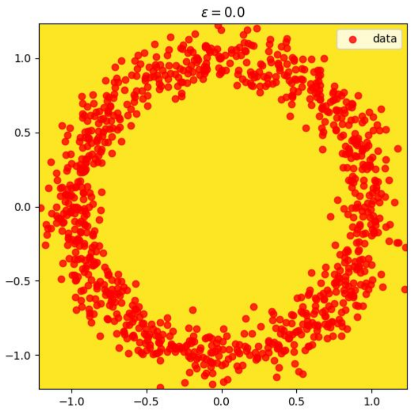
$0 \longleftarrow \epsilon \longrightarrow \infty$

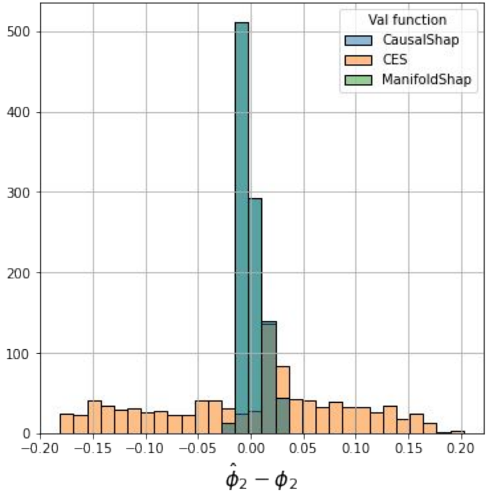
Robustness, causal accuracy trade-off

$$Z = \mathcal{D}_\epsilon := \{x \in \mathbb{R}^d : p(x) > \epsilon\}$$

$0 \longleftarrow \epsilon \longrightarrow \infty$

- ManifoldShap = InterventionalShap
- No robustness to off-manifold perturbations



