Project 3

Jose Alfaro

Due Date: March 11, 2019

Question: 1

Part: A

Exploratory data analysis was conducted on the admissions dataset. Here, we created a scatterplot in which we plotted GMAT scores against GPA and color coded these values based on their admissions decisions where green represented admitted individuals, red represented rejected individuals, and yellow represented individuals that were borderline (*Figure 1*). We see that there is a clear division between these three groups since there is barely any mixing between the observations. To get deeper insights on the effects that GPA and GMAT have on admission status, we created boxplots to explore the means and variances of these variables (*Figure 2*). Here, we observed that individuals that had a GPA of 3.0 or higher typically got accepted or were under consideration while those individuals who had a GPA between 2.3 and 2.7 were rejected. Also, we saw that individuals who had an average GMAT score of about 550 were accepted while individuals who had a GMAT score of about 455 were either placed rejected or placed under the borderline category. From this, we can see that GPA could potentially serve as a good predictor in predicting admission status since there seems to be clear divisions between the three categories. We also see that GMAT could serve as a good admissions indicator as well since those individuals that were admitted had exceptional GMAT scores as opposed to those who were rejected or borderline.

Part: B

$$\text{Decision Boundary Equation: } GMAT = 7,069.09 - 2,436.9331 * GPA$$

$$\text{Decision Boundary Equation: } GMAT = 1,734 - 388.1961836 * GPA$$

After conducting the linear discriminant analysis, we were ablet to get an equation for the decision boundary that separated these observations into three general groups. The decision boundary was superimposed onto a scatterplot of the data which plotted the two predictors (*Figure 3*). By eyeballing the superimposed decision boundary on the data, we can see that this is a very sensible boundary since all but one of the observations land within their prospective groups. Furthermore, the only observation that was not categorized under the correct admission category was only misgrouped by a marginal amount. The confusion matrices were computed for both test and training data and are displayed below (*Table 1 and Table 2*). Here we see that the overall misclassification rate for the training and test datasets are:

$$MR(Train) = \frac{1+2}{1+2+20+23+24} = \frac{3}{70} = 4.29\%$$

$$MR(Test) = \frac{1+2}{1+2+3+4+5} = \frac{3}{15} = 20\%$$

We observed that the overall misclassification rate for the testing dataset was about 4.66 times larger than that of the training dataset. This was expected since prediction rates will always be better when using the training dataset as opposed to the testing dataset.

Part: C

$$\mathbf{X} = \begin{bmatrix} 3.375 & 3.375^2 & 1 & 561.3846 \\ 2.43 & 2.43^2 & 1 & 453.5652 \\ 2.991905 & 2.991905^2 & 1 & 447.9524 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & -1766.15826 \\ 0 & 1 & 0 & 323.90260 \\ 0 & 0 & 1 & 2832.71733 \end{bmatrix}$$

Decision Boundary Equation: $GMAT = -1,766 * GPA + 323.90260 * GPA^2 + 2,832.71733$

After conducting the quadratic discriminant analysis, we were able to get an equation for the decision boundary that separated these observations into three general groups. The decision boundary was superimposed onto a scatterplot of the data which plotted the two predictors (*Figure 4*). When comparing the decision boundary produced by LDA with the decision boundary produced by QDA we noticed that the decision boundary produced by LDA was quite linear and didn't gave a general shape to the data. On the other hand, the decision boundary produced by QDA had more curvature and defined the groups in a more flexible manner. By eyeballing the superimposed QDA decision boundary on the data, we can see that this is a very sensible boundary as well since all but one of the observations landed within their prospective groups. Furthermore, the only observation that was not categorized under the correct admission category was only misgrouped by a marginal amount. Also, one may note that this miscategorized observation was closer to its prospective group than was the misclassified observation in the LDA case. The confusion matrices were computed for both test and training data and ara displayed below (*Table 3 and Table 4*). Here we see that the overall misclassification rate for the training and test datasets are:

$$MR(Train) = \frac{1+1}{1+1+20+23+25} = \frac{2}{70} = 2.86\%$$

$$MR(Test) = \frac{2}{2+3+5+5} = \frac{2}{15} = 13.33\%$$

We observed that the overall misclassification rate for the testing dataset was about 4.66 times larger than that of the training dataset. This was expected since prediction rates will always be better when using the training dataset as opposed to the testing dataset. However, the fact that the testing misclassification rates are 4.66 times larger than the training misclassification rates in the LDA and QDA case is a coincidence. Also, we note that the QDA procedure has the lowest misclassification rates for both the testing and training datasets. This hints that QDA preforms better than LDA in this situation.

Part: D

We conducted the KNN algorithm several times for values K = 1 through K = 70. This allowed us to observe the "bigger picture" and ensure that there is indeed a "U-shaped" curve when plotting error rates vs K (*Figures 5 and 6*). After confirming the concave up curve in the testing error rates, we then found *K = 16* to have the smallest testing error rate (*K =16* had $Test_{Err}$ = 20% and $Train_{Err}$ = 40%). Lastly, we superimposed the decision boundary from the KNN algorithm onto the scatterplot data (*Figure 7*).

Part: E

We found that out of the three techniques explored in this problem, LDA and KNN both produced a test error rate of 20% while QDA produced the lowest test error rate of 13.33%. This would imply that the best preforming classifier for this problem is the QDA classifier.

