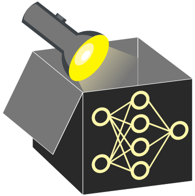
Interpretable Machine Learning



Permutation Feature Importance (PFI)

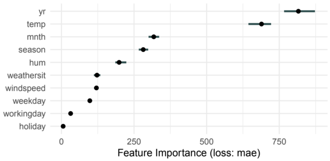
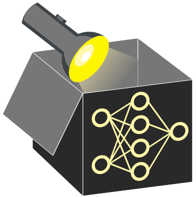


Figure: Bike Sharing Dataset

Learning goals

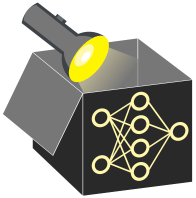
- Understand how PFI is computed
- Understanding strengths and weaknesses

MOTIVATION FOR PFI



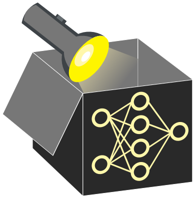
- **Goal:** Assess how important feature(s) X_S are for predictive performance of a **fixed trained model** $\hat{f}$ on a given dataset $\mathcal{D}$
- **Idea:** Estimate change in model performance when X_S is "made uninformative"

MOTIVATION FOR PFI



- **Goal:** Assess how important feature(s) X_S are for predictive performance of a **fixed trained model** $\hat{f}$ on a given dataset $\mathcal{D}$
- **Idea:** Estimate change in model performance when X_S is "made uninformative"
- **Question:** Can we make X_S uninformative by removing it from the model?
  $\rightsquigarrow$ No, $\hat{f}$ was trained with X_S and retraining without X_S gives a different model

MOTIVATION FOR PFI



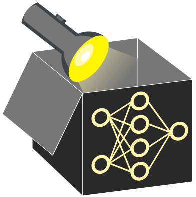
- **Goal:** Assess how important feature(s) X_S are for predictive performance of a **fixed trained model** $\hat{f}$ on a given dataset $\mathcal{D}$
- **Idea:** Estimate change in model performance when X_S is "made uninformative"
- **Question:** Can we make X_S uninformative by removing it from the model?
  $\rightsquigarrow$ No, $\hat{f}$ was trained with X_S and retraining without X_S gives a different model
- **Solution:** Simulate feature removal by replacing X_S with a perturbed version $\tilde{X}_S$ that is independent of (X_{-S}, Y) but preserves distribution $\mathbb{P}(X_S)$
  $\rightsquigarrow$ Compare baseline predictions $\hat{f}(X)$ with perturbed predictions $\hat{f}(\tilde{X}_S, X_{-S})$

$$\text{PFI}_S := \underbrace{\mathbb{E}\Big[L\big(\hat{f}(\tilde{X}_S, X_{-S}), Y\big)\Big]}_{\text{risk after "destroying" } X_S} - \underbrace{\mathbb{E}\Big[L\big(\hat{f}(X), Y\big)\Big]}_{\text{baseline risk}},$$

- **How to perturb X_S?**
  - Add random noise: distorts $\mathbb{P}(X_S)$ (not used)
  - Permutation: preserves marginal $\mathbb{P}(X_S)$, breaks dependence with Y (used)

PERMUTATION FEATURE IMPORTANCE (PFI)

▸ Breiman (2001)



Sample estimator (using independent test set $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$)

- Measure the error **with feat. values** x_S and **with permuted feat. values** $\tilde{x}_S$
- Repeat permutation (e.g., m times) and average difference of both errors:

$$\widehat{PFI}_S = \frac{1}{m} \sum_{k=1}^{m} \left[\mathcal{R}_{\text{emp}}(\hat{f}, \tilde{\mathcal{D}}_{(k)}^{S}) - \mathcal{R}_{\text{emp}}(\hat{f}, \mathcal{D})\right]$$

- $\mathcal{D}_S^{(k)}$: dataset where column(s) x_S are **permuted** once (in repetition k)
- $\mathcal{R}_{\text{emp}}(\hat{f}, \mathcal{D}) = \frac{1}{n} \sum_{(x,y) \in \mathcal{D}} L(\hat{f}(x), y)$: Measures performance of $\hat{f}$ using $\mathcal{D}$
- Average over m permutations to reduce Monte-Carlo variance