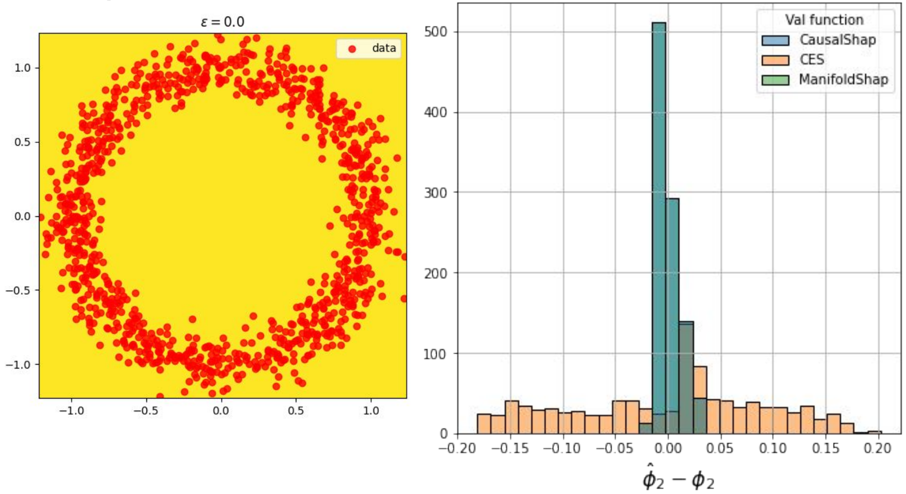


Robustness, causal accuracy trade-off

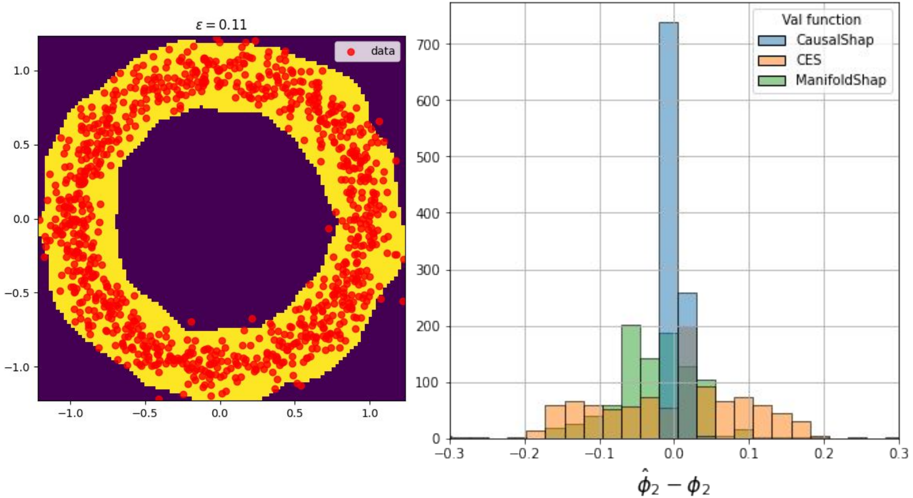
$$Z = \mathcal{D}_\epsilon := \{x \in \mathbb{R}^d : p(x) > \epsilon\}$$

$0 \longleftarrow \epsilon \longrightarrow \infty$

- ManifoldShap = InterventionalShap
- No robustness to off-manifold perturbations



- Increasing robustness to off-manifold perturbations
- Reducing Causal Accuracy, as fewer points are considered when computing Shapley values



Experimental Results

COMPAS Dataset Results

This dataset captures detailed information about the criminal history, jail and prison time, demographic attributes, and COMPAS risk scores for 6172 defendants from Broward County.

Ground Truth Function: only uses 'race' to make predictions.

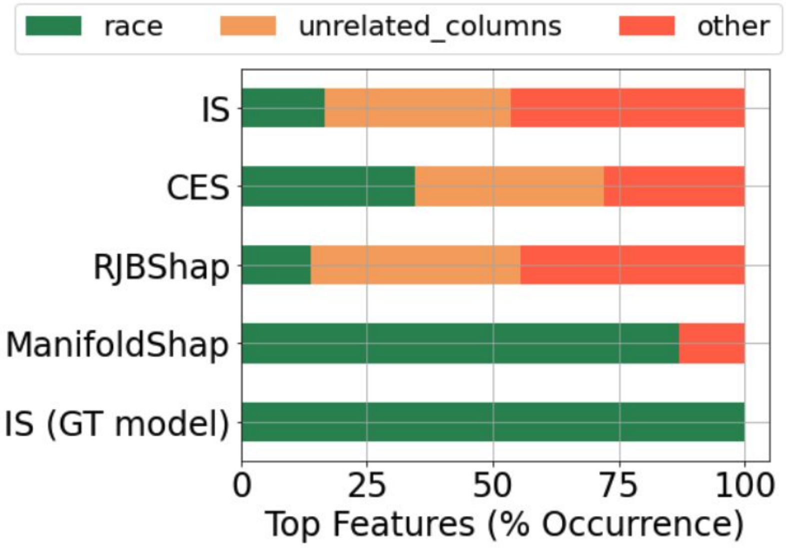
Perturbed model: Perturbs the model off-manifold to only use a synthetic positively correlated feature named 'unrelated_column' to make predictions.

