

# Interpretable Machine Learning

## Conditional Feature Importance (CFI)

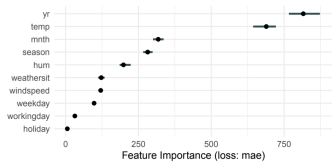


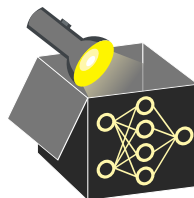
Figure: Bike Sharing Dataset

### Learning goals

- Extrapolation and Conditional Sampling
- Conditional Feature Importance (CFI)
- Interpretation of CFI and difference to PFI

# CONDITIONAL FEATURE IMPORTANCE IDEA

- **Permutation Feature Importance Idea:** Replace the feat. of interest  $x_j$  with an indep. sample from the marginal dist.  $\mathbb{P}(x_j)$ , e.g. by randomly perm. obs. in  $x_j$



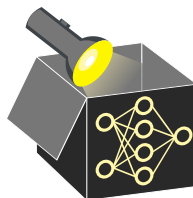
# CONDITIONAL FEATURE IMPORTANCE IDEA

- **Permutation Feature Importance Idea:** Replace the feat. of interest  $x_j$  with an indep. sample from the marginal dist.  $\mathbb{P}(x_j)$ , e.g. by randomly perm. obs. in  $x_j$
- **Problem:** Under dependent features, permutation leads to extrapolation



# CONDITIONAL FEATURE IMPORTANCE IDEA

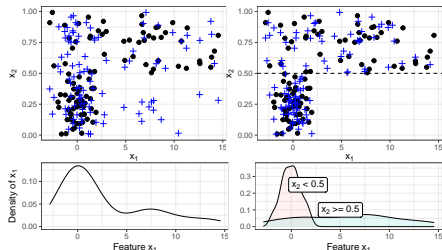
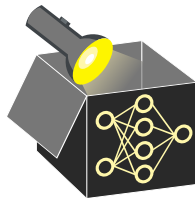
- **Permutation Feature Importance Idea:** Replace the feat. of interest  $x_j$  with an indep. sample from the marginal dist.  $\mathbb{P}(x_j)$ , e.g. by randomly perm. obs. in  $x_j$
- **Problem:** Under dependent features, permutation leads to extrapolation
- **Conditional Feature Importance Idea:** Resample  $x_j$  from the cond. dist.  $\mathbb{P}(x_j|x_{-j})$ , s.t. the joint dist. is preserved, i.e.,  $\mathbb{P}(x_j|x_{-j})\mathbb{P}(x_{-j}) = \mathbb{P}(x_j, x_{-j})$



# CONDITIONAL FEATURE IMPORTANCE IDEA

- **Permutation Feature Importance Idea:** Replace the feat. of interest  $x_j$  with an indep. sample from the marginal dist.  $\mathbb{P}(x_j)$ , e.g. by randomly perm. obs. in  $x_j$
- **Problem:** Under dependent features, permutation leads to extrapolation
- **Conditional Feature Importance Idea:** Resample  $x_j$  from the cond. dist.  $\mathbb{P}(x_j|x_{-j})$ , s.t. the joint dist. is preserved, i.e.,  $\mathbb{P}(x_j|x_{-j})\mathbb{P}(x_{-j}) = \mathbb{P}(x_j, x_{-j})$

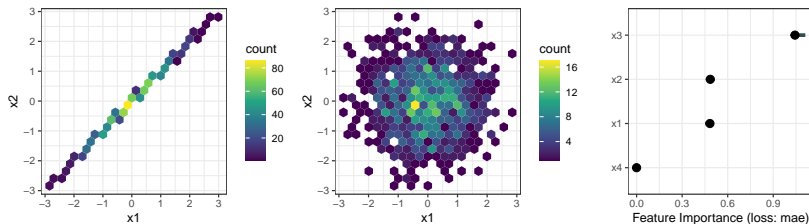
**Example:** Conditional permutation scheme ▶ Molnar et. al (2020)



- $X_2 \sim U(0, 1)$  and  $X_1 \sim N(0, 1)$  if  $X_2 < 0.5$ , else  $X_1 \sim N(4, 4)$  (black dots)
- **Left:** For  $X_2 < 0.5$ , permuting  $X_1$  (crosses) preserves marginal (but not joint) distribution  
~> Bottom: Marginal density of  $X_1$
- **Right:** Permuting  $X_1$  within subgroups  $X_2 < 0.5$  &  $X_2 \geq 0.5$  reduces extrapolation  
~> Bottom: Density of  $X_1$  conditional on groups