Example of permuting feature x_S with $S = \{1\}$ and $m = 6$ permutations:

$\mathcal{D}$

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	4	7
2	5	8
3	6	9

$\Rightarrow$

$\tilde{\mathcal{D}}_{(1)}^{S}$

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	4	7
2	5	8
3	6	9

$\tilde{\mathcal{D}}_{(2)}^{S}$

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
2	4	7
1	5	8
3	6	9

$\tilde{\mathcal{D}}_{(3)}^{S}$

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
2	4	7
3	5	8
1	6	9

$\tilde{\mathcal{D}}_{(4)}^{S}$

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	4	7
3	5	8
2	6	9

$\tilde{\mathcal{D}}_{(5)}^{S}$

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
3	4	7
1	5	8
2	6	9

$\tilde{\mathcal{D}}_{(6)}^{S}$

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
3	4	7
2	5	8
1	6	9

Note: S refers to a subset of features, here $|S| = 1$ to measure impact of permuting x_1 on performance

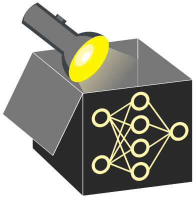
PERMUTATION FEATURE IMPORTANCE

$\tilde{\mathcal{D}}^{S}_{(k)}$

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	2	4	7
⋮	1	5	8
n	3	6	9

$\mathcal{D}$

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	4	7
2	5	8
3	6	9



1. **Perturbation:** Sample feature values from the distribution of x_S ($P(X_S)$).
   $\Rightarrow$ Randomly permute feature x_S
   $\Rightarrow$ Replace x_S with permuted feature $\tilde{x}_S$ and create data $\tilde{\mathcal{D}}^S$ containing $\tilde{x}_S$

PERMUTATION FEATURE IMPORTANCE

$\tilde{\mathcal{D}}^S_{(k)}$

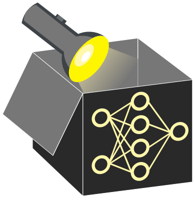
i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	2	4	7
$\vdots$	1	5	8
n	3	6	9

$\hat{f}$
0.6
0.6
0.6

$\mathcal{D}$

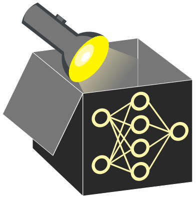
$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	4	7
2	5	8
3	6	9

$\hat{f}$
0.4
0.8
0.6



1. **Perturbation:** Sample feature values from the distribution of x_S ($P(X_S)$).
   $\Rightarrow$ Randomly permute feature x_S
   $\Rightarrow$ Replace x_S with permuted feature $\tilde{x}_S$ and create data $\tilde{\mathcal{D}}^S$ containing $\tilde{x}_S$
2. **Prediction:** Make predictions for both data, i.e., $\mathcal{D}$ and $\tilde{\mathcal{D}}^S$

PERMUTATION FEATURE IMPORTANCE



$\tilde{\mathcal{D}}^{S}_{(k)}$

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	2	4	7
⋮	1	5	8
n	3	6	9

$L(\hat{f}, y)$
0.9
0.5
0.1

$\mathcal{D}$

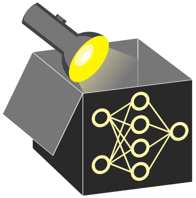
$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	4	7
2	5	8
3	6	9

$L(\hat{f}, y)$
0.25
0.35
0.1

3. Aggregation:

- Compute the loss for each observation in both data sets

PERMUTATION FEATURE IMPORTANCE



$\tilde{\mathcal{D}}^S_{(k)}$

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	2	4	7
$\vdots$	1	5	8
n	3	6	9

$\mathcal{D}$

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	4	7
2	5	8
3	6	9

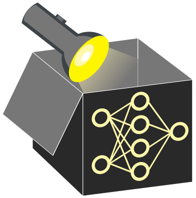
ΔL
0.65
0.15
0

$L(\hat{f}, y)$		$L(\hat{f}, y)$
0.9		0.25
0.5	-	0.35
0.1		0.1

3. Aggregation:

- Compute the loss for each observation in both data sets
- Take the difference of both losses ΔL for each observation

