# DSC5103 Statistics

Session 8. Bagging and Random Forest

# Last time

- The Basics of Decision Trees
  - Regression Trees
  - Building and Pruning Trees
  - Classification Trees
  - Trees vs. Linear Models
  - Advantages and Disadvantages of Trees
    - Suffer from big categorical variables
    - Low accuracy, High variance

# Plan for today

- The Bootstrap
- Bagging and Random Forest
- Ensemble methods in general

# Dealing with Uncertainties

- Uncertainties in predictive analytics
  - Sample uncertainty: data sample is a random draw from the population
  - Data partition uncertainty: due to validation/cross-validation
  - Tool related uncertainty: randomization inside the algorithm
  - Test data uncertainty

- All estimation/learning methods depend on the data sample, which is random. If we could repeatedly and independently sample from the population, we can
  - estimate the variance of the output (e.g., β_hat, trees)
  - average over multiple outputs to reduce variance

  - but …

# The Bootstrap

- The bootstrap: a flexible and powerful statistical tool for generating new samples ***from the current sample***

- The idea: to mimic the process of obtaining new data
  - Obtain distinct datasets by repeatedly sampling *n* observations from the original dataset **with replacement** => multiple bootstrap samples
  - Obtain estimates/predictions for each bootstrap sample => multiple bootstrap estimates
    - Calculate standard error/confidence interval (called Bootstrap Percentile) of the bootstrap estimates, which approximates the true standard error/confidence interval

    - Average over multiple predictions to reduce variance/overfitting

# The Bootstrap

# Bagging -- Motivation

- Decision trees suffer from high variance!
  - If we randomly split the training data into 2 parts, and fit decision trees on both parts, the results could be quite different

- To reduce variance
  - By regularization: pruning

  - By averaging:
    - Averaging over a set of trees if we have multiple training sets

    - If not, we can use bootstrap => bagging (***b***ootstrap ***agg***regat***ing***)

# Bagging in General

- Bagging is an extremely powerful idea based on two things:
  - Bootstrapping: generate plenty of training datasets
    - Lead to many parallel models with similar bias
  - Aggregating: reduce variance by averaging

