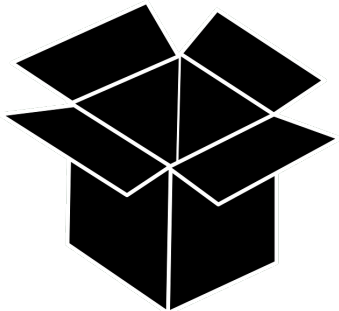


Interpretable Machine Learning

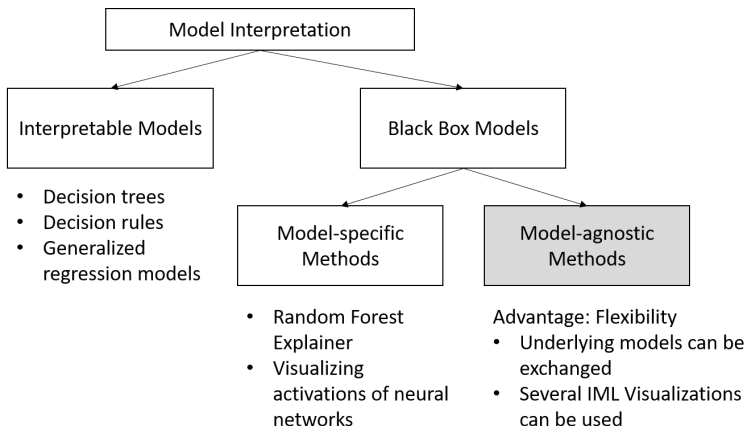
Fundamental Terms and Concepts



Learning goals

- Fundamental Terms and Concepts

WHAT TOOLS DO WE HAVE?



⇒ We will focus on model-agnostic interpretability!

INTRINSIC AND MODEL-AGNOSTIC INTERPRETATION

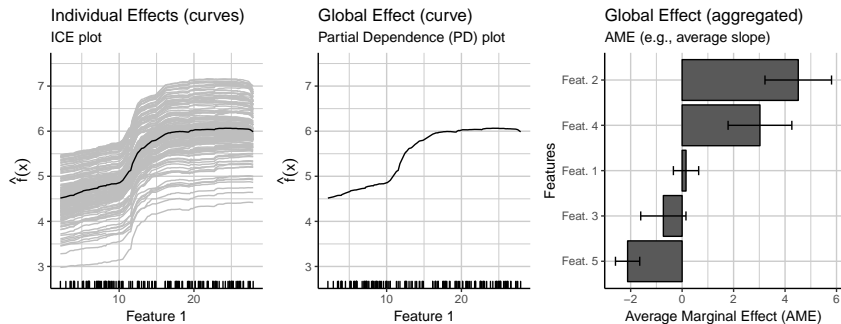
- Intrinsically interpretable models:
 - Examples are linear models and decision trees.
 - They are interpretable because of their simple structures, e.g. weighted combination of feature values or tree structure.
 - They are difficult to interpret with many features or complex interaction terms.
- Model-agnostic interpretation methods:
 - They are applied after training (post-hoc).
 - They also work for more complex black box models.
 - They can also be applied to intrinsically interpretable models, e.g. feature importance for decision trees.

MODEL-AGNOSTIC INTERPRETABILITY

- Model-agnostic interpretability methods work for **any** kind of machine learning model.
- Explanation type is not tied to the underlying model type.
- Often, only access to data and fitted predictor is required. No further knowledge about the model itself is necessary.
- We usually distinguish between **feature effect** and **feature importance** methods.

FEATURE EFFECTS VS. FEATURE IMPORTANCE

Feature Effects visualize or quantify the (average) relationship or contribution of a feature to the model prediction.

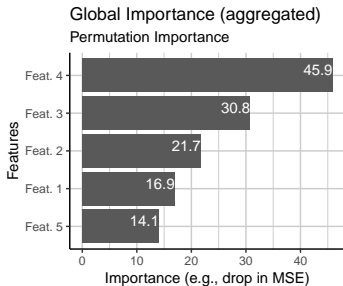


- Methods: Partial Dependence Plots, Individual Conditional Expectation, Accumulated Local Effects (ALE)
- Pendant in linear models: Regression coefficient $\hat{\theta}_j$

FEATURE EFFECTS VS. FEATURE IMPORTANCE

Feature importance methods rank features by how much they contribute to the predictive performance or prediction variance of the model.

- Methods: Permutation Feature Importance, Functional Anova
- Analog in linear models: Absolute t-statistic $\left| \frac{\hat{\theta}_j}{SE(\hat{\theta}_j)} \right|$



GLOBAL AND LOCAL INTERPRETABILITY

Global interpretability methods explain the expected model behavior for the entire input space by considering all available observations (or representative subsets). For example:

- Permutation Feature Importance
- Partial Dependence Plot
- Functional Anova
- ...

Local interpretability methods explain single predictions or a group of similar observations. For example:

- Individual Conditional Expectation (ICE) Plots
- Local Interpretable Model-Agnostic Explanations (LIME)
- Shapley Values
- ...

FIXED MODEL VS. REFITS

- Most methods presented in this lecture analyze a fixed, trained model (e.g., permutation feature importance).
- Some methods require refitting the model (e.g., PIMP).
- Trained model \Rightarrow Model is the object of analysis.
- Refitting \Rightarrow Learning process is the object of analysis.
- The advantage of refitting is that it includes information about the variability in the learning process.