Question: 2

Part: A

An exploratory data analysis was conducted for the bankruptcy dataset in which we created scatterplots for all possible combinations of the variables X1, X2, X3, and X4 (*Figure 8*). Here, we observe that X2 and X1 seem to be correlated in some manner since the variance within the data appears to be minimal. Furthermore, variables X1 and X3 seem to do the best job in distinguishing those observations that went bankrupt and those that didn't since there seems to be a clear separation between the two responses. All of the other variable combinations don't appear to present a clear distinction between bankrupt and non-bankrupt observations. Therefore, our initial suspicions are that variables X1, X2, and X3 would be important.

Part: B

After creating a full model for containing all predictors (X1, X2, X3, and X4), we conducted a the stepwise model selection procedure in order to select the best subset of variables that explain most of the variation of the data. The stepwise selection procedure selected the model with the lowest AIC value (AIC = 34.64), which happened to contain only two predictors: X1 and X3 .Therefore the final model produced was:

$$log(\frac{p(y=1)}{1-(p=1)}) = \beta_0 + \beta_1 * X1 + \beta_2 * X3$$

with the following β_i coefficients: $\beta_0 = 5.940, \beta_1 = -6.556$, and $\beta_2 = -3.019$. Lastly, we interpreted these coefficients as:
- For every one unit change in X1, the log odds of going bankrupt (versus not going bankrupt) decreases by 6.556.
- For every one unit change in X3, the log odds of going bankrupt (versus not going bankrupt) decreases by 3.019.

Question: 3

Part: A

We used the logistic regression model built in the last step to provide and equation for the decision boundary which is represented as: X_3 = -($\frac{\beta_0}{\beta_2}$) - ($\frac{\beta_1}{\beta_2}$) * X_1. This decision boundary was then superimposed onto the scatterplot of X3 Vs. X1 since these were the predictors selected for the reduced model ($Figure$ 9). Furthermore, the confusion matrix was calculated and is represented in $Table$ 5. From the $Table$ 5, we see that:

$$Sensitivity = \frac{18}{21} = 85.71\%$$

$$Specificity = \frac{24}{25} = 96\%$$

$$MR = \frac{1+3}{1+3+18+24} = \frac{4}{46} = 8.7\%$$

Lastly, we plotted the ROC curve which can be found in $Figure$ 12 and saw that the area under the curve was 0.9371. Which implies that the test is accurate.

Part: B

We used the logistic regression model containing all predictors to provide and equation for the decision boundary which is represented as: $X_3 = \frac{-\beta_0}{\beta_3} - \frac{\beta_1}{\beta_3} * X_1 - \frac{\beta_2}{\beta_3} * X_2 - \frac{\beta_4}{\beta_3} * X_4)$. This decision boundary was then superimposed onto the scatterplot of X3 Vs. X1 since these were the predictors selected for the reduced model ($Figure$ 9). Furthermore, the confusion matrix was calculated and is represented in $Table$ 6. From $Table$ 6, we see that:

$$Sensitivity = \frac{18}{21} = 85.71\%$$

$$Specificity = \frac{24}{25} = 96\%$$

$$MR = \frac{1+3}{1+3+18+24} = \frac{4}{46} = 8.7\%$$

It is important to note that these values are exactly the same as the reduced model. Lastly, we plotted the ROC curve which can be found in $Figure$ 12 and saw that the area under the curve was 0.9410. Which implies that the full model is slightly more accurate than the reduced model when predicting whether an observation will be bankrupt or not.

Part: C

$$\text{Decision Boundary Equation:} X_3 = 1.689724 - 2.197994 * X1$$

We repeated all of part A except this time we used all of the predictors using the LDA method. The decision boundary was superimposed onto a scatterplot of the two most important predictors (*Figure 10*). The confusion matrix can be found in *Table 7*. Note here that this confusion matrix is the same as the two for logistic regression. with that being said, we see from *Table 7* that:

$$Sensitivity = \frac{18}{21} = 85.71\%$$

$$Specificity = \frac{24}{25} = 96\%$$

$$MR = \frac{1+3}{1+3+18+24} = \frac{4}{46} = 8.7\%$$

Lastly, we plotted the ROC curve which can be found in *Figure 12* and saw that the area under the curve was 0.9333 which implies that LDA is fairly accurate.

Part: D

Again, we repeated all of part A except this time we used all of the predictors using the QDA method. The decision boundary was computed and plotted over the scatterplot of the two important predictors (*Figure 11*). The confusion matrix can be found in *Table 8*. We see from *Table 8* that:

$$Sensitivity = \frac{19}{21} = 90.48\%$$

$$Specificity = \frac{24}{25} = 96\%$$

$$MR = \frac{1+2}{1+2+19+24} = \frac{3}{46} = 6.52\%$$

Lastly, we plotted the ROC curve which can be found in *Figure 12* and saw that the area under the curve was 0.9695 which implies that QDA is more accurate than LDA. Furthermore, we see that QDA appears to have better accuracy than every other method discussed in this problem. Therefore, QDA is the better classifier in this case as well.

Part: E

Since QDA has the lowest misclassification rate than all of the other methods discussed, this suggests that QDA is the optimal method for this problem. Furthermore, QDA had the highest AUC than both of the logistic models and the LDA method. This reinforces our conclusion that QDA is the optimal method to use for this problem. As a reminder, the reduced logistic model produced an AUC of 0.9371, the full logistic model produced an AUC of .9410, LDA produced an AUC of 0.9333, and QDA produced an AUC of 0.9696.