PERMUTATION FEATURE IMPORTANCE



$$\mathcal{R}_{\text{emp}}(\hat{f}, \tilde{\mathcal{D}}_{(k)}^{S}) - \mathcal{R}_{\text{emp}}(\hat{f}, \mathcal{D})$$

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	2	4	7
⋮	1	5	8
n	3	6	9

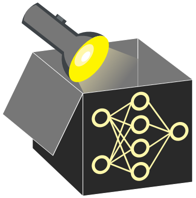
$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	4	7
2	5	8
3	6	9

ΔL
0.65
0.15
0

= 0.267

3. **Aggregation:**
   - Compute the loss for each observation in both data sets
   - Take the difference of both losses ΔL for each observation
   - Average this change in loss across all observations
     Note: Same as computing $\mathcal{R}_{\text{emp}}$ on both data sets and taking difference

PERMUTATION FEATURE IMPORTANCE



$$\mathcal{R}_{\text{emp}}(\hat{f}, \tilde{\mathcal{D}}^{S}_{(k)}) - \mathcal{R}_{\text{emp}}(\hat{f}, \mathcal{D})$$

1

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	2	4	7
⋮	1	5	8
n	3	6	9

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	4	7
2	5	8
3	6	9

ΔL
0.65
0.15
0

= 0.267

⋮

m

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	3	4	7
⋮	2	5	8
n	1	6	9

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	4	7
2	5	8
3	6	9

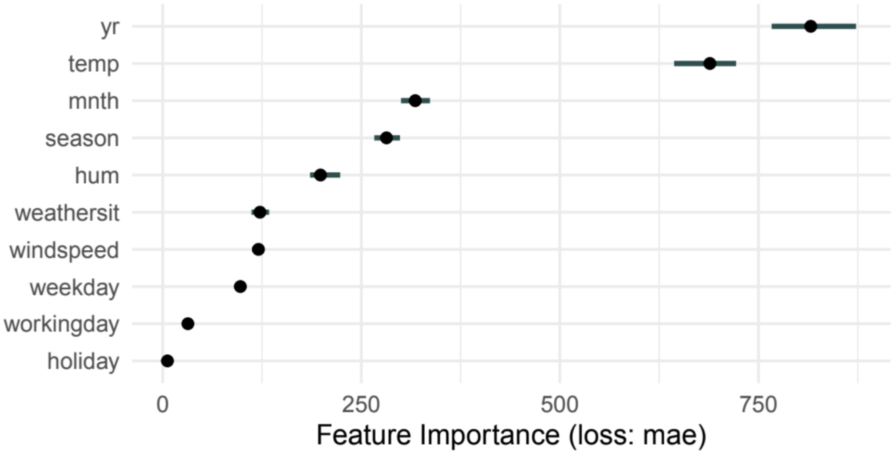
ΔL
0.85
0
0.35

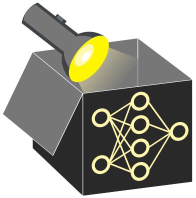
= 0.4

$\widehat{PFI}_S$ **= ½ (0.267 + 0.4)**

3. **Aggregation:**
   - Compute the loss for each observation in both data sets
   - Take the difference of both losses ΔL for each observation
   - Average this change in loss across all observations
   - Repeat perturbation and average over multiple repetitions

EXAMPLE: BIKE SHARING DATASET

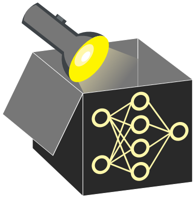




Interpretation:

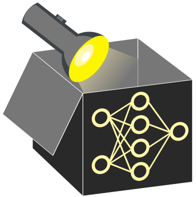
- `yr` and `temp` are most important features using mean absolute error (MAE)
- Destroying information about `yr` by permuting it increases MAE of model by 816
- Error bars show 5% and 95% quantiles over multiple permutations

COMMENTS ON PFI



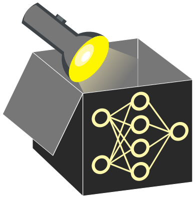
- Interpretation: Increase in error when feature's information is destroyed

