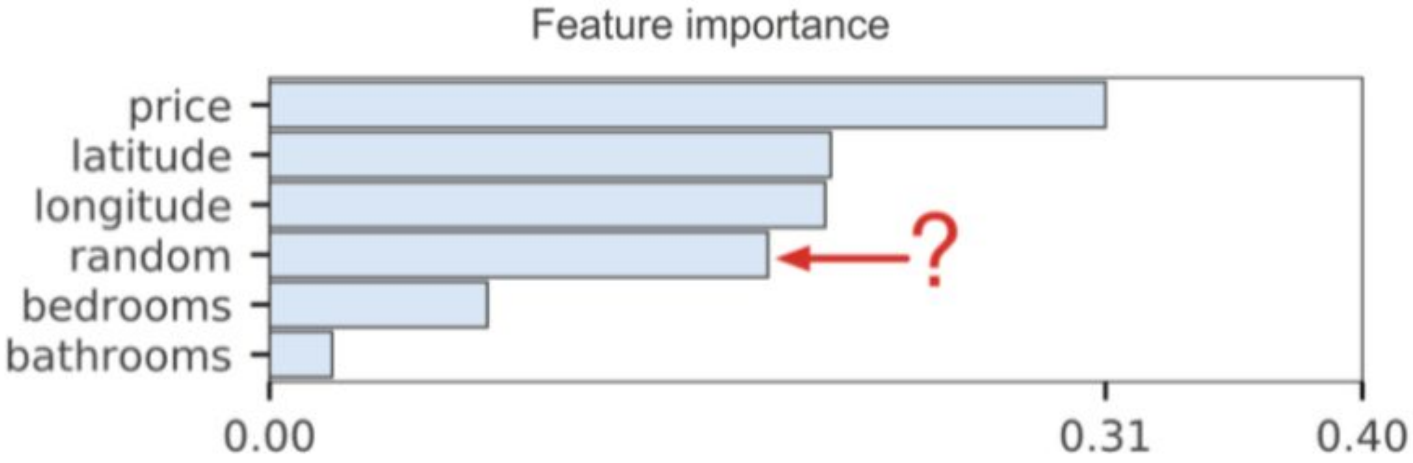
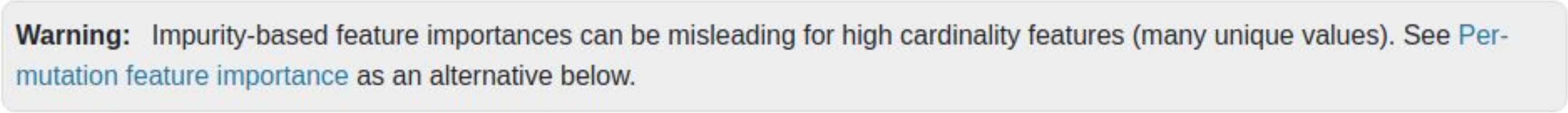
Beware default RF Feature Importances



Based on: https://explained.ai/rf-importance/index.html
Terence Parr, Kerem Turgutlu, Christopher Csiszar,
Jeremy Howard

Olaf Placha
Piotr Nawrot

When the problem occurs?



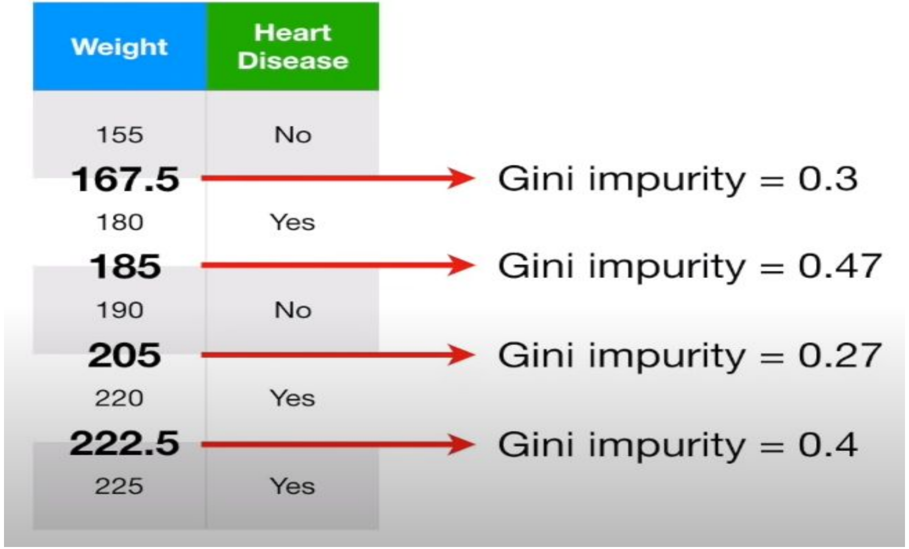
Warning: Impurity-based feature importances can be misleading for high cardinality features (many unique values). See Permutation feature importance as an alternative below.