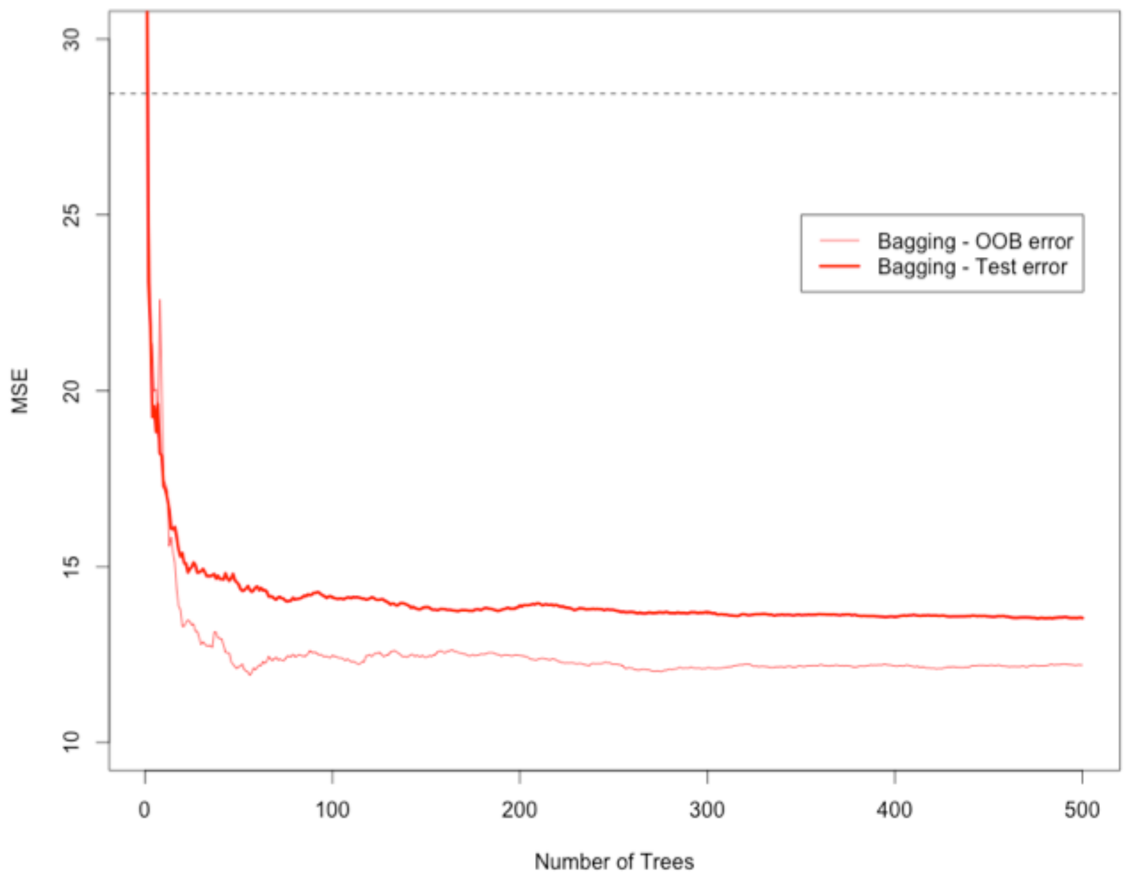
# Bagging for Trees

- Generate B different bootstrapped training datasets
- Construct B trees using the bootstrapped training datasets <u>without pruning</u>
  - Each individual tree has high variance but low bias
- Averaging these trees reduces variance, and thus we end up lowering both variance and bias ☺
- Prediction:
  - For regression: average the resulting predictions
  - For classification: majority vote by the class that each bootstrapped data set

# Bagging for Trees

# Example: Boston Housing Data

- Bagging does not overfit as the number of trees increases



# Out-of-Bag (OOB) Error Estimation

- A very straightforward way to estimate the test error of a bagged model
  - On average, each bootstrap sample takes around 2/3 of the observations, so we end up having the rest 1/3 <u>Out-of-Bag</u> observations
  - The remaining non-selected part could be used as the validation data