COMMENTS ON PFI



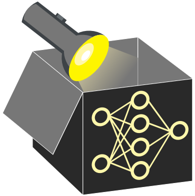
- Interpretation: Increase in error when feature's information is destroyed
- Results can be unreliable due to random permutations
  $\Rightarrow$ Solution: Average results over multiple repetitions

COMMENTS ON PFI



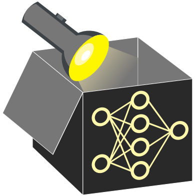
- Interpretation: Increase in error when feature's information is destroyed
- Results can be unreliable due to random permutations
  $\Rightarrow$ Solution: Average results over multiple repetitions
- Permuting features despite correlation/dependence with other features can lead to unrealistic combinations of feature values $\rightsquigarrow$ Extrapolation issue

COMMENTS ON PFI



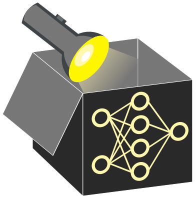
- Interpretation: Increase in error when feature's information is destroyed
- Results can be unreliable due to random permutations
  $\Rightarrow$ Solution: Average results over multiple repetitions
- Permuting features despite correlation/dependence with other features can lead to unrealistic combinations of feature values $\rightsquigarrow$ Extrapolation issue
- PFI automatically includes importance of interaction effects with other features
  $\Rightarrow$ Permuting x_j also destroys interactions with permuted feature
  $\Rightarrow$ PFI score contains importance of all interactions with permuted feature

COMMENTS ON PFI



- Interpretation: Increase in error when feature's information is destroyed
- Results can be unreliable due to random permutations
  $\Rightarrow$ Solution: Average results over multiple repetitions
- Permuting features despite correlation/dependence with other features can lead to unrealistic combinations of feature values $\rightsquigarrow$ Extrapolation issue
- PFI automatically includes importance of interaction effects with other features
  $\Rightarrow$ Permuting x_j also destroys interactions with permuted feature
  $\Rightarrow$ PFI score contains importance of all interactions with permuted feature
- Interpretation of PFI depends on whether training or test data is used

COMMENTS ON PFI - EXTRAPOLATION



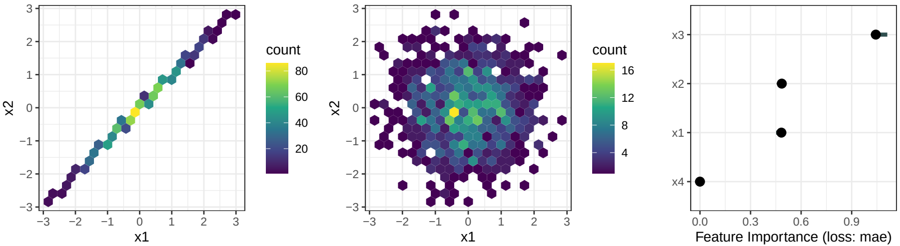
Example: Let $y = x_3 + \epsilon_y$, with $\epsilon_y \sim \mathcal{N}(0, 0.1)$.

- $x_1 := \epsilon_1$, $x_2 := x_1 + \epsilon_2$ are highly correlated ($\epsilon_1 \sim \mathcal{N}(0, 1)$, $\epsilon_2 \sim \mathcal{N}(0, 0.01)$)
- $x_3 := \epsilon_3$, $x_4 := \epsilon_4$, with $\epsilon_3, \epsilon_4 \sim \mathcal{N}(0, 1)$ and all noise terms ϵ_j are independent
- Fitting a linear model yields $\hat{f}(\mathbf{x}) \approx 0.3x_1 - 0.3x_2 + x_3$

COMMENTS ON PFI - EXTRAPOLATION

Example: Let $y = x_3 + \epsilon_y$, with $\epsilon_y \sim \mathcal{N}(0, 0.1)$.