# RECALL: EXTRAPOLATION IN PFI

**Example:** Let  $y = x_3 + \epsilon_y$  with  $\epsilon_y \sim N(0, 0.1)$  where  $x_1 := \epsilon_1$ ,  $x_2 := x_1 + \epsilon_2$  are highly correlated ( $\epsilon_1 \sim N(0, 1)$ ,  $\epsilon_2 \sim N(0, 0.01)$ ) and  $x_3 := \epsilon_3$ ,  $x_4 := \epsilon_4$ , with  $\epsilon_3, \epsilon_4 \sim N(0, 1)$ . All noise terms are independent. Fitting a LM yields  $\hat{f}(\mathbf{x}) \approx 0.3x_1 - 0.3x_2 + x_3$ .



Hexbin plot of  $x_1, x_2$  before permuting  $x_1$  (left), after permuting  $x_1$  (center), and PFI scores (right)  $\Rightarrow x_1$  and  $x_2$  should be irrelevant for the prediction  $\hat{f}(\mathbf{x})$  for

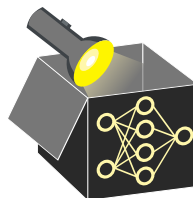
$\{\mathbf{x} : \mathbb{P}(\mathbf{x}) > 0\}$  as  $0.3x_1 - 0.3x_2 \approx 0$

$\Rightarrow$  PFI evaluates model on unrealistic obs. outside  $\mathbb{P}(\mathbf{x}) \rightsquigarrow x_1$  and  $x_2$  are considered relevant

# CONDITIONAL FEATURE IMPORTANCE

► Strobl et al. (2008)

► Hooker et al. (2021)



Conditional feature importance (CFI) for features  $x_S$  using test data  $\mathcal{D}$ :

- Measure the error **with unperturbed features**.
- Measure the error **with perturbed feature values**  $\tilde{x}^{S|-S}$ , where  $\tilde{x}_S^{S|-S} \sim \mathbb{P}(x_S|x_{-S})$
- Repeat permuting the feature (e.g.,  $m$  times) and average the difference of both errors:

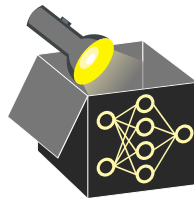
$$\widehat{CFI}_S = \frac{1}{m} \sum_{k=1}^m \mathcal{R}_{\text{emp}}(\hat{f}, \tilde{\mathcal{D}}_{(k)}^{S|-S}) - \mathcal{R}_{\text{emp}}(\hat{f}, \mathcal{D})$$

Here,  $\tilde{\mathcal{D}}^{S|-S}$  denotes the dataset where features  $x_S$  were sampled conditional on the remaining features  $x_{-S}$ .

# IMPLICATIONS OF CFI

► König et al. (2020)

**Interpretation:** Due to the conditional sampling w.r.t. all other features, CFI quantifies a feature's unique contribution to the model performance.





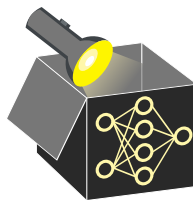
# IMPLICATIONS OF CFI

► König et al. (2020)

**Interpretation:** Due to the conditional sampling w.r.t. all other features, CFI quantifies a feature's unique contribution to the model performance.

## Entanglement with data:

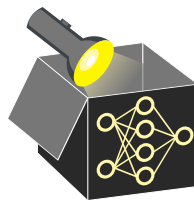
- If feature  $x_S$  does not contribute unique information about  $y$ , i.e.,  $x_S \perp\!\!\!\perp y | x_{-S} \Rightarrow \text{CFI} = 0$
- Why? Under the conditional independence  $\mathbb{P}(\tilde{x}^{S|-S}, y) = \mathbb{P}(x, y)$   
 $\rightsquigarrow$  no prediction-relevant information is destroyed by permutation of  $x_S$  conditional on  $x_{-S}$



# IMPLICATIONS OF CFI

► König et al. (2020)

**Interpretation:** Due to the conditional sampling w.r.t. all other features, CFI quantifies a feature's unique contribution to the model performance.



## Entanglement with data:

- If feature  $x_S$  does not contribute unique information about  $y$ , i.e.,  $x_S \perp\!\!\!\perp y | x_{-S} \Rightarrow \text{CFI} = 0$
- Why? Under the conditional independence  $\mathbb{P}(\tilde{x}^{S|-S}, y) = \mathbb{P}(x, y)$   
 $\rightsquigarrow$  no prediction-relevant information is destroyed by permutation of  $x_S$  conditional on  $x_{-S}$

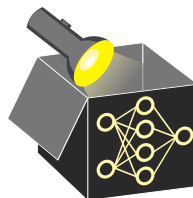
## Entanglement with model:

- If the model does not use a feature  $\Rightarrow \text{CFI} = 0$
- Why? Then the prediction is not affected by any perturbation of the feature  
 $\rightsquigarrow$  model performance does not change after conditional permutation

# IMPLICATIONS OF CFI

Can we gain insight into whether ...

