Problem: variance trees

Solution is ensembling:

**Bagging – Bootstrap Aggregation**

Step bootstrap the data, create data set 1 > build dec tree 1

Step 2, BS data set 2 > dec tree 2

….

Aggregate all the predictions from all bootstrapped trees.

Regression: average all the predictions (mean)

Classification: majority vote.(mode)

OOB Out of Bag observations/error

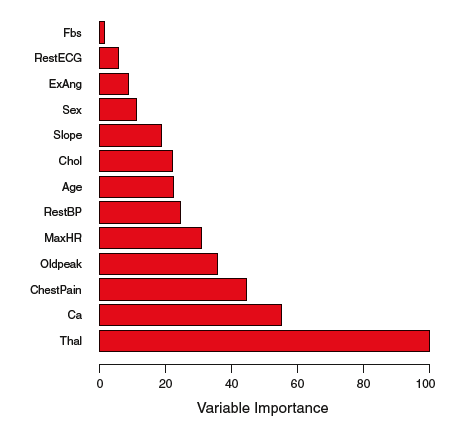
Bootstrap use 2/3

1/3 for test set. .

Grow the tree fully

Interpreting the features is lost bc we have many trees

SO, we use **feature importance**



Problem w bagging. Correlated trees: important variable, is present in almost all the trees and at the top.

Solution is Random Forest.

Random forest is very similar to bagging

Difference is in how we make out splits (which features we consider)

Every time we make a split, we take a random sample of features.

100 predictors. Sqrt =10 predctors

**Boosting**

XGBoost example package

Still ensemble of trees.

Difference is trees are built sequential. Residual output of the first tree is input to the next tree.

Short trees, stump (1 split)

Slow learning. Progressing. Parameter lambda (learning rate. Typically .01 or 0.001)

Dependent on each other.

Difficult to use on Distributed cluster system. RF are better for that.