- $x_1 := \epsilon_1$, $x_2 := x_1 + \epsilon_2$ are highly correlated ($\epsilon_1 \sim \mathcal{N}(0, 1)$, $\epsilon_2 \sim \mathcal{N}(0, 0.01)$)
- $x_3 := \epsilon_3$, $x_4 := \epsilon_4$, with $\epsilon_3, \epsilon_4 \sim \mathcal{N}(0, 1)$ and all noise terms ϵ_j are independent
- Fitting a linear model yields $\hat{f}(\mathbf{x}) \approx 0.3x_1 - 0.3x_2 + x_3$

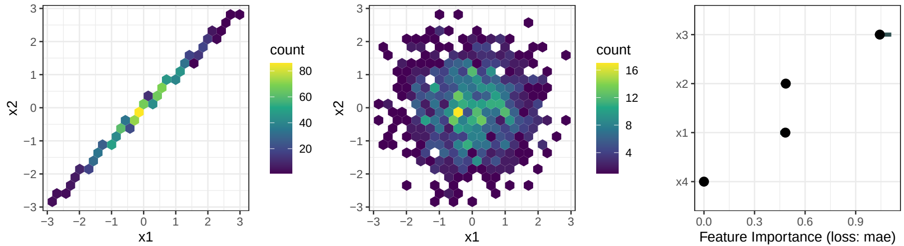


Hexbin plot of (x_1, x_2) before (left) and after (center) permuting x_1; PFI scores (right).

COMMENTS ON PFI - EXTRAPOLATION

Example: Let $y = x_3 + \epsilon_y$, with $\epsilon_y \sim \mathcal{N}(0, 0.1)$.

- $x_1 := \epsilon_1$, $x_2 := x_1 + \epsilon_2$ are highly correlated ($\epsilon_1 \sim \mathcal{N}(0, 1)$, $\epsilon_2 \sim \mathcal{N}(0, 0.01)$)
- $x_3 := \epsilon_3$, $x_4 := \epsilon_4$, with $\epsilon_3, \epsilon_4 \sim \mathcal{N}(0, 1)$ and all noise terms ϵ_j are independent
- Fitting a linear model yields $\hat{f}(\mathbf{x}) \approx 0.3x_1 - 0.3x_2 + x_3$



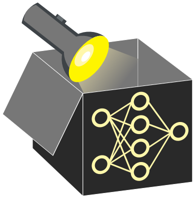
Hexbin plot of (x_1, x_2) before (left) and after (center) permuting x_1; PFI scores (right).

$\Rightarrow$ x_1, x_2 cancel in $\hat{f}$ since $x_1 \approx x_2$, hence $0.3x_1 - 0.3x_2 \approx 0 \rightsquigarrow$ should be irrelevant

$\Rightarrow$ Permuting x_1 breaks joint structure $\rightsquigarrow$ unrealistic inputs

$\Rightarrow$ *PFI* > 0 due to extrapolation (PFI evaluates model on unrealistic inputs)
$\rightsquigarrow$ x_1, x_2 are misleadingly considered relevant

COMMENTS ON PFI - INTERACTIONS

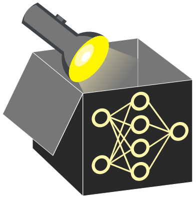


Example: Let $x_1, \ldots, x_4$ be independently and uniformly sampled from $\{-1, 1\}$ and

$$y := x_1 x_2 + x_3 + \epsilon_Y \text{ with } \epsilon_Y \sim N(0, 1)$$

Fitting a LM yields $\hat{f}(x) \approx x_1 x_2 + x_3$.

COMMENTS ON PFI - INTERACTIONS



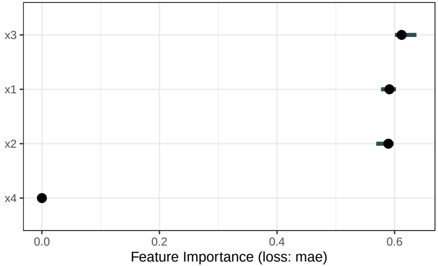
Example: Let $x_1, \ldots, x_4$ be independently and uniformly sampled from $\{-1, 1\}$ and

$$y := x_1 x_2 + x_3 + \epsilon_Y \text{ with } \epsilon_Y \sim N(0, 1)$$

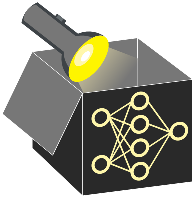
Fitting a LM yields $\hat{f}(x) \approx x_1 x_2 + x_3$.

Although x_3 alone contributes as much to the prediction as x_1 and x_2 jointly, all three are considered equally relevant.

$\Rightarrow$ PFI does not fairly attribute the performance to the individual features.

