IE 425

Homework 2 (due May 7, Sunday @23:59)

1. Consider the dataset given in the file "bank.csv". The data is related with direct marketing campaigns of a Portuguese bank. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required in order to access if the product (bank term deposit) would be subscribed or not. The classification goal is to predict if a client will subscribe to a term deposit, called "y" in the dataset. Use a seed value of 425 whereever you need a seed.

a) Partition the dataset using the caTools package into training and test sets where 80% of the observations go into the training set and 20% goes into the test set.

b) Determine the best random forest (based on the random forest package) by using 10-fold cross validation five times with the caret package on the training set by playing with the mtry and ntree parameters. What are the best values of these two parameters?

c) What is the out-of-bag accuracy? Comment on which input attributes are important in making predictions.

d) Provide the Confusion Matrix along with sensitivity, specificity, precision, recall, and the F measure on the test set obtained by the best random forest. Does the out-of-bag accuracy provide a good estimate for the accuracy on the test set?

e) Repeat part b with the gradient boosting machine using the caret and gbm packages by playing with the interaction.depth, n.trees, shrinkage, and n.minobsinnode parameters. What are the best values of these four parameters?

f) Provide the Confusion Matrix along with sensitivity, specificity, precision, recall, and the F measure on the test set obtained by the best boosting tree.

2. Consider the dataset given in the file "SeoulBikeData.csv". The output attribute to be predicted is the "Rented Bike Count". Use a seed value of 425 whereever you need a seed.

a) Partition the dataset into training and test sets where 80% of goes into the training set and 20% goes into the test set.

b) Determine the best random forest (based on the random forest package) by using 10-fold cross validation five times with the caret package on the training set by playing with the mtry and ntree parameters. What are the best values of these two parameters?

c) Comment on which input attributes are important in making predictions.

d) Make predictions in the test set and report the root mean square error rate and mean absolute error.

e) Repeat part b with the gradient boosting using the caret and gbm packages by playing with the interaction.depth, n.trees, shrinkage, and n.minobsinnode parameters. What are the best values of these four parameters?

f) Make predictions in the test set and report the root mean square error rate and mean absolute error.