APPM 4580 Homework 4

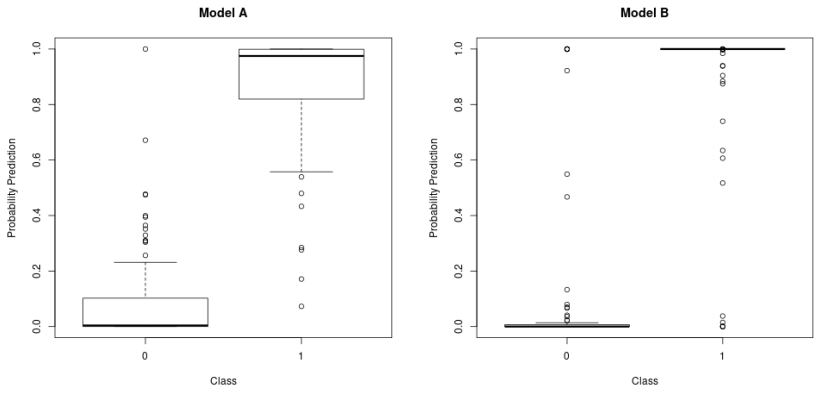
Michael Voecks, ID: 102584009

Problem 1

a)

For the Logistic Model (Model A) I tested bidirectional, forward, and backward stepwise selection and determined that the bidirectional stepwise selection technique was the best for this model because it had the best AIC score. For our data, the model found using the bidirectional stepwise selection technique produced the best model. For the Discriminant Analysis Model (Model B) I tested bidirectional and backward class stepwise selection techniques and looked at the accuracy score as my measure for which model worked best for both linear and quadratic discriminant analysis. It was found that the model produced from performing backwards class selection with quadratic discriminant analysis had the highest accuracy score, and so this was used for model B. This model also had fewer covariates than its LDA counterpart.

b)



As the boxplots above show, both models have some clear outliers that were produced from the misclassification of emails.

c)

The confusion matrix for both models are shown below

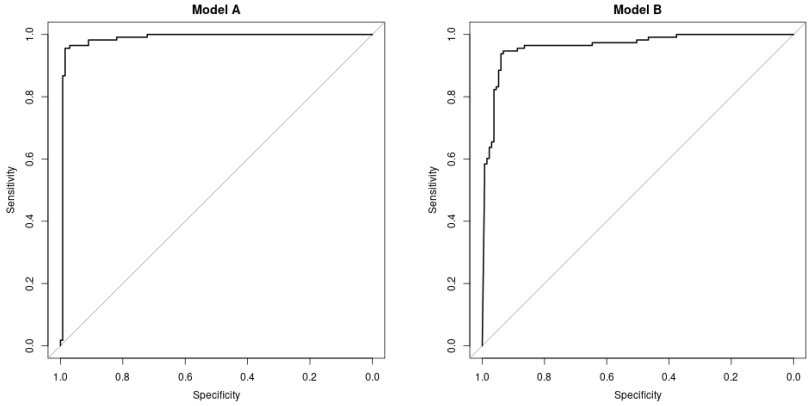
$$ModelA = \begin{bmatrix} 131 & 6 \\ 2 & 107 \end{bmatrix}$$

$$ModelB = \begin{bmatrix} 124 & 6 \\ 9 & 107 \end{bmatrix}$$

Both models have the same sensitivity with a value of 0.9469027, however, Model A has a higher specificity at 0.9849624 than Model B's 0.9323308, therefore I would prefer to use Model A.

d)

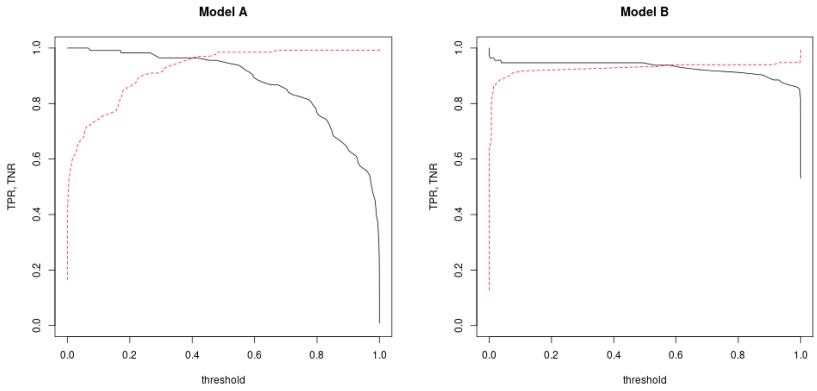
The ROC curves for each model are shown below.



The AUC for Model A is: 0.9864
The AUC for Model B is: 0.9627
It seems that Model A performs better overall than Model B since its AUC is larger. In addition to this, a visual comparison of the two ROC curves shows that Model A has a much better True Positive Rate and

e)

Generated below are plots of the True Positive and Negative Rates for both models as the threshold varies.



Based of the intersection of the curves on the graphs, it was decided that both models needed updated thresholds that will be different from one another, specifically Model A will have a threshold of 0.4 and Model B will have a threshold of 0.6 so that the True Positive Rate and the True Negative Rate match each other.

f)

The confusion matrix for the updated models are:

$$ModelA = \begin{bmatrix} 129 & 4 \\ 4 & 109 \end{bmatrix}$$

$$ModelB = \begin{bmatrix} 125 & 7 \\ 8 & 106 \end{bmatrix}$$

Model A has a sensitivity of 0.9646 and a specificity of 0.9699. Model B has a sensitivity of 0.9381 and a specificity of 0.9398.

g)

Model A has a Brier score of 0.03861887 while Model B has a Brier score of 0.06071819.