  - We can predict the response for the **i-**th observation using each of the trees in which that observation was OOB. This will yield around **B/**3 predictions for the **i-**th observation, which we average.
- This estimate is essentially the LOOCV error for bagging, if **B** is large
  - No more CV!

# Random Forest

- Motivation
  - Bagging works because of the variance reduction
    - Averaging the prediction of B *correlated* models
  - Random Forest: further de-correlate the trees to improve variance reduction
- How does it work?
  - Each time a split in a tree is considered, a random sample of ***m*** predictors is chosen as split candidates from the full set of ***p*** predictors
  - How to choose m?
    - If m=p, it becomes Bagging
    - Rule of thumb: m=p/3 for regression; m=sqrt(p) for classification
    - Can be easily tuned by OOB error

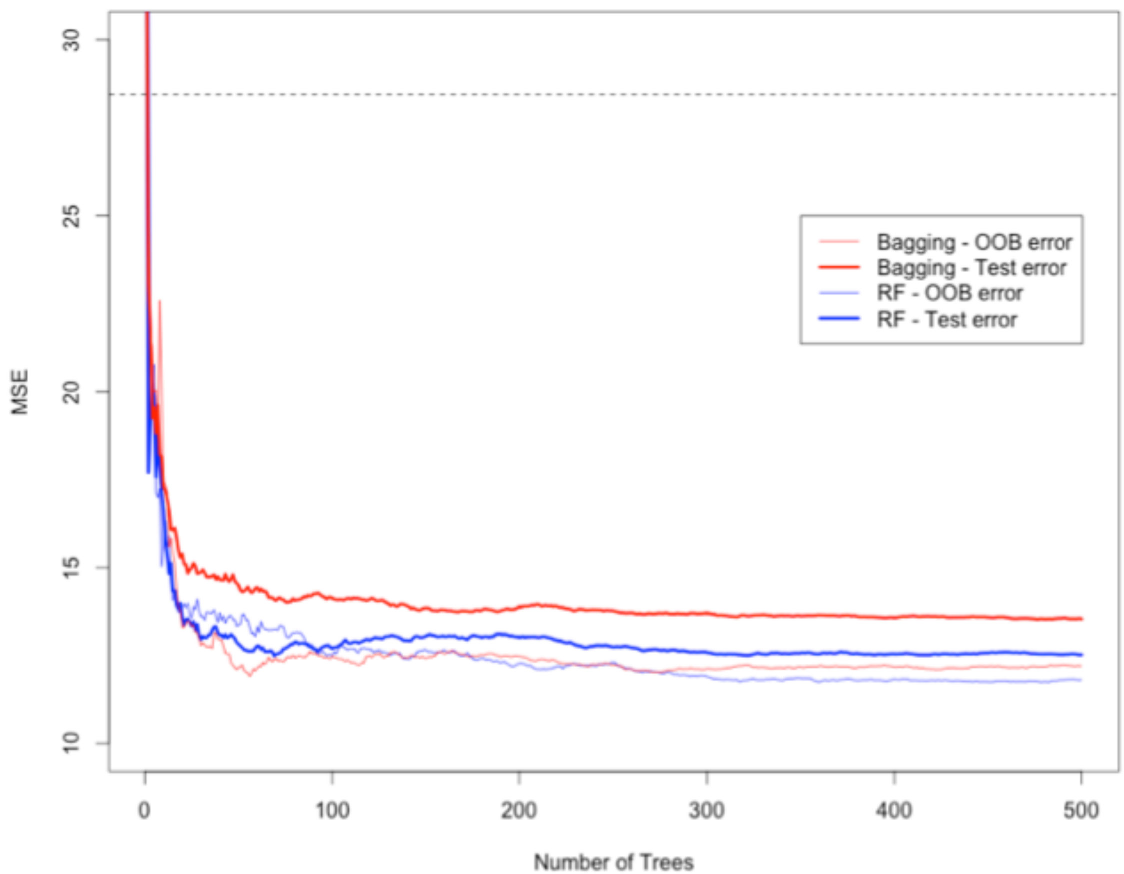
# Why it works?