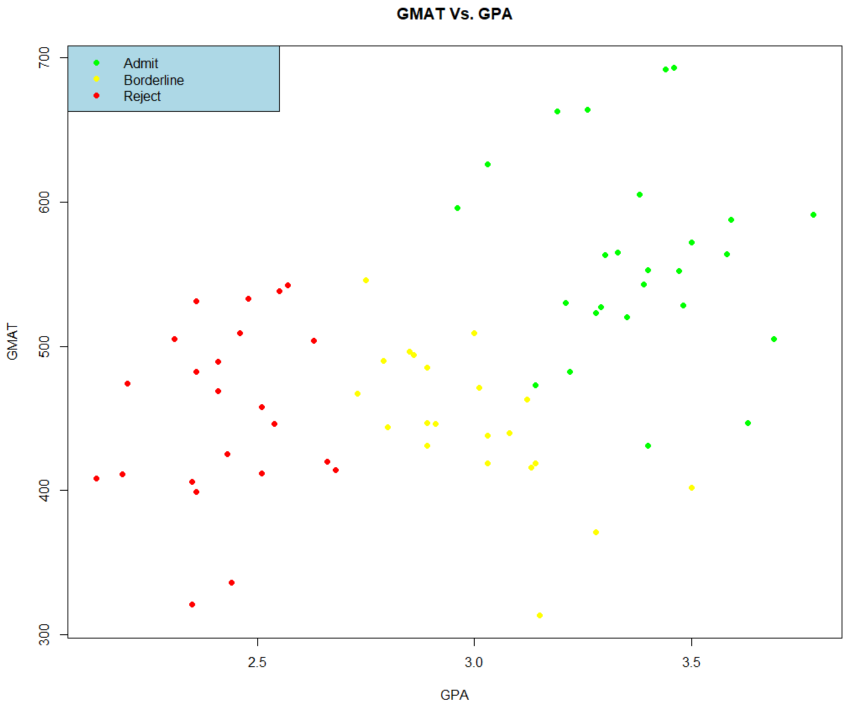

Figure 1: Scatterplot of GMAT Vs. GPA

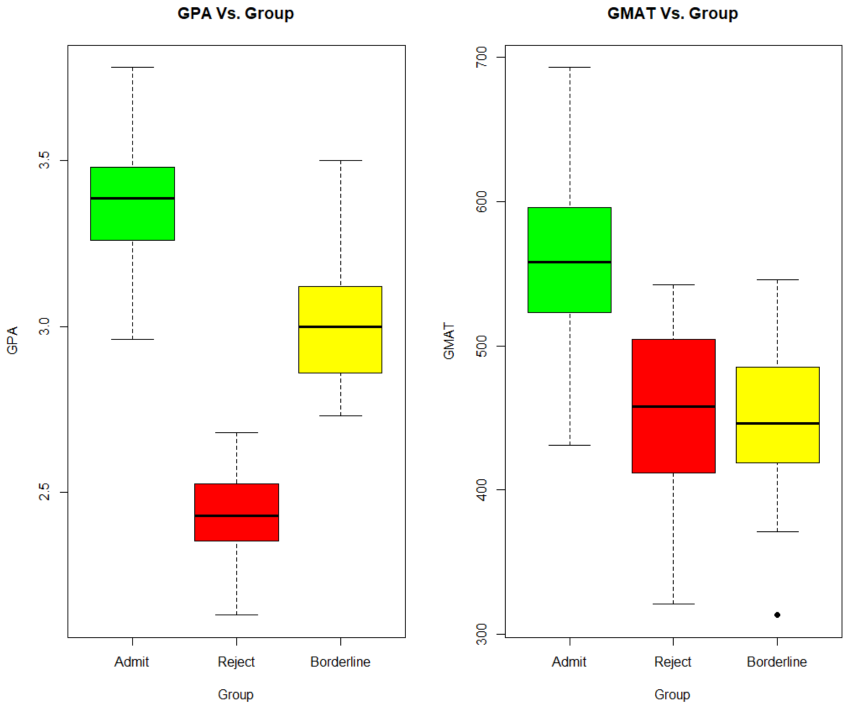

Figure 2: Boxplots for GPA Vs. Group and GMAT Vs. Group

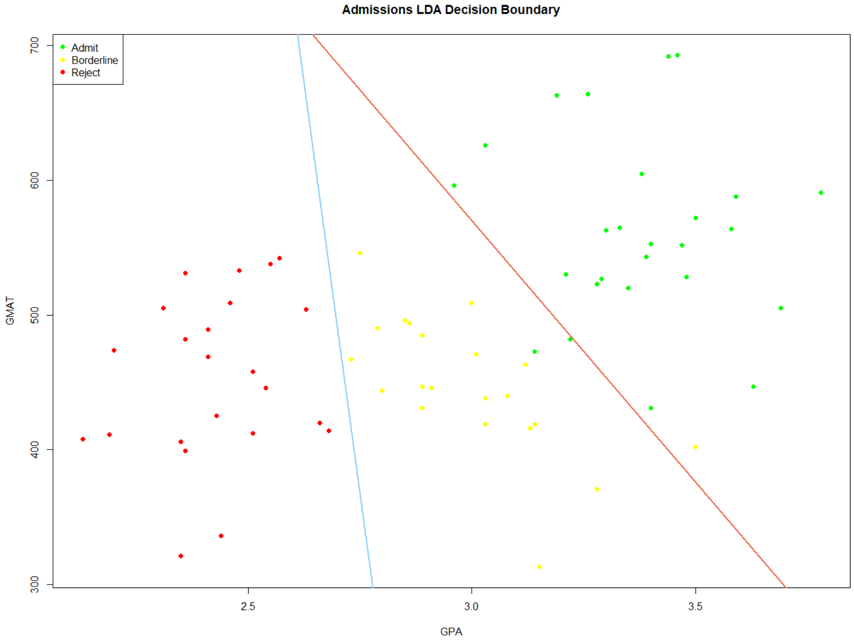

Figure 3: Decision Boundary for LDA

