

# Interpretable Machine Learning

## Introduction to loss-based feature importance

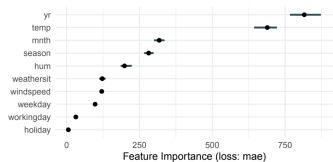


Figure: Bike Sharing Dataset

### Learning goals

- Understand motivation for feature importance
- Develop an intuition for possible use-cases
- Know characteristics of feature importance methods

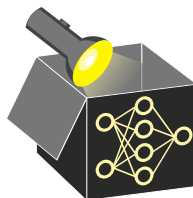
# MOTIVATION

- **Feature effects** describe the relationship of features  $x$  with the prediction  $\hat{y}$ 
  - requires one plot per feature
  - does not take the true target  $y$  into account



# MOTIVATION

- **Feature effects** describe the relationship of features  $x$  with the prediction  $\hat{y}$ 
  - requires one plot per feature
  - does not take the true target  $y$  into account
- **Feature importance** methods quantify the relevance of features w.r.t. prediction performance
  - condensed to one number per feature
  - provides insight into the relationship with  $y$



# MOTIVATION

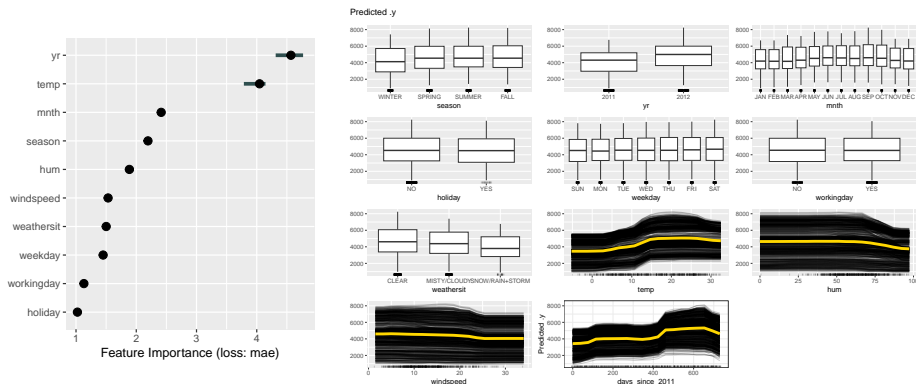
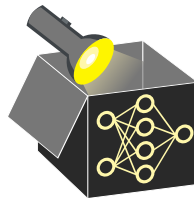
- **Feature effects** describe the relationship of features  $x$  with the prediction  $\hat{y}$ 
  - requires one plot per feature
  - does not take the true target  $y$  into account
- **Feature importance** methods quantify the relevance of features w.r.t. prediction performance
  - condensed to one number per feature
  - provides insight into the relationship with  $y$
- **N.B.:** Here, we use the term feature importance to describe loss-based feature importance methods. In the literature, you may find other notions of “feature importance” (e.g., variance-based methods derived from feature effect methods, see also [Greenwell et al. \(2020\)](#))



# EXAMPLE

Feature importance offers condensed summary of feat. relevance w.r.t. performance

- Fit random forest on bike sharing data
- Left: Feature importance ranking by permutation feature importance (PFI)
- Right: Feature effects for all features



# FEATURE IMPORTANCE SCHEME

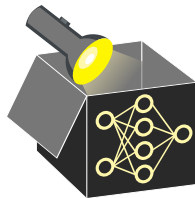
Loss-based feature importance methods are often based on two concepts

## ❶ **Perturbation/Removal:**

Generate predictions for which the feature of interest has been perturbed or removed

## ❷ **Performance Comparison:**

Compare performance under perturbation/removal with the original model performance



Depending on the type of perturbation/removal, feature importance methods provide insight into different aspects of model and data.

# POTENTIAL INTERPRETATION GOALS

Feature importance methods provide condensed insights, but can only highlight certain aspects of model and data. There are different interpretation goals one might be interested in whose question of interest do not necessarily coincide (except for special cases).



