Lecture 24: Random Forests

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner

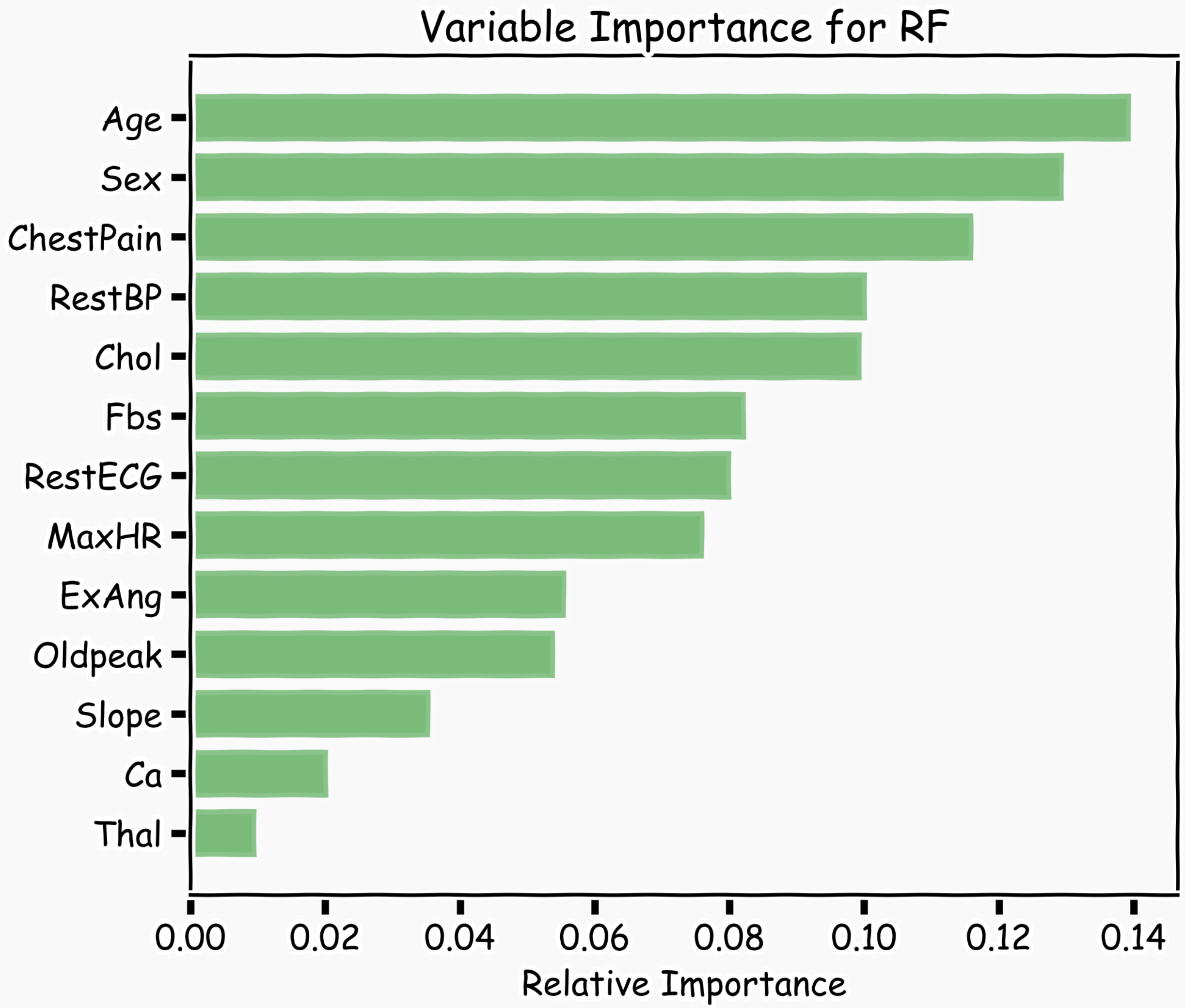
Outline

- Random Forest (RF)
- Tuning the hyperparameters of a RF
- **Feature interpretation in a RF**

