FE-582 – Assignment 3

Problem 1

Follow the example in Lecture 5 R code to analyze the pair trading strategy for several pairs taken from the following list: PEP, KO, DPS for a period of at least 10 years. You may download stock data from WRDS (you can register and create an account as Stevens student at https://wrds-web.wharton.upenn.edu/wrds/) or other free data sources such that Yahoo using library("quantmod"). Include transaction costs of 0.1% (or 10 basis points) for each transaction. Determine the optimal k for each pair and report the profit and loss in each case. Discuss the results.

Problem 2

Use the data set Default.csv which has 7,000 observations on the following 4 variables:

- default - A factor with levels No and Yes indicating whether the customer defaulted on their debt
- student - A factor with levels No and Yes indicating whether the customer is a student
- balance - The average balance that the customer has remaining on their credit card after making their monthly payment
- income - Income of customer

Apply logistic regression, linear discriminant analysis, quadratic discriminant analysis and K-nearest neighbor classification methods to predict customers that are likely to default in DefaultPredict.csv dataset. Please use several values of K in the KNN classification method such that you can minimize the errors. Compare the errors for all the methods and draw conclusions.

Problem 3

This problem use the OJ data set which is part of the ISLR package.

a) Create a training set containing a random sample of 700 observations, and a test set containing the remaining observations.
b) Fit a tree to the training data, with Purchase as the response and the other variables as predictors. Use the summary() function to produce summary statistics about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?
c) Type in the name of the tree object in order to get a detailed text output. Pick one of the terminal nodes, and interpret the information displayed.

d) Create a plot of the tree, and interpret the results.
e) Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?
f) Apply the cv.tree() function to the training set in order to determine the optimal tree size.
g) Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis.
h) Apply random forests method and check the performance of the model.
i) Apply boosting method and check the performance of the model.
j) Discuss the comparison in performance by applying different tree methods.