- In Bagging, a strong predictor will dominate in all the trees
- All bagged trees will look similar => highly correlated predictions
- Averaging many highly correlated quantities does not lead to a large variance reduction

- Random forest “de-correlates” the bagged trees by giving chance to weak predictors
  - less correlated predictions
  - more reduction in variance

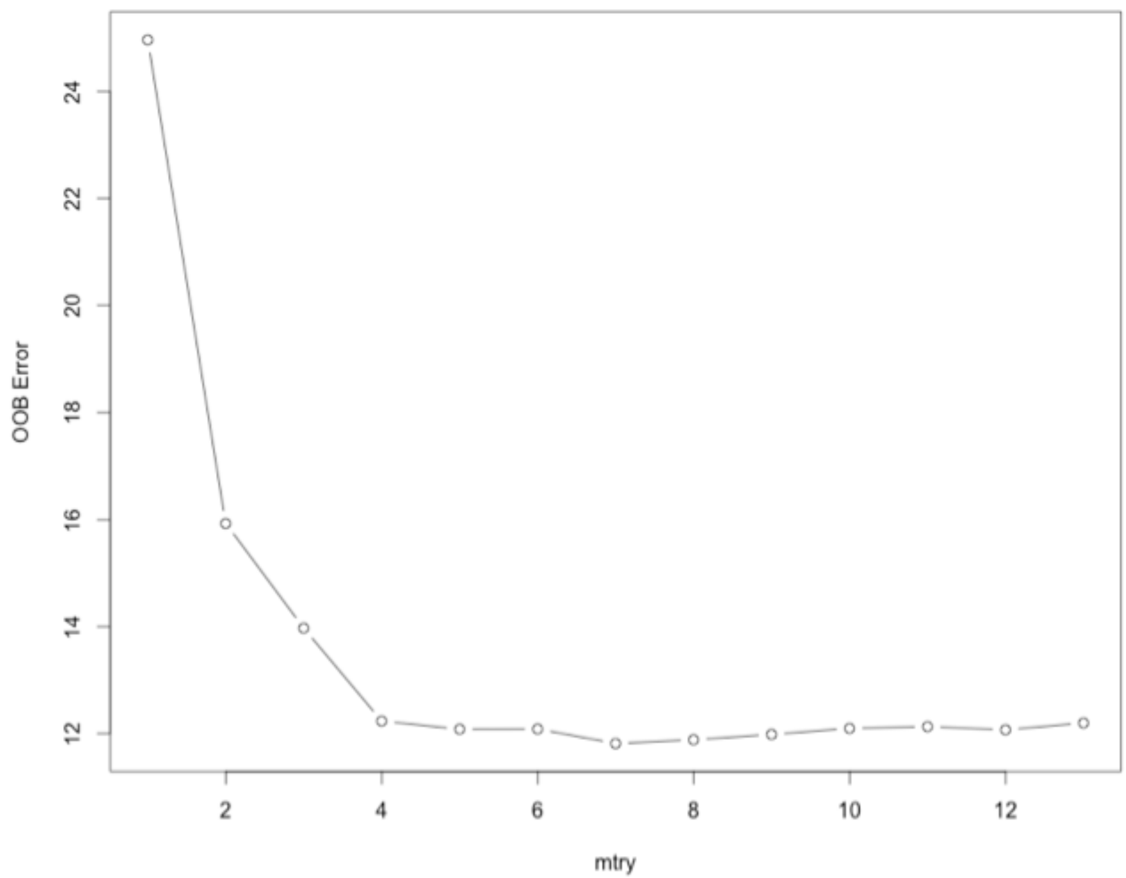
# Example: Boston Housing Data

- RF outperforms because it is a generalization of Bagging



# Random Forest with different values of “m”

- Tuning parameter m



# Interpretability

- Bagging/RF improves prediction accuracy at the expense of interpretability
  - Bagging/RF typically improves the accuracy over prediction using a single tree
  - It is no longer clear how to interpret the forest of trees

- But, we can still get an overall summary of the impact of each predictor
  - Relative Influence Plots
  - Partial Dependency plots

# Relative Influence Plots

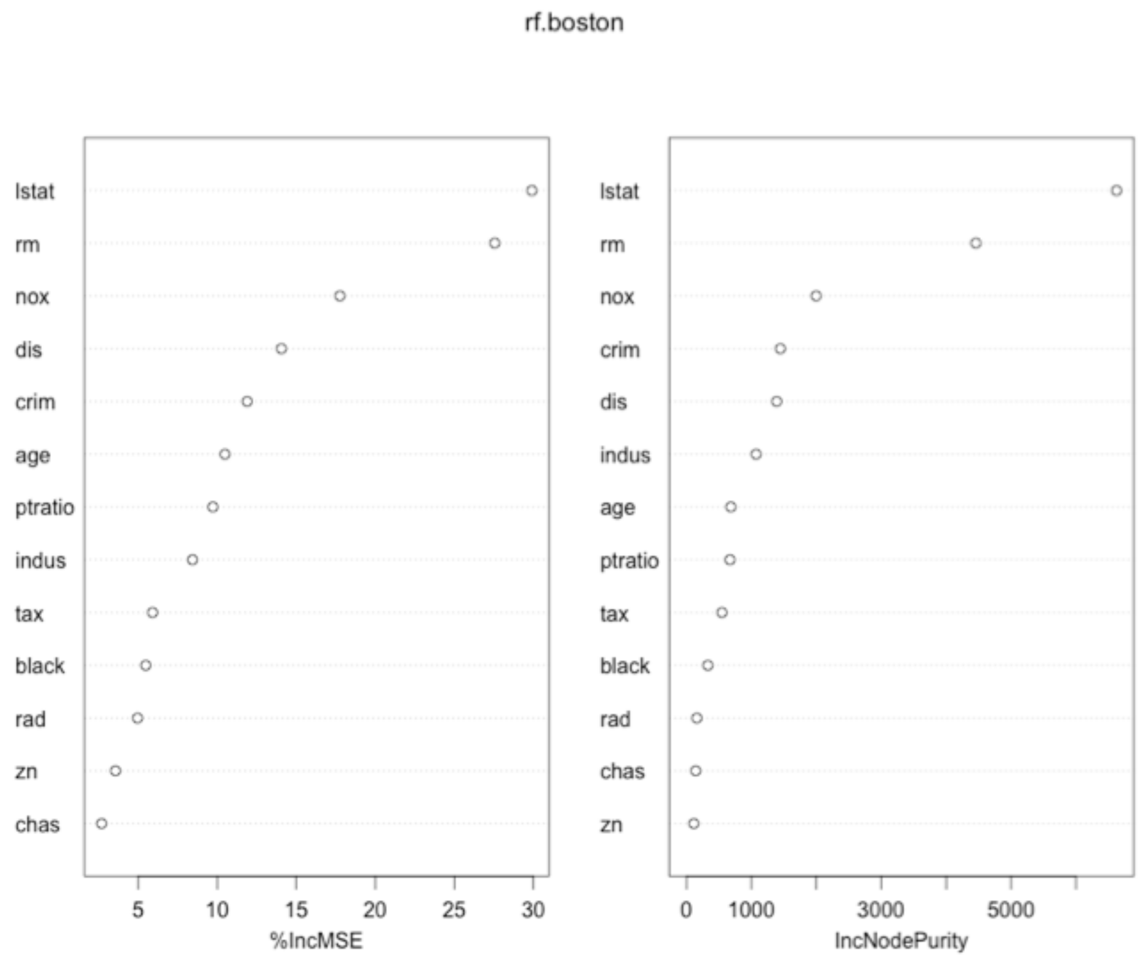
- How do we decide which variables are most useful in predicting the response?

