1 Interpretability Methods

Complex machine learning algorithms (e.g. NNs) are hard/impossible to interpret. Interpretability methods help with debugging, trust and taking appropriate action on the results.

1.1 Local interpretability methods

Explain, why your model made this/these exact decisions.

1.1.1 Shapley values

Find attributes that determine the deviation of your output from the *expected value*. **How:** Calculate how much each feature pushes the prediction away from the expected value, by shuffling through all combinations of features having the sample-value or expected value respectively.

SHapley Additive exPlanations (SHAP) SHAP calculates Shapley values

1.1.2 Local interpretable model-agnostic explanations (LIME)

! → LIME interpretaions are not always consistent.

1.1.3 Example-based explanations

Influence Functions What would happen to the model parameters, if you would up-weight an instance? (model is function of training data.)