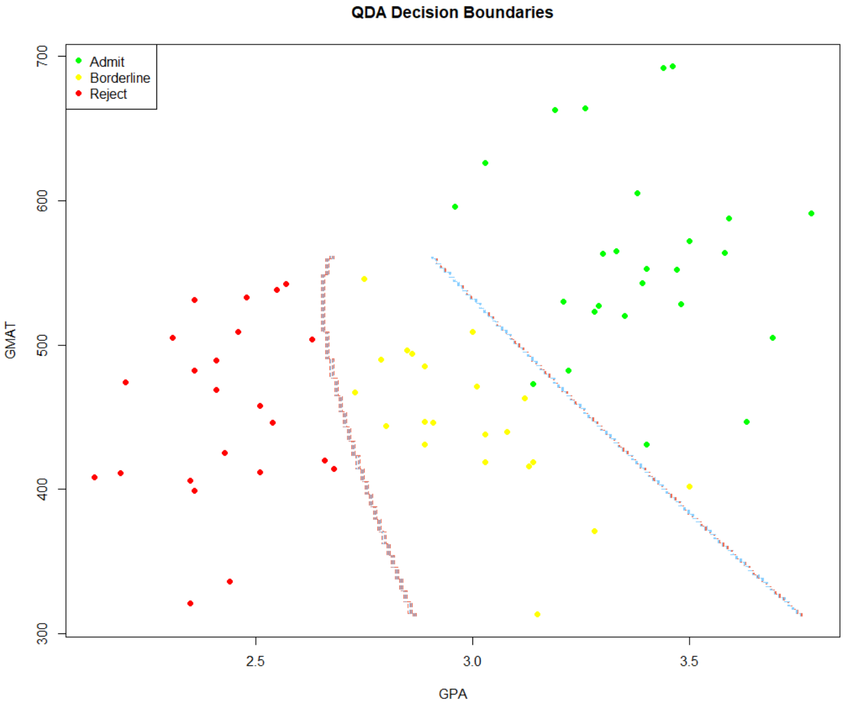

Figure 4: Decision Boundary for QDA

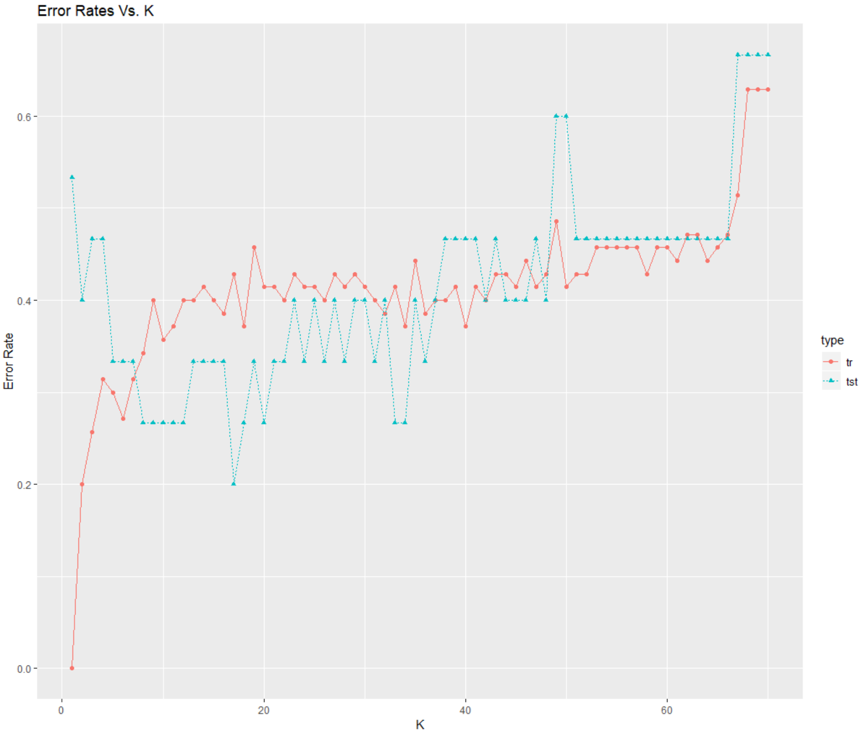

Figure 5: Error Rates Vs. K (KNN)

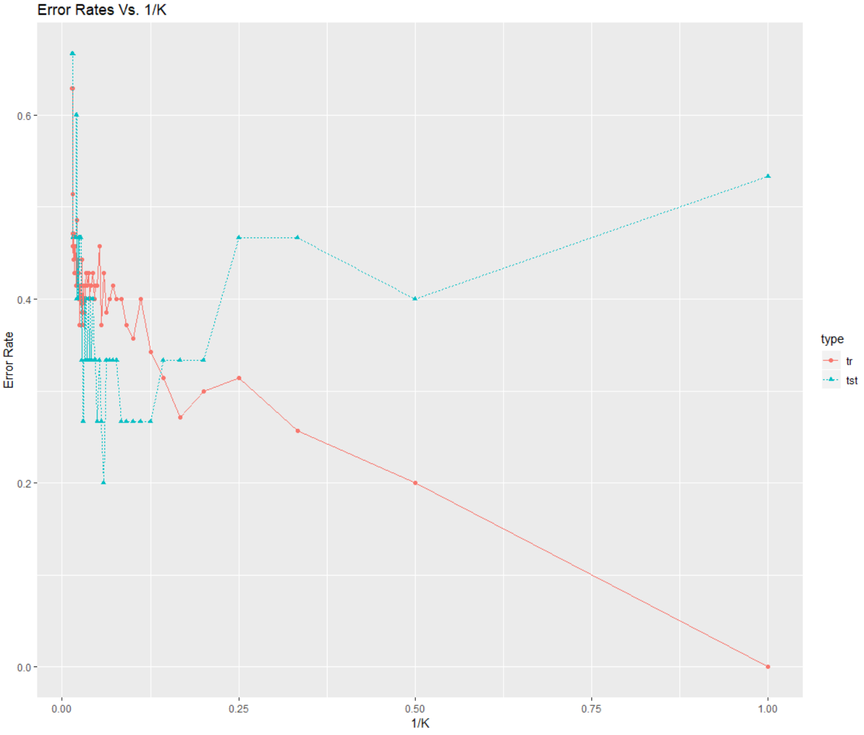

Figure 6: Error Rates Vs. 1/K (KNN)

