- SESSION 9 -
MACHINE LEARNING INTERPRETABILITY

HERTIE DATA SCIENCE SOCIETY

OUR MAIN TOPICS TODAY

Why the need for interpretation?

What features are important?

How does feature affect prediction?

How to understand individual prediction?

WHY THE NEED FOR INTERPRETATION?

WHY IS IT IMPORTANT TO US?

WHAT INSIGHTS ARE POSSIBLE?

FEATURE IMPORTANCE

What features in the data did the model think are most important?

EFFECTS ON PREDICTION

For any single prediction from a model, how did each feature in the data affect that particular prediction?

BIG PICTURE

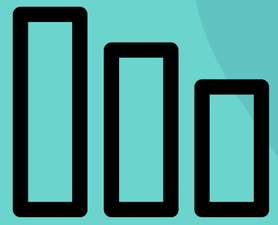
How does each feature affect the model's predictions in a big-picture sense (what is its typical effect when considered over a large number of possible predictions)?

BUILDING TRUST

Understanding a model helps to build trust in its recommendation and insights as well as its fairness, privacy, robustness, causality

DIRECTING FUTURE DATA COLLECTION

Model-based insights give a good understanding what new data are useful to collect



INFORMING FEATURE ENGINEERING

When working with new domain knowledge, you will need more than your intuition to

INFORMING DECISION-MAKING

Many important decisions are made by humans, for which insights can be more valuable than predictions.

GLOBAL INTERPRETATION

Helps to deliver an understanding of how the model makes predictions, what features are influential, how these features affect target in the model **as a whole**. Methods used:

- Feature Importance (Permutation Importance)
- Feature Effect (Partial Plots)

LOCAL INTERPRETATION

Helps to deliver an understanding of how features are influencing the target for a **specific** observation (or small group of observations). Methods used:

- LIME: Local interpretable model-agnostic explanations
- SHAP Values

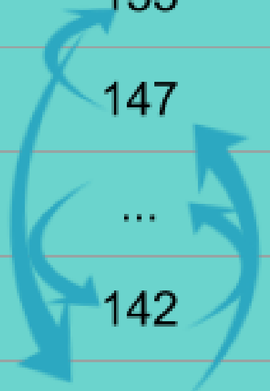
Feature Importance

WHAT FEATURES HAVE THE BIGGEST IMPACT ON PREDICTIONS?

PERMUTATION IMPORTANCE

Sample: Predict height at age 20 based on information at age 10

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24



STEP 1

Train a model

STEP 2

- Shuffle the values in a single column, make predictions using the new dataset.
- Use these predictions and the true target values to calculate how much the model error increases due to shuffling.
- That performance deterioration measures the importance of the variable you just shuffled.

STEP 3

- Return the data to the original order
- Repeat step 2 with the next column in the dataset, until you have calculated the importance of each column.

PARTIAL DEPDENCE PLOTS

HOW A FEATURE AFFECTS PREDICTIONS

STEP 1

Train a model. Use the trained model to predict outcomes;

STEP 2

Repeatedly alter the value for one variable to make a series of predictions;

STEP 3

Trace out predicted outcomes (on the vertical axis) as we move from small values to large values

LIME

HOW INDIVIDUAL PREDICTIONS ARE FORMED

LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

LIME PROVIDE METHODS TO EXPLAIN WHY AN INDIVIDUAL PREDICTION WAS MADE FOR A GIVEN OBSERVATION.

STEP 1

Create replicated feature data based on training data with slight value modifications;

STEP 2

Compute how close the original observation that you want to investigate to the observations in the replicated data

STEP 3

Apply machine learning model to predict outcomes of replicated data

STEP 4

Select a number of features that best describe predicted outcomes

STEP 5

Fit simple model to the replicated data with the best features which can help to explain how the outcome is formed

STEP 6

Explain the predicted outcome of the origin observation that you want to investigate using the model in step 5

SHAP VALUE

THE IMPACT OF SPECIFIC VALUE OF SPECIFIC FEATURE ON THE OUTCOME

STEP 1

Choose one specific feature X

STEP 2

Test the accuracy of models using the combinations of all other features, except X

STEP 3

Add in X to each combination to test how the model accuracy improve or deteriorate

PRESENTING RESULTS