FE-582 – Assignment 2

Problem 1

Use the data set Default.csv which has 10000 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.

Problem 2

Use the Lecture_7_data.zip file.
Certain words like "and," "the," and "of," are very common in all English sentences and are not very meaningful in deciding spam/nonspam status, so these words have to be removed from the emails. Download a list of stop words from http://www.ranks.nl/resources/stopwords.html and remove them from the emails (word list).
Do the following;

- Transform a message body into a set of the words.
- Combine the words across messages into a bag of words.
- Tally the frequencies in the training data of words in spam and ham separately to estimate the probability of a word appears in a message given it is spam (or ham) from the proportion of spam (or ham) messages containing that word.
- Estimate the likelihood that a new test message is spam (or ham) given its contents, i.e., given the message's words compute the naïve Bayes version of the log likelihood ratio.
- Find a threshold for the log likelihood ratio, where a message with a value above the threshold is classified as spam. Choose this threshold by examining the error rates for the test data.

Compare the classification results obtained by removing the stop words with the case when you don't remove words from analysis.

Problem 3

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. Include transaction costs of 0.02% (or 2 basis points) for each transaction. Discuss the results