For example, one may be interested in getting insight into whether the ...

- (1) feature  $x_j$  is causal for the prediction?
- (2) feature  $x_j$  contains prediction-relevant information about  $y$ ?
- (3) model requires access to  $x_j$  to achieve its prediction performance?

# POTENTIAL INTERPRETATION GOALS

Feature importance methods provide condensed insights, but can only highlight certain aspects of model and data. There are different interpretation goals one might be interested in whose question of interest do not necessarily coincide (except for special cases).



For example, one may be interested in getting insight into whether the ...

(1) feature  $x_j$  is causal for the prediction?

- Changing feature value  $x_j$  has an effect on prediction  $\hat{y} = \hat{f}(x)$
- In LM: non-zero coefficient, in ML: present feature effect
- **Note:** If  $x_j$  is causal for prediction  $\hat{y} \not\Rightarrow$  causal for the ground truth  $y$ , e.g.:
  - A disease symptom may be used in a model to predict disease status  
 $\rightsquigarrow$  causal for prediction  $\hat{y}$
  - But intervening on disease symptom does not have an effect on the disease  
 $\rightsquigarrow$  not causal for the ground truth  $y$

(2) feature  $x_j$  contains prediction-relevant information about  $y$ ?

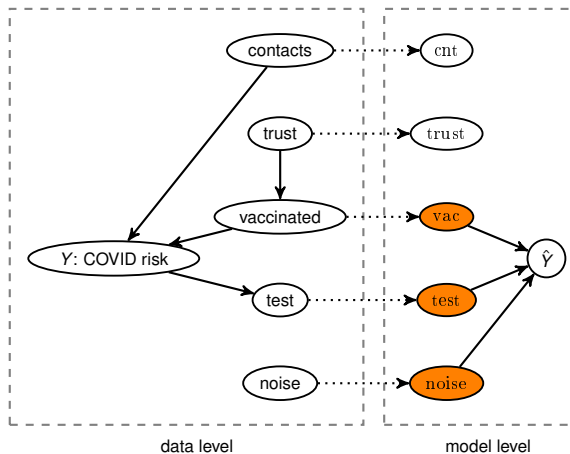
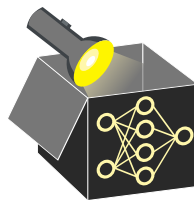
(3) model requires access to  $x_j$  to achieve its prediction performance?



# EXAMPLE: CAUSAL FOR THE PREDICTION (1)

A feature may be causal for the prediction  $\hat{y}$  (1) without containing prediction-relevant information about  $y$  (2)

*Examples:* overfitting due noisy features

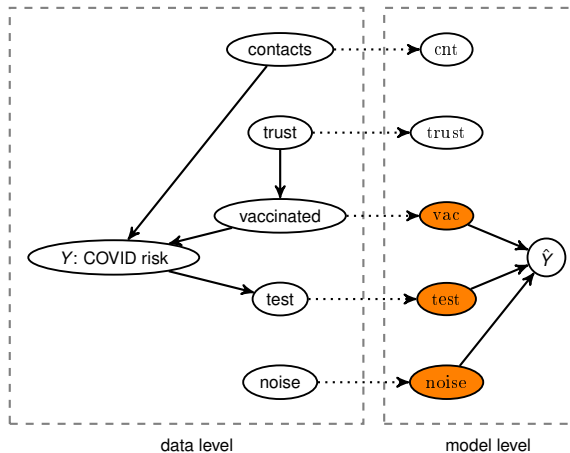


# EXAMPLE: CAUSAL FOR THE PREDICTION (1)



A feature may be causal for the prediction  $\hat{y}$  (1) without containing prediction-relevant information about  $y$  (2)

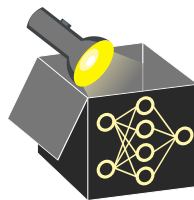
*Examples:* overfitting due noisy features



