8️⃣

Interpretable Models

Status	Done
Files	
Link	
Type	All Around
Course	Data Analysis

Interpretable models

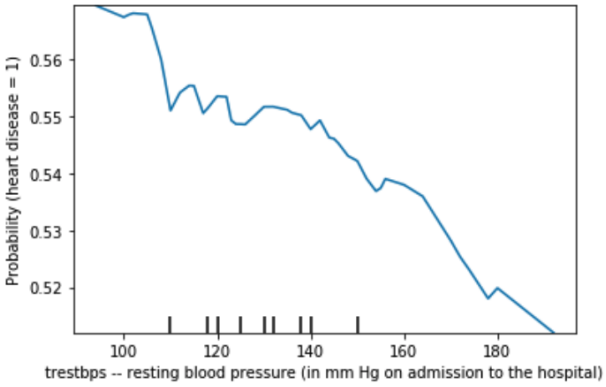
- The notebook has examples on how to interpret the fitted models (feature importance) for
  - Linear regression
  - Logistic regression
  - Trees

Model agnostic methods

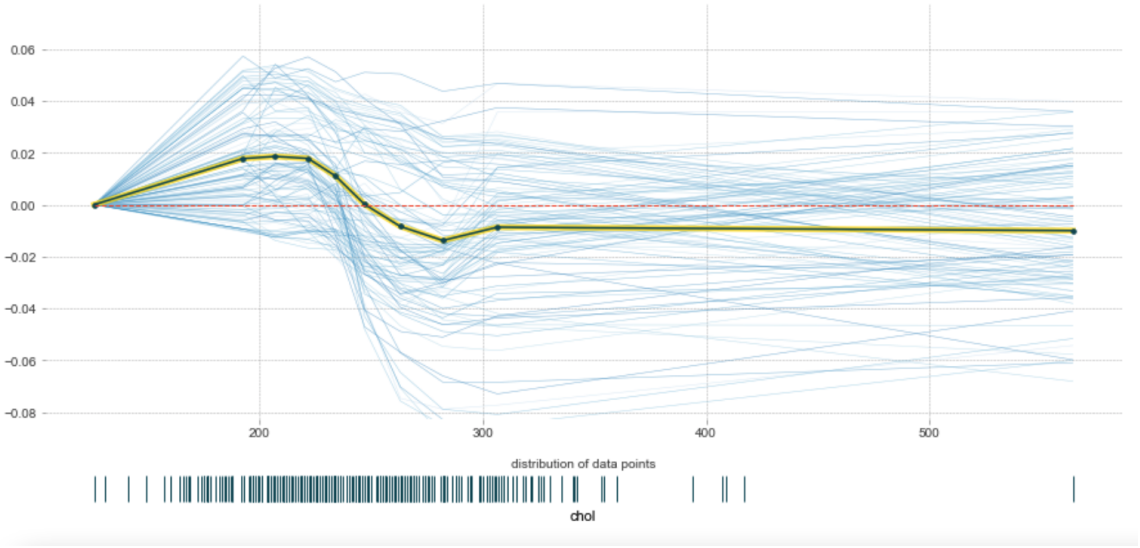
Can be applied to any model

- Permutation Feature Importance
  - Exactly what it sounds like! Permutate the feature value, and see the difference in prediction error, if there is more error in prediction, that means the algorithm is relying heavily on the feature, and therefore it has a high importance
- Partial Dependence Plots (PDP)

- Limited to one or two features per visualisations



- Individual Conditional Expectation (ICE)
  - Similar as PDP but instead of averaging, the plot is for every instance



> ! With both PDP and ICE its important to show the distribution of the instance feature values as if they are skewed it might introduce biases

- Local Interpretable Model-agnostic Explanations (LIME)

1. Sample points around x_i (the point that I want to interpret)

- I use my big/complex model to predict labels for each sample
- I weigh samples according to the distance from x_i
- Learn a new simple model (e.g. decision tree or linear regression) on weighted samples
- Use simple model to explain the decision

LIME caveats

- If you explain the same record twice, the explanations can be different!
- You can only explain one instance, so it's not possible to interpret the whole black box model.

- Shapley Additive Explanations (SHAP)

  (+)

  - We can use it for local and global explanations.
  - Has a strong theoretical foundation.
  - Recent research has showed that they can work with dependent features.

  (-)

  - Computation speed can be very slowwwww (especially with many features)