COMPAS Dataset Results

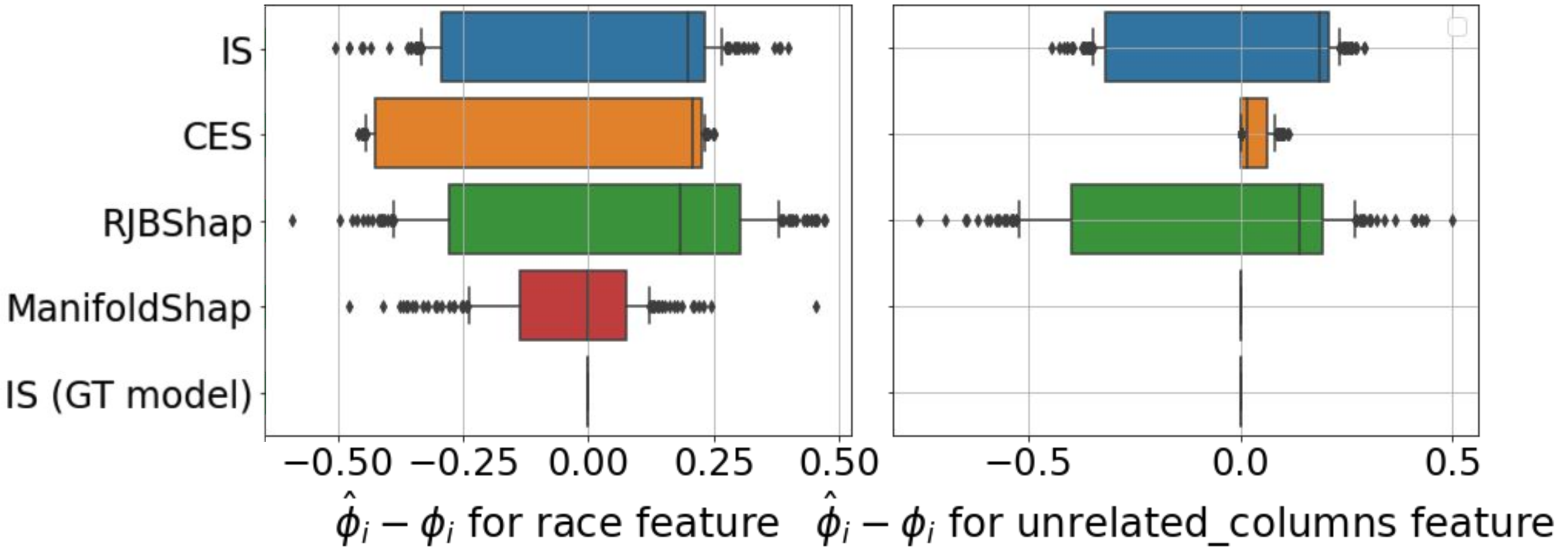
This dataset captures detailed information about the criminal history, jail and prison time, demographic attributes, and COMPAS risk scores for 6172 defendants from Broward County.

Ground Truth Function: only uses ‘race’ to make predictions.

Perturbed model: Perturbs the model off-manifold to only use a synthetic positively correlated feature named ‘unrelated_column’ to make predictions.



COMPAS Dataset Results

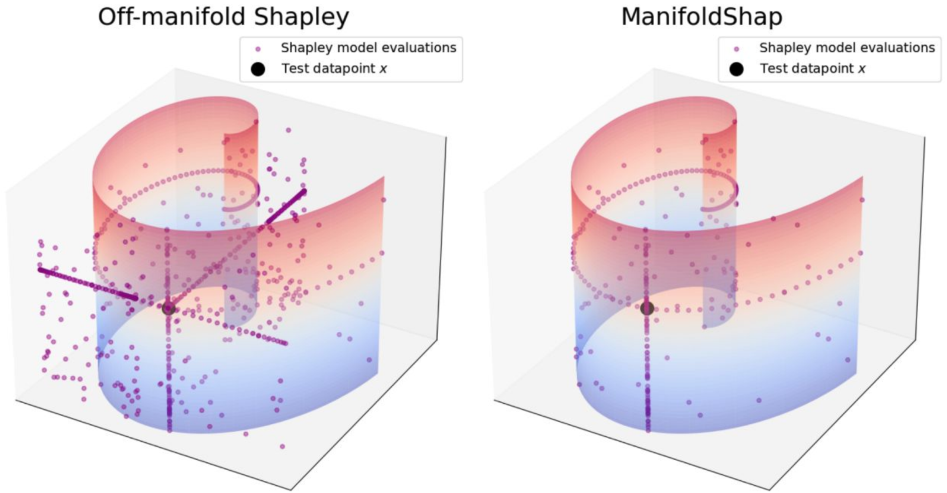


Conclusions

ManifoldShap properties

Robustness:

- **ManifoldShap** explanations are robust to model changes outside the data manifold.
- **ManifoldShap** is the only value function which satsifies this property.



Accuracy:

- **ManifoldShap** remains close to ground truth Interventional Shapley values.

Check out our paper!