- All features used by the model are of interest
  - Here: Model uses feature noise, although it does not contain prediction-relevant information about  $y$  (data level)
- ⇒ Overfitted models may use many noise features which are deemed relevant on model level (but not on data level)

# POTENTIAL INTERPRETATION GOALS

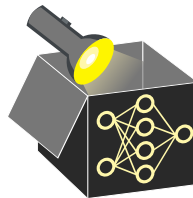
Feature importance methods provide condensed insights, but can only highlight certain aspects of model and data. There are different interpretation goals one might be interested in whose question of interest do not necessarily coincide (except for special cases).



For example, one may be interested in getting insight into whether the ...

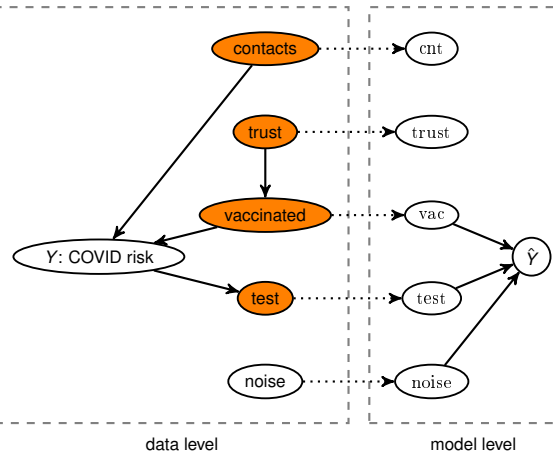
- (1) feature  $x_j$  is causal for the prediction?
- (2) feature  $x_j$  contains prediction-relevant information about  $y$ ?
  - Feature  $x_j$  helps to predict the target  $y$  (e.g., conditional expectation) w.r.t. performance
  - If  $x_j \perp\!\!\!\perp y$  (independent) then  $x_j$  and  $y$  have zero mutual information (since  $\mathbb{E}[y|x_j] = \mathbb{E}[y]$ )  
 $\leadsto x_j$  has no prediction-relevant information
- (3) model requires access to  $x_j$  to achieve its prediction performance?

# EXAMPLE: CONTAINS PREDICTION-RELEVANT INFORMATION (2)

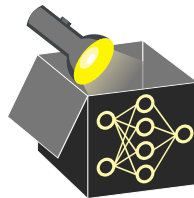


A feature may contain  
prediction-relevant information (2)  
without causing the prediction (1)

*Examples:* underfitting, model  
multiplicity

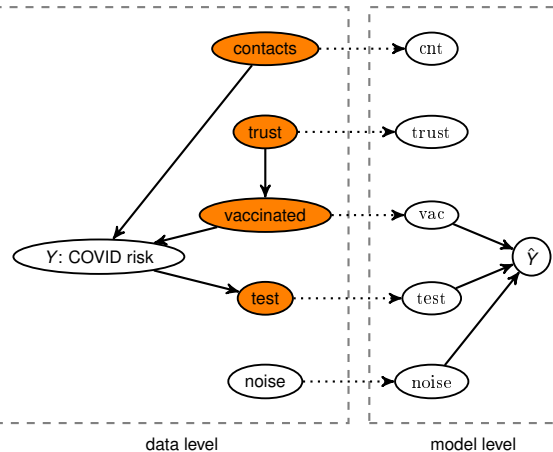


## EXAMPLE: CONTAINS PREDICTION-RELEVANT INFORMATION (2)



A feature may contain prediction-relevant information (2) without causing the prediction (1)

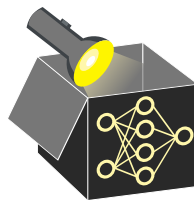
*Examples:* underfitting, model multiplicity



- All prediction-relevant features for  $y$  are of interest
  - Example: All features that are directly or indirectly (i.e., via another feature) connected to  $y$
- ⇒ Underfitted models may ignore prediction-relevant features such as **contacts** here

# POTENTIAL INTERPRETATION GOALS

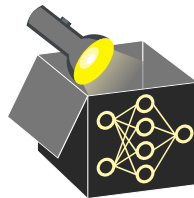
Feature importance methods provide condensed insights, but can only highlight certain aspects of model and data. There are different interpretation goals one might be interested in whose question of interest do not necessarily coincide (except for special cases).