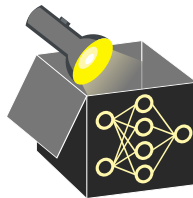
- 1 the feature  $x_j$  is causal for the prediction?
  - $CFI_j \neq 0 \Rightarrow$  model relies on  $x_j$  (converse does not hold, see next slide)



# IMPLICATIONS OF CFI

Can we gain insight into whether ...

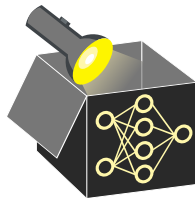
- ❶ the feature  $x_j$  is causal for the prediction?
  - $CFI_j \neq 0 \Rightarrow$  model relies on  $x_j$  (converse does not hold, see next slide)
- ❷ the variable  $x_j$  contains prediction-relevant information?
  - If  $x_j \not\perp\!\!\!\perp y$  but  $x_j \perp\!\!\!\perp y|x_{-j}$  (e.g.,  $x_j$  and  $x_{-j}$  share information)  $\Rightarrow CFI_j = 0$
  - $x_j$  is not exploited by model (regardless of whether it is useful for  $y$  or not)  
 $\Rightarrow CFI_j = 0$



# IMPLICATIONS OF CFI

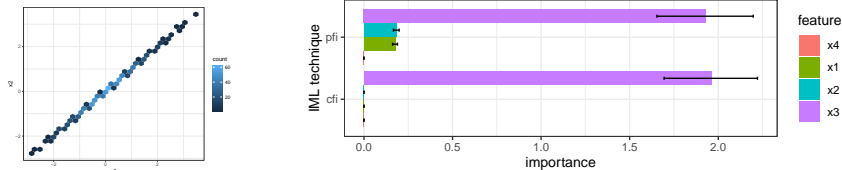
Can we gain insight into whether ...

- ❶ the feature  $x_j$  is causal for the prediction?
  - $CFI_j \neq 0 \Rightarrow$  model relies on  $x_j$  (converse does not hold, see next slide)
- ❷ the variable  $x_j$  contains prediction-relevant information?
  - If  $x_j \not\perp\!\!\!\perp y$  but  $x_j \perp\!\!\!\perp y|x_{-j}$  (e.g.,  $x_j$  and  $x_{-j}$  share information)  $\Rightarrow CFI_j = 0$
  - $x_j$  is not exploited by model (regardless of whether it is useful for  $y$  or not)  $\Rightarrow CFI_j = 0$
- ❸ Does the model require access to  $x_j$  to achieve its prediction performance?
  - $CFI_j \neq 0 \Rightarrow x_j$  contributes unique information (meaning  $x_j \not\perp\!\!\!\perp y|x_{-j}$ )
  - Only uncovers the relationships that were exploited by the model



# COMPARISON: PFI AND CFI

**Example:** Let  $y = x_3 + \epsilon_y$  with  $\epsilon_y \sim N(0, 0.1)$  where  $x_1 := \epsilon_1$ ,  $x_2 := x_1 + \epsilon_2$  are highly correlated ( $\epsilon_1 \sim N(0, 1)$ ,  $\epsilon_2 \sim N(0, 0.01)$ ) and  $x_3 := \epsilon_3$ ,  $x_4 := \epsilon_4$ , with  $\epsilon_3, \epsilon_4 \sim N(0, 1)$ . All noise terms are independent. Fitting a LM yields  $\hat{f}(\mathbf{x}) \approx 0.3x_1 - 0.3x_2 + x_3$ .



**Figure:** Density plot for  $x_1, x_2$  before permuting  $x_1$  (left). PFI and CFI (right).

$\Rightarrow x_1$  and  $x_2$  are irrelevant for the prediction  $\hat{f}(\mathbf{x})$  for  $\{\mathbf{x} : \mathbb{P}(\mathbf{x}) > 0\}$  as

$$0.3x_1 - 0.3x_2 \approx 0$$

$\Rightarrow$  PFI evaluates model on unrealistic obs. outside  $\mathbb{P}(\mathbf{x}) \rightsquigarrow x_1, x_2$  are considered relevant (PFI  $> 0$ )

$\Rightarrow$  Since  $x_1$  can be reconstructed from  $x_2$  and vice versa, CFI considers  $x_1$  and  $x_2$  to be irrelevant