

STAT 473 Group Project

Credit Default Risk Assessment

**Submitted to:**

Dr. Xiyue Liao

**Submitted by:**

Julja Eco

Nate Talampas

May 12, 2023

**I. Introduction**

Credit default risk is the risk a lender takes that a borrower will not make the required payments on a debt obligation, which can lead to a potential loss for a lender. Earlier credit and risk management analysis would be conducted by analyzing the borrower's credentials and capabilities, which was more prone to error. Machine learning algorithms are more efficient in performing credit risk assessments with better precision and at faster speeds. We compared different machine learning algorithms and their performance metrics. To do this, we needed to identify which predictors to include and omit in each model. The dataset included 32,581 observations and 12 variables. The featured variables were age, annual income, home ownership, employment length (in years), loan intent, loan grade, loan amount, interest rate, loan status, loan to income percent, historical default, and credit history length. We found that while the Classification Tree model has the best accuracy, the Ridge Regression model has the highest AUC value. Classification Tree should be applied to low-risk clients or low loan amounts, while Ridge Regression should be applied to high-risk clients or high loan amounts.

**II. Questions of Interest**

Our questions of interest after looking at the dataset were:

* What variables are most significant in predicting credit default risk?
* How do different machine learning algorithms perform in predicting credit default risk?

**III. Analysis**

We used Logistic Regression to determine what variables are most significant in predicting credit default. We used Logistic Regression, Linear Discriminant Analysis, Ridge Regression, and Classification Trees to estimate whether an individual will default on their loan. Performance was evaluated using metrics like Accuracy and Area Under the Curve for the ROC Curve. The data contained missing values. Because of this the data was tidied— rows with the missing values were removed across the different variables. Any character variables were also changed into factor variables. The summary statistics were computed. Loan status is a binary variable that determines whether an individual defaults on their credit loan or not. Annual income, employment length, loan amount, interest rate, and loan to income percent are numerical variables. Home ownership, loan intent, and loan grade are character variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **loan\_status**  Min. :0.0000  1st Qu.:0.0000  Median: 0.0000  Mean: 0.2166  3rd Qu: 0.0000  Max: 1.0000 | **person\_income**  Min: 4000  1st Qu: 39474  Median: 55910  Mean: 66645  3rd Qu: 80000  Max: 6000000 | **person\_home\_ownership**  Length: 38636  Class: character  Mode: character | **Person\_emp\_length**  Min: 0.00  1st Qu: 2.00  Median: 4.00  Mean: 4.78  3rd Qu: 7.00  Max: 41.00 | **Loan\_intent**  Length: 38636  Class: character  Mode: character |
| **Loan\_grade**  Length: 38636  Class: character  Mode: character | **Loan\_amnt**  Min: 500  1st Qu: 5000  Median: 8000  Mean: 9655  3rd Qu: 12500  Max: 35000 | **Loan\_int\_rate**  Min: 5.42  1st Qu: 7.90  Median: 10.99  Mean: 11.04  3rd Qu: 13.48  Max: 23.22 | **loan\_percent\_income**  Min. :0.0000  1st Qu.:0.0900  Median: 0.1500  Mean: 0.1695  3rd Qu: 0.2300  Max: 0.8300 |  |