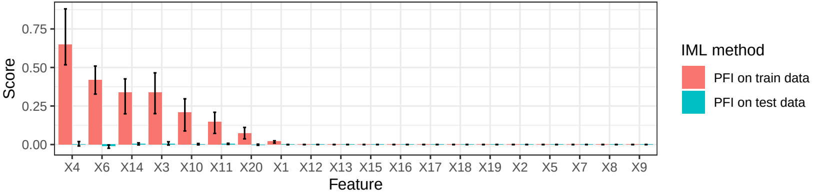


COMMENTS ON PFI - TRAIN VS. TEST DATA



Example:

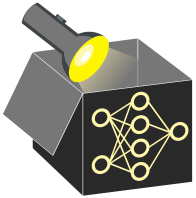
- $x_1, \ldots, x_{20}, y$ are independently sampled from $\mathcal{U}(-10, 10)$
- Train set: $n = 50$ (intentionally small) and large test set
- Model: `xgboost` with default settings (overfits strongly)



Observation:

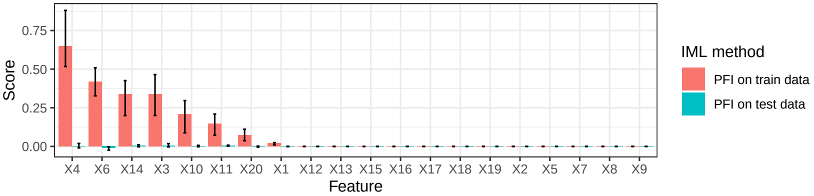
- PFI on train data highlights features that the model overfitted to.
- PFI on test data detects no relevant features.

COMMENTS ON PFI - TRAIN VS. TEST DATA



Example:

- $x_1, \ldots, x_{20}, y$ are independently sampled from $\mathcal{U}(-10, 10)$
- Train set: $n = 50$ (intentionally small) and large test set
- Model: `xgboost` with default settings (overfits strongly)

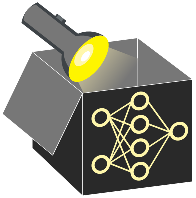


Observation:

- PFI on train data highlights features that the model overfitted to.
- PFI on test data detects no relevant features.

Why? *PFI* $\neq 0$ if permuting a feature breaks a dependency the model relies on.
Model overfits due to spurious feature-target dependencies in train that vanish on test.
$\Rightarrow$ To identify features that help the model to generalize, compute PFI on test data.

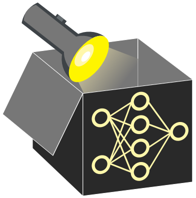
IMPLICATIONS OF PFI



Can we get insight into whether the ...

1. feature x_j is causal for the prediction?
   - $PFI_j \neq 0 \Rightarrow$ model relies on x_j
   - As the train vs. test data example shows, the converse does not hold

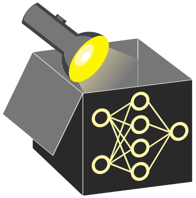
IMPLICATIONS OF PFI



Can we get insight into whether the ...

1. feature x_j is causal for the prediction?
   - $PFI_j \neq 0 \Rightarrow$ model relies on x_j
   - As the train vs. test data example shows, the converse does not hold
2. feature x_j contains prediction-relevant information?
   - $PFI_j \neq 0 \Rightarrow x_j$ is dependent on y, x_{-j}, or both (due to extrapolation)
   - x_j is not exploited by model (regardless of whether it is useful for y or not) $\Rightarrow PFI_j = 0$

IMPLICATIONS OF PFI



Can we get insight into whether the ...

1. feature x_j is causal for the prediction?
   - $PFI_j \neq 0 \Rightarrow$ model relies on x_j
   - As the train vs. test data example shows, the converse does not hold
2. feature x_j contains prediction-relevant information?
   - $PFI_j \neq 0 \Rightarrow x_j$ is dependent on y, x_{-j}, or both (due to extrapolation)
   - x_j is not exploited by model (regardless of whether it is useful for y or not) $\Rightarrow PFI_j = 0$
3. model requires access to x_j to achieve it's prediction performance?
   - As the extrapolation example demonstrates, such insight is not possible