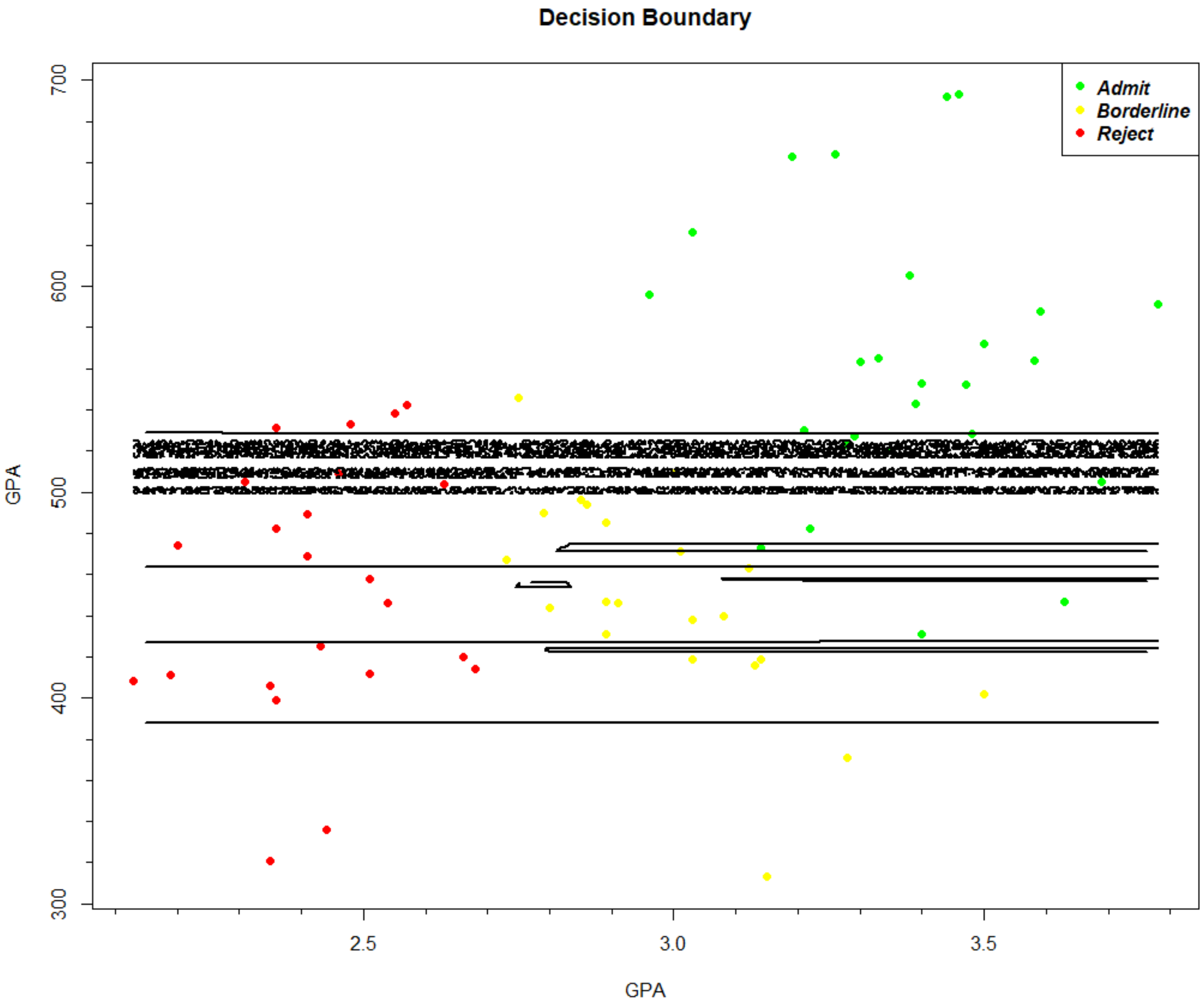

Figure 7: KNN Decision Boundary

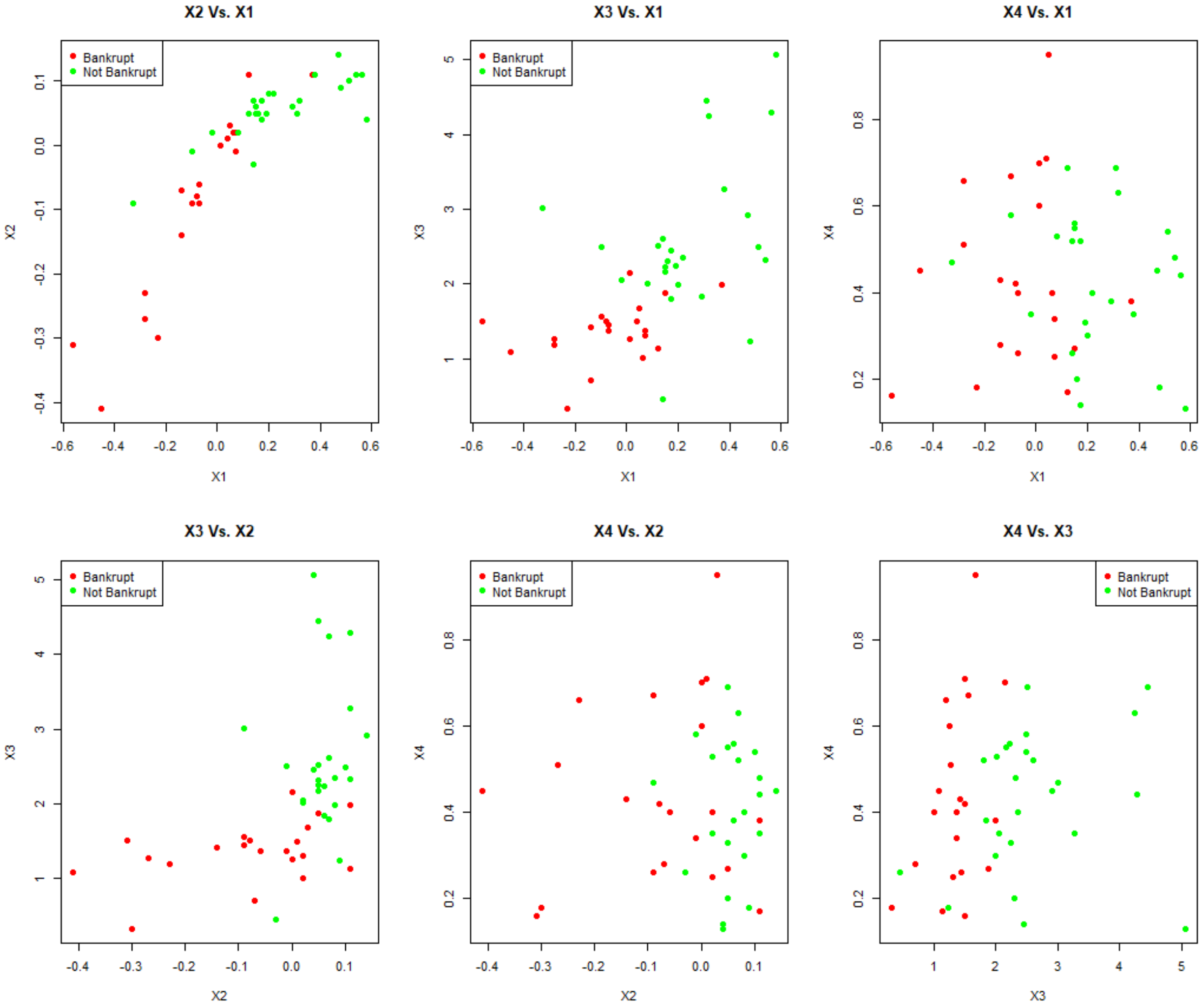

Figure 8: Question 2 Scatterplots (EDA)