[1]

“We found that for the original random forest method the variable importance measures are affected by the ***number of categories*** and ***scale of measurement*** of the predictor variables, which are ***no direct indicators of the true importance of the variable***” [2]

[1] https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html
[2] Bias in random forest variable importance measures: Illustrations, sources and a solution
Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis and Torsten Hothorn

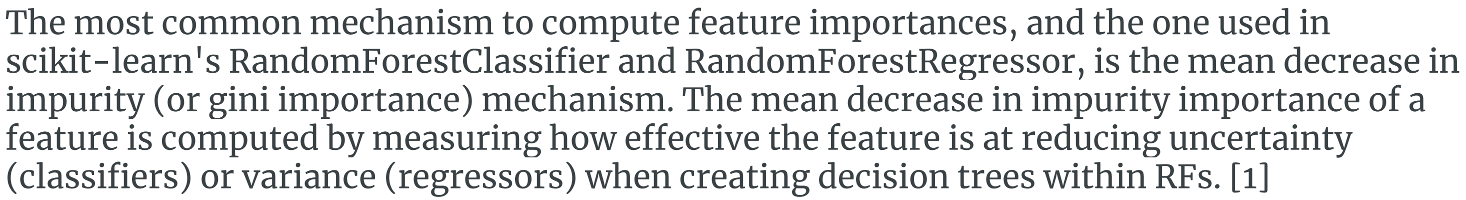
How are we deciding on a split in tree predictors?



Weight	Heart Disease	
155	No	
167.5		→ Gini impurity = 0.3
180	Yes	
185		→ Gini impurity = 0.47
190	No	
205		→ Gini impurity = 0.27
220	Yes	
222.5		→ Gini impurity = 0.4
225	Yes	

Source: https://www.youtube.com/watch?v=7VeUPuFGJHk

How are importances in sklearn’s tree computed?



The most common mechanism to compute feature importances, and the one used in scikit-learn's RandomForestClassifier and RandomForestRegressor, is the mean decrease in impurity (or gini importance) mechanism. The mean decrease in impurity importance of a feature is computed by measuring how effective the feature is at reducing uncertainty (classifiers) or variance (regressors) when creating decision trees within RFs. [1]

[1] https://explained.ai/rf-importance/index.html

Why the problem occurs?

“Testing more split points means there's a higher probability of finding a split that, purely by chance, happens to predict the dependent variable well.” [1]

[2]

[1] https://explained.ai/rf-importance/index.html
[2] https://www.startupdonut.co.uk/sites/default/files/

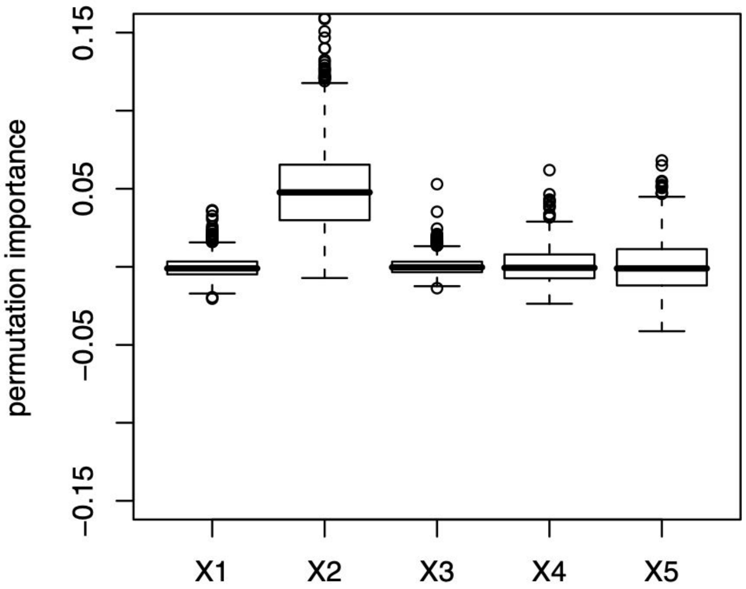
Are there better options?

