Random Forest

Wisdom of Crowd- The Idea

In 1906, the statistician Sir Francis Galton was visiting a county fair in England, at which a contest was being held to guess the dressed weight of an ox that was on exhibit. There were 800 guesses, and, while the individual guesses varied widely, both the mean Guess of all People came out within 1% of the ox's true weight.

James Suroweicki has explored this phenomenon in his book *The Wisdom of Crowds* (Doubleday, 2004). This principle applies to predictive models, as well: averaging (or taking majority votes) of multiple models—an *ensemble* of models—turns out to be more accurate than just selecting one model.

The Word *Ensemble means* Forming a prediction by using a collection of models.

Random Forest is one of the several Ensemble Models that can be used for both Regression & Classification. A RF model builds several Decision Trees (hence called a Forest) and finally combines their results to make a prediction.

Random Forest is also called a Bagging Model (rf uses a technique called Bagging to create the different Trees).

Decision trees discussed in previous Module suffer from high variance.This means that if we split the training data into two parts at random, and fit a decision tree to both halves, the results that we get could be quite different.

In contrast, a procedure with low variance will yield similar results if applied repeatedly to distinct data sets; linear regression tends to have low variance, [if the ratio of n (number of rows) to p (number of columns) is moderately large].

Bootstrap aggregation, or bagging, is a general-purpose procedure for reducing the variance of a statistical learning method; it is particularly useful and frequently used in the context of decision trees. A natural way to reduce the variance and hence increase the prediction accuracy of a statistical learning method is to take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions. This is called Bagging.

Bagging has been demonstrated to give impressive improvements in accuracy by combining together hundreds or even thousands of trees.

Random forests provide an improvement over bagged trees by way of a random small tweak that decorrelates the trees. As in bagging, we build a number of decision trees on bootstrapped training samples. *But when building these decision trees, 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. The split is allowed to use only one of those m predictors. A fresh sample of m predictors is taken at each split,*

Random Forest OOB Error

Each bagged tree makes use of around two-thirds of the observations. The remaining one-third of the observations not used to fit a given bagged tree are referred to as the out-of-bag (OOB) observations. The OOB (out of Bag) estimate of error is the error Rate for the trained models, applied to the data left out of the training set for that tree. When we aggregate this over all the trees, we can get the OOB Error for the RF model.

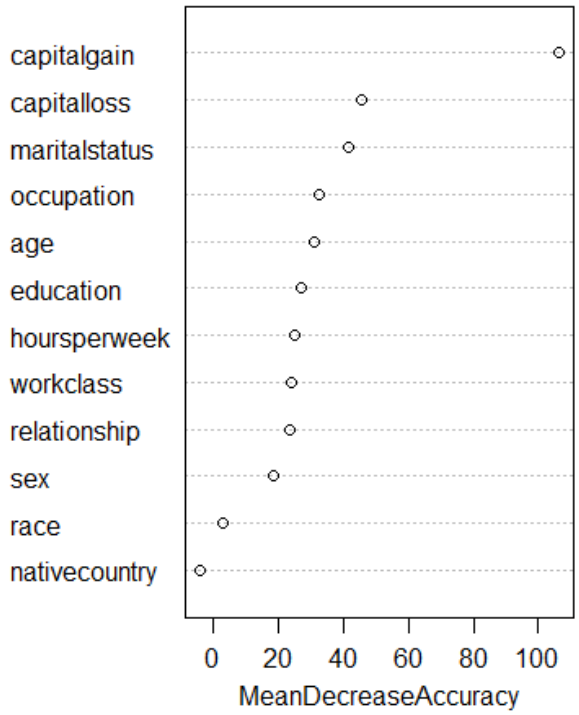
This resulting OOB error is a valid estimate of the test error for the bagged model, since the response for each observation is predicted using only the trees that were not fit using that observation. This is a good proxy error for the test data even before predicting on it.

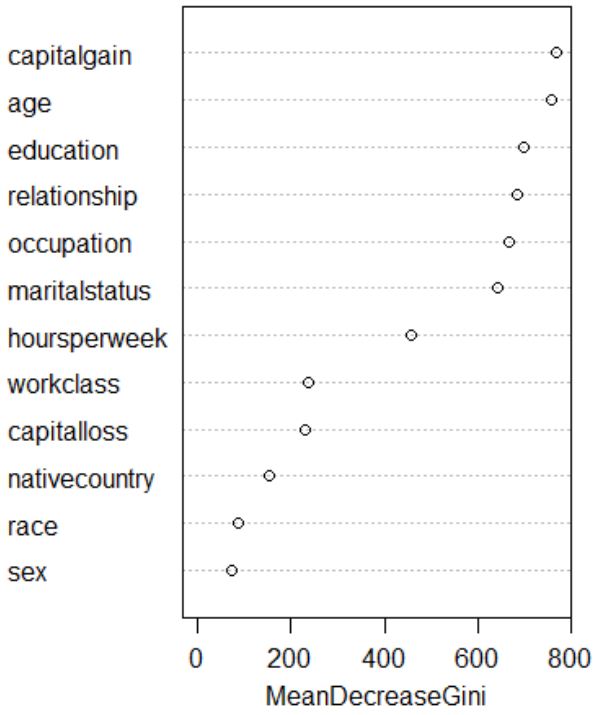
Variable Importance Plot in RF

Random Forest has the ability to automatically determine which predictors are important and discover complex relationships between predictors. The Variable Importance Plot can be used to find the relative Importance of the predictors in the Data. The Variable Importance Plot is found using the following-

1) The values of each variable is randomly permuted (this is like taking junk values for the variables). Then the decrease in Accuracy of the Model is meausured on the out of bag (OOB) sample. If the accuracy decreases a lot, it means the variable is very important. (LHS plot below)

2) For each predictor, the decrease in Gini Index due to the Predictor being selected as a Decision Node is measured (for regression tree the Decrease in SSE is measured). This measures how much improvement to purity of the buckets that variable contributes. A large value of this parameter indicates an important predictor. *This is measured on the training set and hence is less reliable.* (RHS plot below)


capitalgain
capitalloss
maritalstatus
occupation
age
education
hoursperweek
workclass
relationship
sex
race
nativecountry
0
20
40
60
80
100
MeanDecreaseAccuracy


capitalgain
age
education
relationship
occupation
maritalstatus
hoursperweek
workclass
capitalloss
nativecountry
race
sex
0
200
400
600
800
MeanDecreaseGini