IE 425

Homework 2 (due April 21, Thursday @23:59)

1. Consider the "Human Resources Analytics" problem which is based on the data set "HR.csv".Why are our best and most experienced employees leaving prematurely? Have fun with this database and try to predict which valuable employees will leave next. Fields in the dataset include:

- Satisfaction Level
- Last evaluation
- Number of projects
- Average monthly hours
- Time spent at the company
- Whether they have had a work accident
- Whether they have had a promotion in the last 5 years
- Departments (column sales)
- Salary
- Whether the employee has left

Our goal is to predict the profile of employees who are likely to quit.

a) Partition the data set into training and test sets with 80% going into the training set by using a seed value of 425. Whenever you need to use set.seed function, use set.seed(425).

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 and what is the out-of-bag accuracy? Comment on which input attributes are important in making predictions.

c) Provide the Confusion Matrix along with sensitivity, specificity, precision and recall on the test set obtained by the best random forest.

d) 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?

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

2. Consider the dataset given in the file "ToyotaCorolla.csv". The output attribute to be predicted is the Price attribute. 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 and what is the out-of-bag accuracy? Comment on which input attributes are important in making predictions.

c) Make predictions in the test set and report the root mean square error rate and mean absolute error using the functions in the Metrics package.

d) 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?

e) Make predictions in the test set and report the root mean square error rate and mean absolute error using the functions in the Metrics package.