- Relative Influence Plots

  1. The decrease in accuracy with vs. without a predictor, averaged over all trees

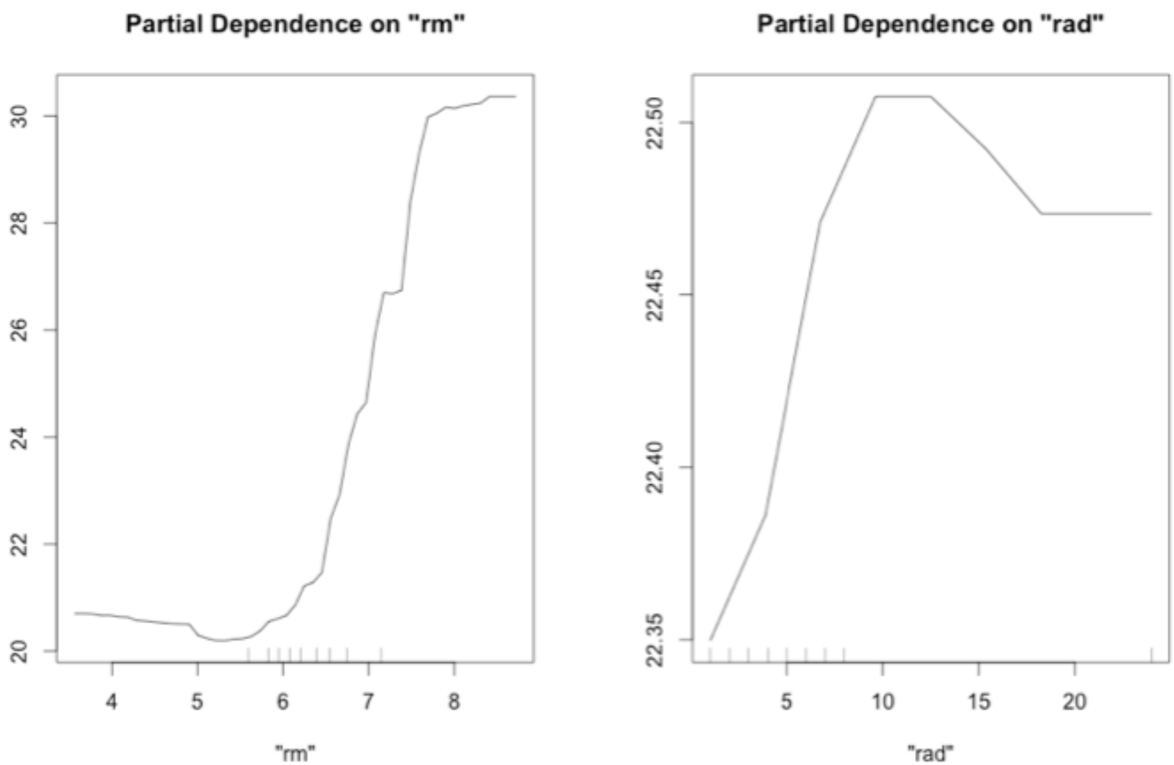
  2. The total amount of deviance/gini that is decreased due to splits over a given predictor, averaged over all B trees. A large value indicates an important predictor.

# Example: Boston Housing Data



# Partial Plot in RF

- Partial dependence on individual predictors



# Ensemble Methods

- Aggregation of predictions of multiple models with the goal of improving accuracy

- Famous example: the Netflix million dollar challenge
  - http://www.netflixprice.com
  - The top solution consists of blending of 107 individual results
  - Two top ranked teams united by merging their results and achieved better accuracy
  - Winning solution based on GBM:http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf

    http://www2.research.att.com/~volinsky/netflix/

# Ensemble Methods

- Intuition: Utility of combining diverse, independent opinions in human decision-making

- Example:
  - Vote by 5 completely independent classifiers with 70% accuracy => 83.7% accuracy
  - Vote by 101 such classifiers => 99.9% accuracy