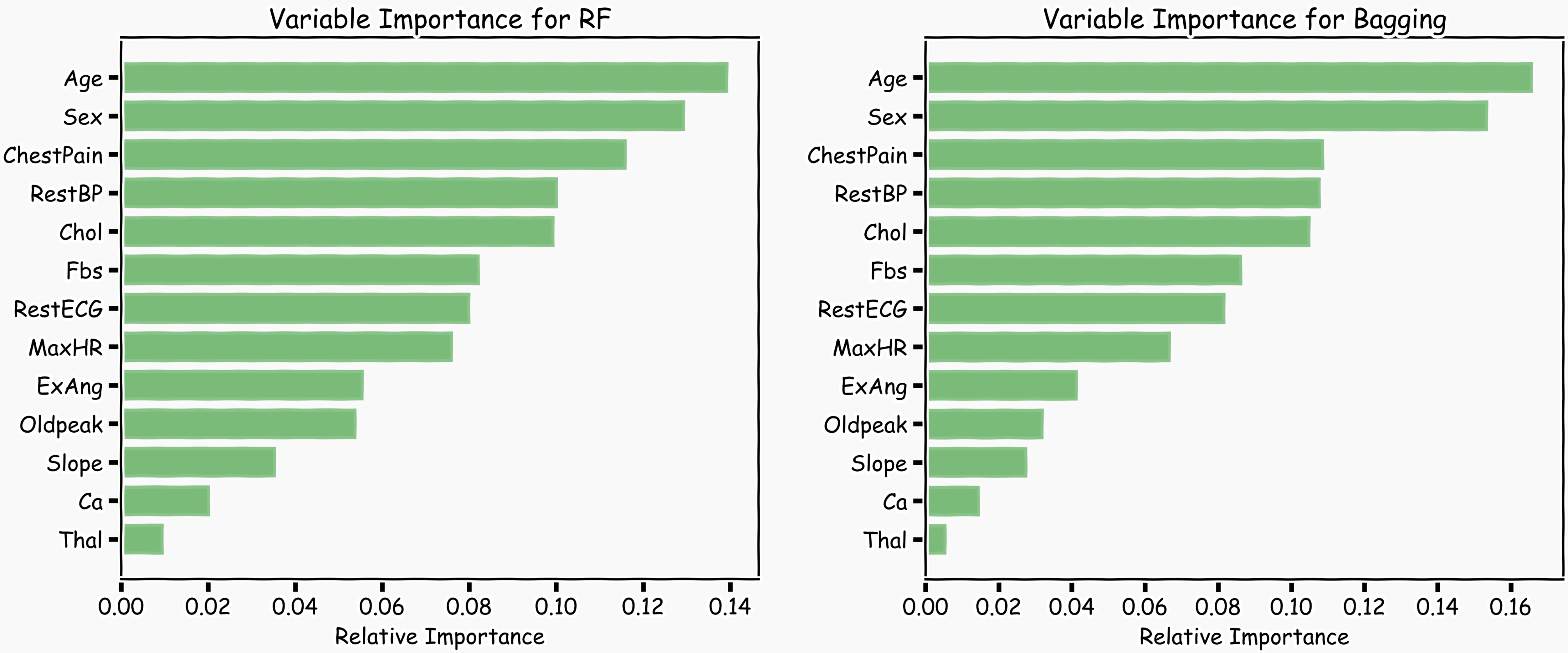
Variable Importance for RF

Explaining predictions from tree models is always desired; the patterns uncovered by a model are, in some applications, more important than the model's prediction performance.

A drawback of RF, Bagging, and other ***ensemble methods,*** is that the averaged model is no longer easily interpretable - i.e. one can no longer trace the *logic* of an output through a series of decisions based on predictor values!



Variable Importance for RF



100 trees, max_depth=10

Variable Importance for RF

1. Mean Decrease in Impurity (MDI)

- Same as Bagging.
- Record the prediction accuracy on the *oob* samples for each tree.
- Calculate the total amount that the RSS (for regression) or Gini index (for classification) is decreased due to splits over a given predictor, averaged over all trees.
- The decrease in accuracy as a result of this permuting is averaged over all trees, and is used as a measure of the importance of variable j in the random forest.
- The default in Scikit-learn `feature_importances_`

Variable Importance for RF

2. Permutation Importance

- Record the prediction accuracy on the *oob* samples for each tree.
- Randomly permute the data for column j in the *oob* samples the record the accuracy again.
- The decrease in accuracy as a result of this permuting is averaged over all trees, and is used as a measure of the importance of variable j in the random forest.