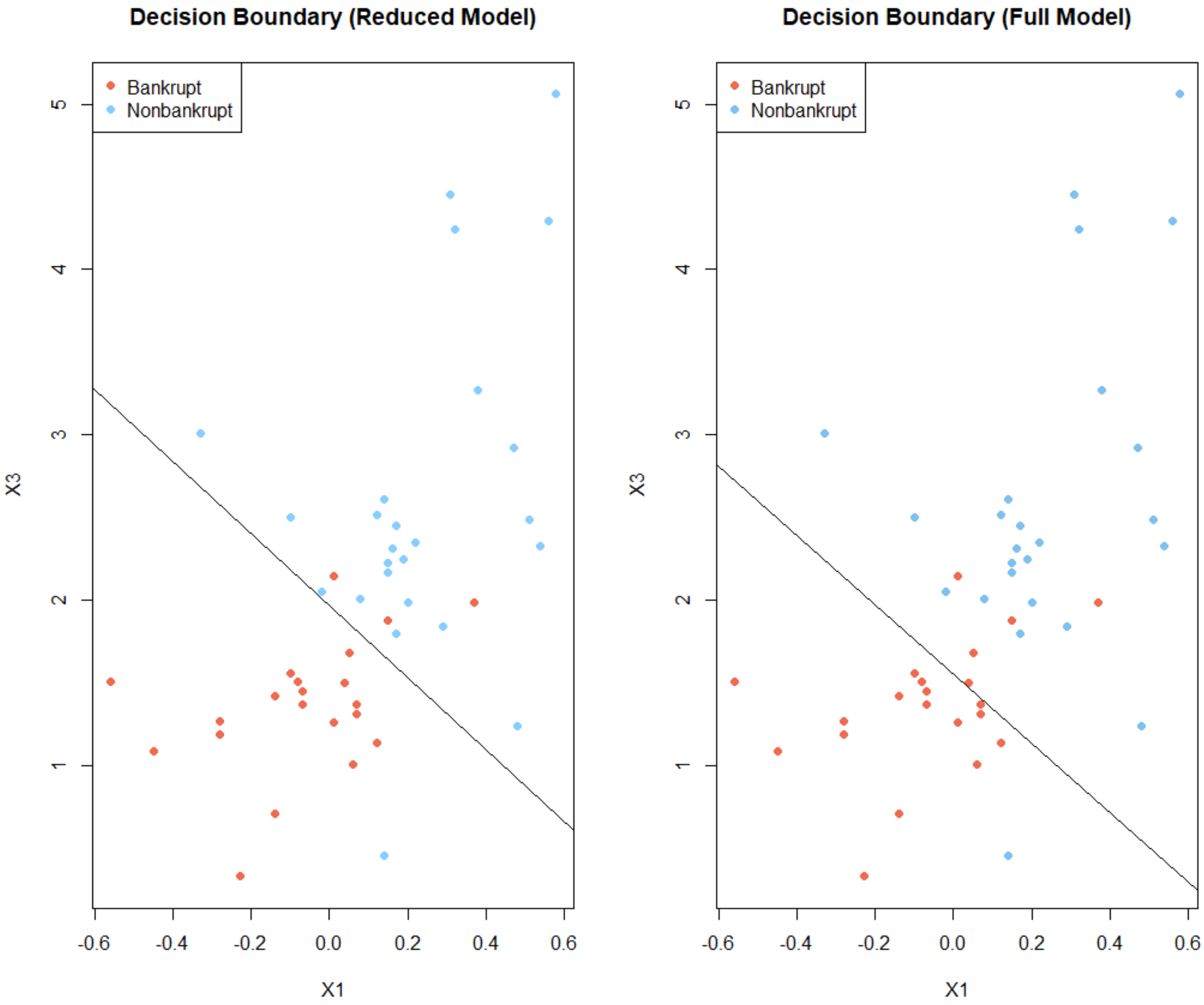

Figure 9: Decision Boundary for Reduced and Full Logistic Models

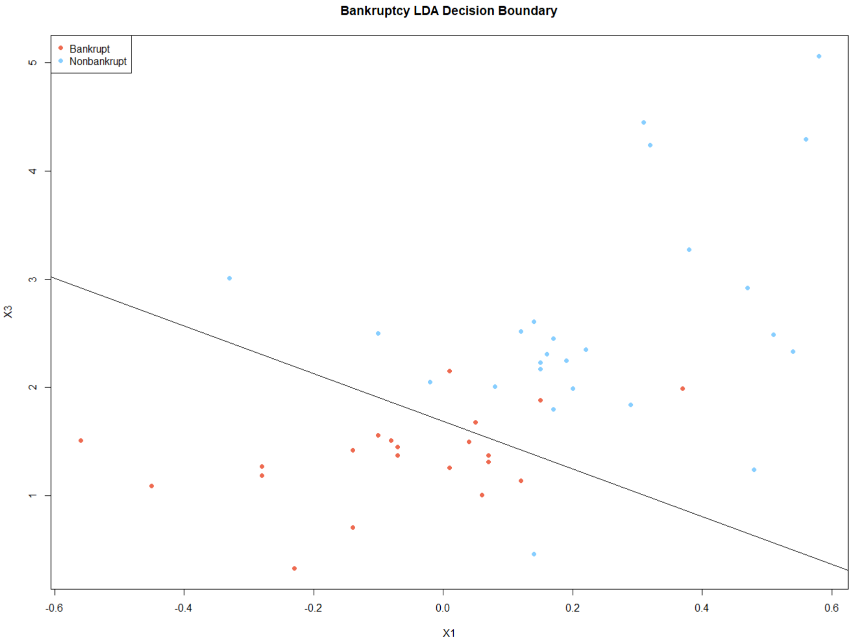