Feature importances methods, ordered by the amount of computations needed:

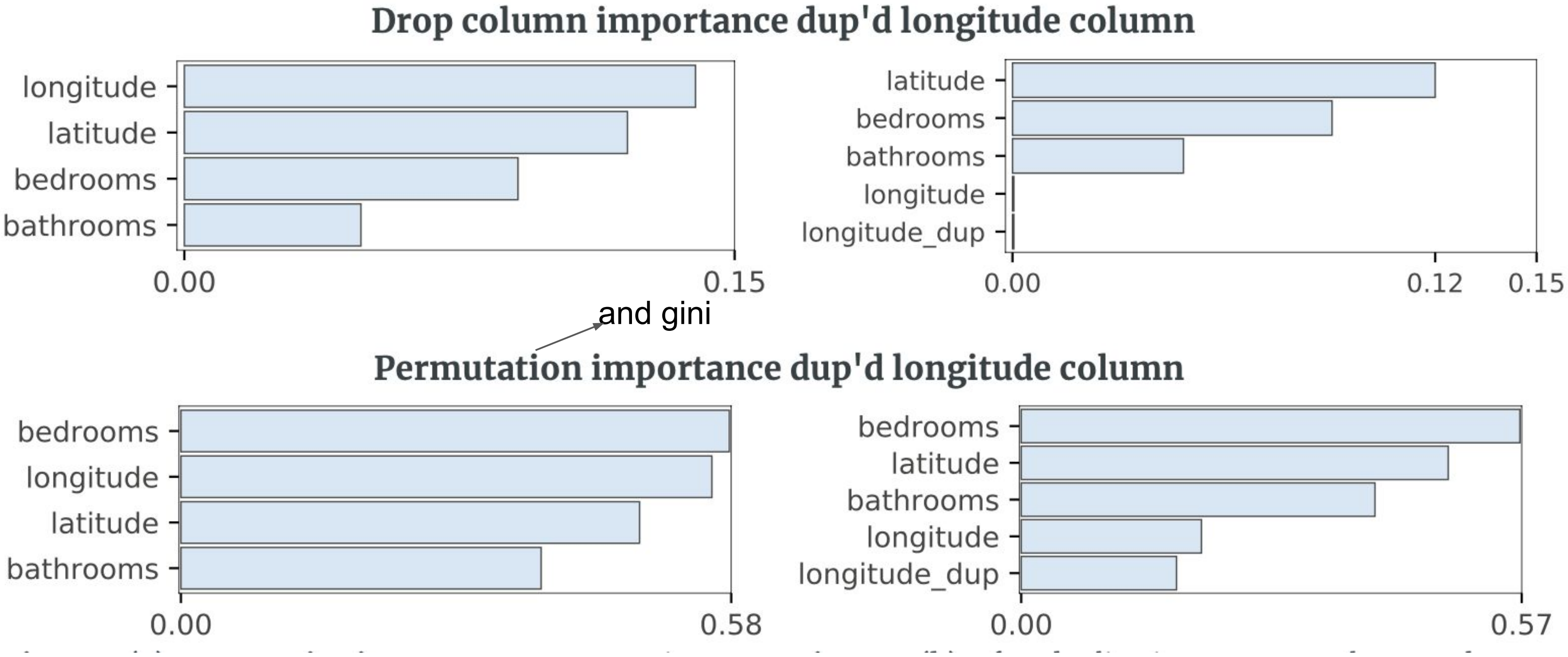
1. Gini importance/Selection frequency
2. Permutation importance
3. Drop-column importance

Different models that should help in theory:

1. Extremely randomized trees
2. Conditional inference trees



Different methods handles correlations differently



Drop column importance dup'd longitude column

and gini

Permutation importance dup'd longitude column