3. One step further (SHAP values, LIME)

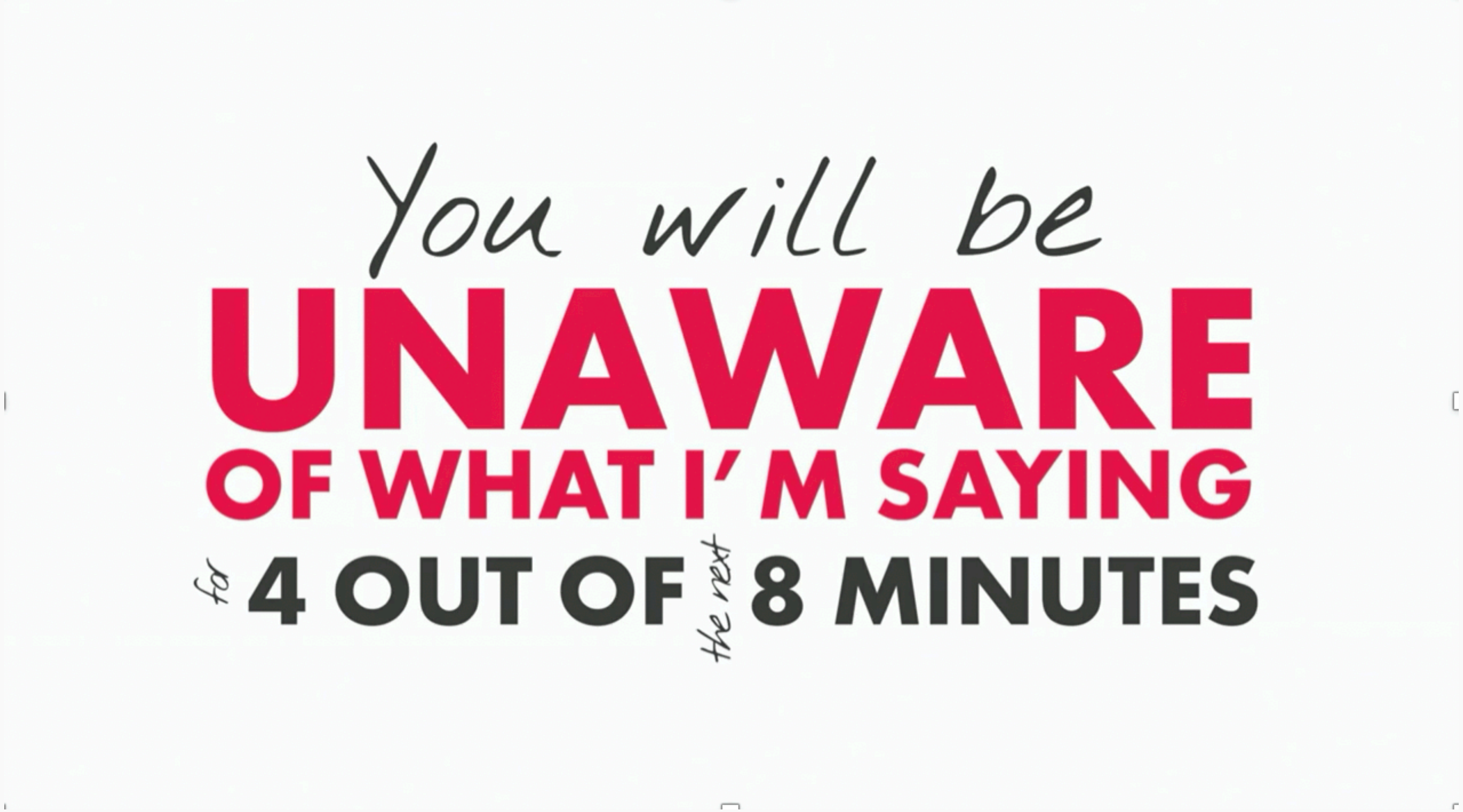
- We will see these methods in later lectures.

Final Thoughts on Random Forests

Increasing the number of trees in the ensemble generally does not increase the risk of overfitting.

Again, by decomposing the generalization error in terms of bias and variance, we see that increasing the number of trees produces a model that is at least as robust as a single tree.

However, if the number of trees is too large, then the trees in the ensemble may become more correlated, increase the variance.



You will be **UNAWARE OF WHAT I' M SAYING** for **4 OUT OF** the next **8 MINUTES**

Zeenat Potia

Next Lecture

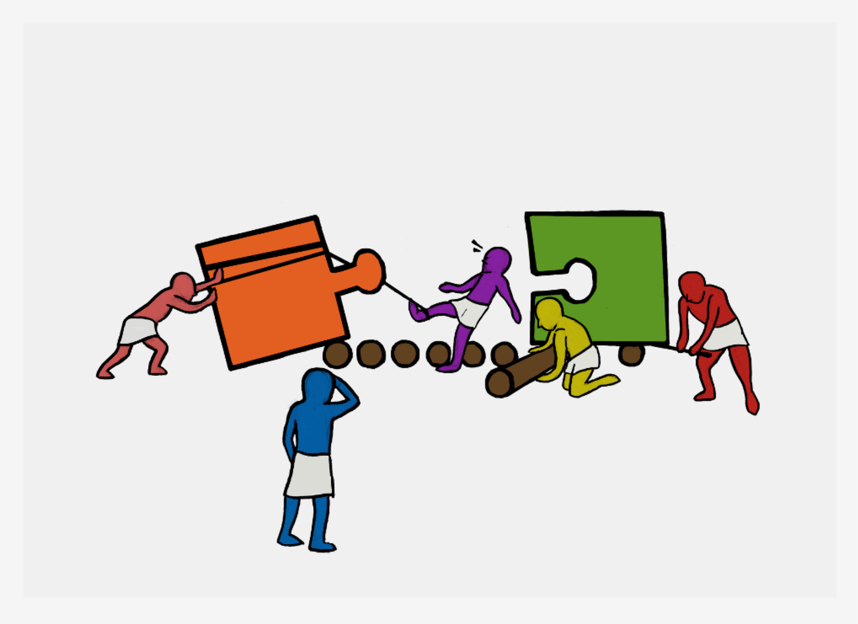
- Imbalanced dataset
  - Weighted samples
- Categorical data
- Missing data

AND BOOSTING!

Exercise 3

- Person sharing their screen is the one located **closest to NYC.**

- Be respectful of each other.



- Exercise 3 is about calculating feature importance in a single tree and a RF, using two methods.