For example, one may be interested in getting insight into whether the ...

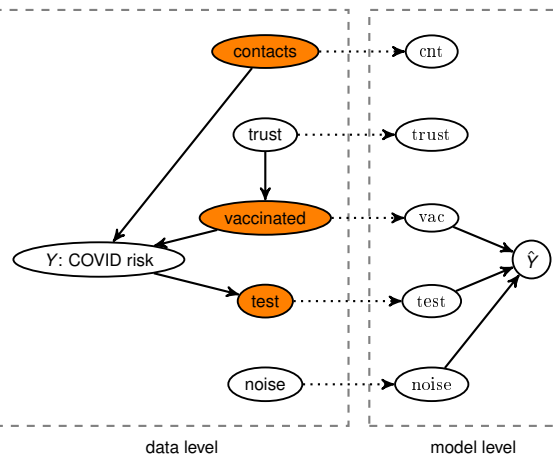
- (1) feature  $x_j$  is causal for the prediction?
- (2) feature  $x_j$  contains prediction-relevant information about  $y$ ?
- (3) model requires access to  $x_j$  to achieve its prediction performance?
  - Feature  $x_j$  helps to predict the target  $y$  w.r.t. performance, compared to using only  $x_{-j}$
  - If  $x_j \perp\!\!\!\perp y | x_{-j}$  (independent) then  $\mathbb{E}[y | x_{-j}] = \mathbb{E}[y | x_j, x_{-j}]$   
 $\leadsto x_j$  does not contribute unique prediction-relevant information about  $y$
  - **Note:** A model may rely on features that can be replaced with others, e.g., a random forest fitted on data with  $\mathbb{E}[y | x_1] \neq \mathbb{E}[y]$  and  $\mathbb{E}[y | x_1] = \mathbb{E}[y | x_1, x_2]$  where  $x_1$  was not used as split variable may rely on  $x_2$

## EXAMPLE: UNIQUE PREDICTION RELEVANT INFORMATION (3)

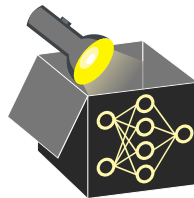


A feature may contain prediction-relevant information (2), without the model requiring access to the feature for (optimal) prediction performance (3)

*Examples: correlated features, confounding*

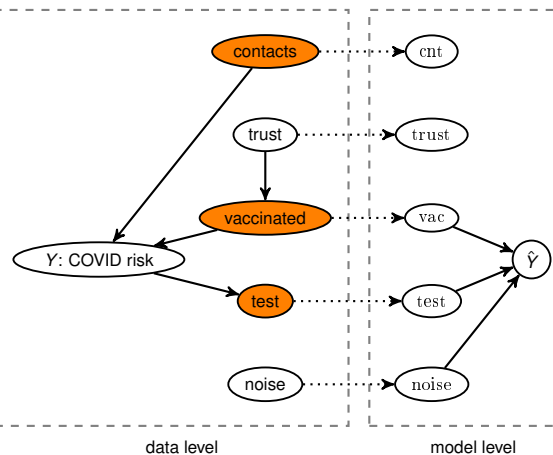


## EXAMPLE: UNIQUE PREDICTION RELEVANT INFORMATION (3)



A feature may contain prediction-relevant information (2), without the model requiring access to the feature for (optimal) prediction performance (3)

*Examples: correlated features, confounding*



- All unique prediction-relevant features for  $y$  are of interest
  - Example: All features that are directly connected to  $y$
- ⇒ **trust** and **vaccinated** may be correlated but only **vaccinated** is directly connected to  $y$