**Homework 6: Predicting German Credit**

**Due Feb 28 11:59 PM (20 points)**

IST 5535 – Spring 2021, Chen

***Instruction:*** Finish the following tasks. Type your answers or explanations directly in R Markdown.

1. The data file "credit-g.csv" contains a German credit dataset. The dataset contains 1000 observations (customers) of 21 variables. The response variable is "class". Read in the dataset and explore it. How many customers in the dataset have good credit?
2. Use ggplot2 package to draw a barchart of 'credit\_history' grouped by 'class'. The barchart should satisfy below criteria:

- The number of customers under each credit history category should be grouped by class

- Has a main title "Distribution of Credit History Grouped by Class"

- Has x axis labeled as "Credit History"

- Has y axis labeled as "Number of Customers"

Explain what insights you can get from the plot regarding predicting customer credit class.

1. Use ggplot2 package to draw a boxplot of 'credit\_amount' grouped by 'class'. Explain what insights you can get from the plot regarding predicting customer credit class.
2. Randomly split the whole dataset as two parts: training set containing 80% of the data, and test set containing 20% of the data.
3. Conduct a logistic regression to predict bad credit customers in the training set, using all other variables as predictors. Treat “bad” as the positive class and “good” as negative class. That is, the following model specification is used:
4. Explain variables that have significant positive or negative effects. Do these directions of these effects make sense? Are the logistic regression results consistent with your findings in steps 2 and 3?
5. Evaluate the performance of the logistic regression on the test set. Calculate overall accuracy, sensitivity, and specificity. Which measure is best to evaluate how the model predicts bad credit customers? Does this logistic regression model do a good job in classifying bad credit customers?
6. Perform LDA on the training data in order to predict class using only the variables that are found significantly impacting the credit class in the logistic regression analysis (include the whole categorical variable even some levels are not significant in logistic regression). Calculate overall accuracy, sensitivity, and specificity on the test data. Does the model do a good job in classifying bad credit customers?
7. Perform QDA on the training data in order to predict class using only the variables that are found significantly impacting the credit class in the logistic regression analysis (include the whole categorical variable even some levels are not significant in logistic regression). Calculate overall accuracy, sensitivity, and specificity on the test data. Does the model do a good job in classifying bad credit customers?
8. Compare logistic regression, LDA, and QDA. Which model performs best for predicting bad credit customers?

Submission:

1. Submit to Canvas both the R Markdown file the HTML report directly generated from the R markdown.
2. Make sure the HTML report is clear and has an easy-to-read structure. You will lost 5 points if the HTML is too long such as displaying all data points, or the HTML report contains other content (such as some content in the R markdown template) that is not relevant to the programming task.