Figure 10: LDA Decision Boundary (Bankruptcy Dataset)

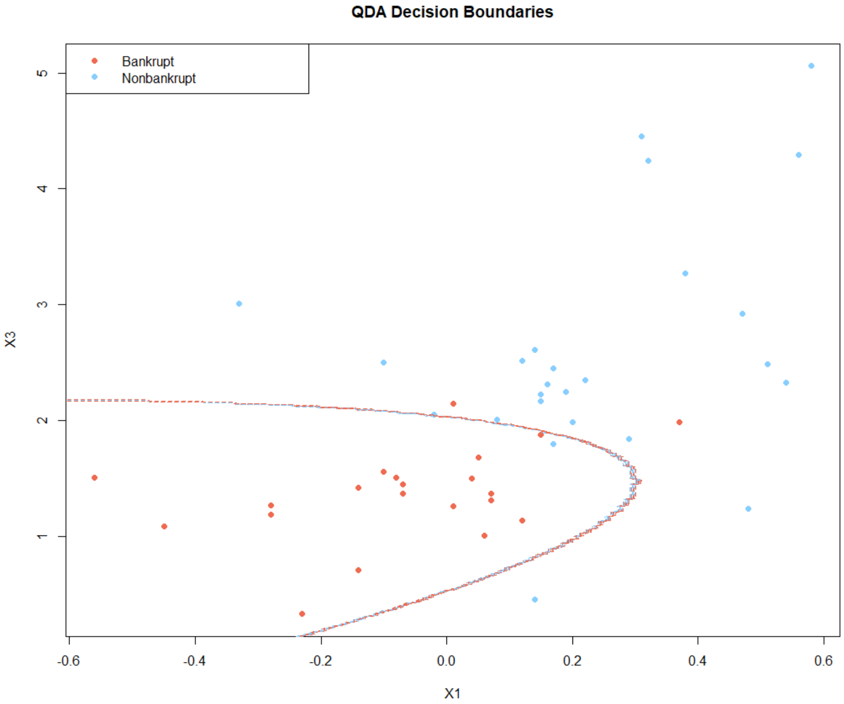

Figure 11: QDA Decision Boundary (Bankruptcy Dataset)

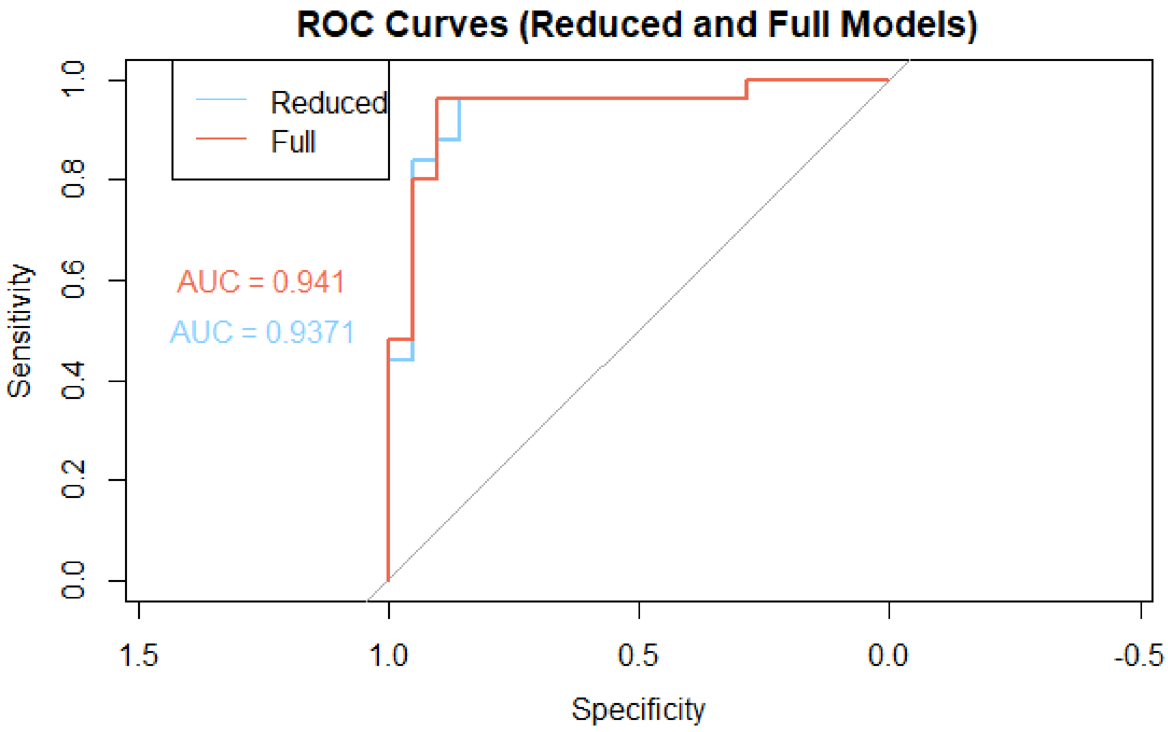

Figure 12: ROC Curves for Reduced and Full Models

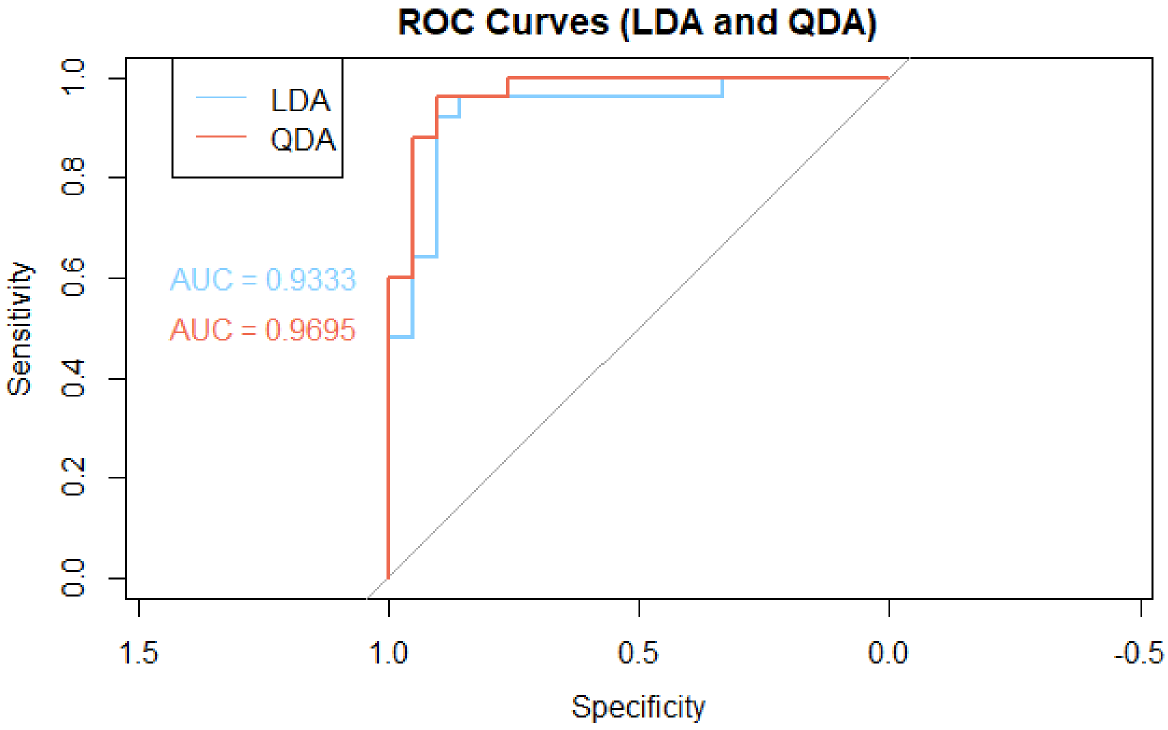

Figure 13: ROC Curve for LDA and QDA