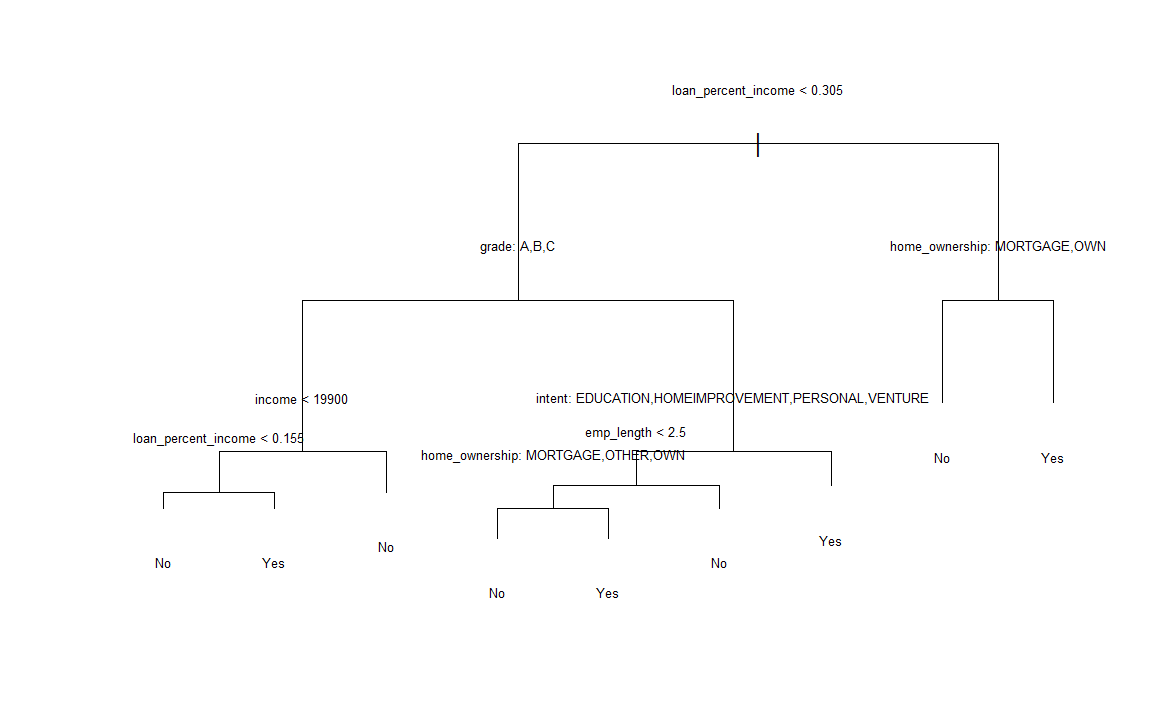
We determined that the data is imbalanced, so the ROC curve would be a significant performance metric. We split our data into 50% test set and 50% training set. We fit a logistic regression using the training set and predict the loan status outcome using the test set. The accuracy of the model is 69.51%, meaning for every 100 predictions the model makes about 69 correct predictions. The area under the ROC curve is 0.5103.

We use a similar technique to fit a linear discriminant analysis model. We train the model using the training set and obtain the prior probabilities of each outcome. The prior probability of no default is 0.782 and the probability of default is 0.218. The accuracy of the model is 86.71%, meaning for every 100 predictions the model makes about 86 correct predictions. The area under the ROC curve is 0.8669.

Additionally for the purpose of the classification tree method, a new categorical variable statusC was created based from the binary variable loan\_status wherein 1 was “Yes” and 0 was “No” and the loan\_status variable was dropped. Once the data was clean, we checked if the data was imbalanced. Using loan\_status as our response variable, the dataset was imbalanced because 78.34% of individuals did not default. Because of this we split our data into 50% test set and 50% train set.

Next, we use the shrinking method of Ridge Regression that helps improve the performance of the logistic regression model by shrinking the coefficient estimates as close to 0 as possible. By using this method, we hope to avoid overfitting the model as well as address any multicollinearity of the dataset. To build the ridge regression model, we first standardized the predictors using the scale function. Then the glmnet function is used to build the model with alpha set to 0 and the family set to binomial. Then we perform cross-validation to select the tuning parameter or best lambda in which the error is the smallest, giving us the best results. The model is re-fit again with the additional argument lambda set to the best lambda from the cross-validation results. The accuracy of the model is 86.63% meaning for every 100 predictions, the model will make about 86 correct predictions. Then we construct the ROC curve to calculate the AUC value which is 0.8716, thus the model performs better than the logistic regression model.

We also use the tree-based method Classification Tree. This is a different classification method than LDA since this algorithm selects the input feature that best separates the training data at each node, and then recursively splits the data based on that feature. To build the tree, the tree function is used with statusC set as the response variable against all the other variables. The unpruned tree has 9 terminal nodes. Then we perform cross-validation again using the cv.tree function to see if the tree needs to be pruned. Since 9 has the lowest cross validation error rate, we do not need to prune the tree. Them, we plot a visual of the tree for interpretation as shown below:



The root node is the loan to income proportion meaning that this is the most important factor in predicting whether or not a customer will default on their loan. A low loan to income proportion means your loan is a small percentage of your income, which makes it easier for borrower to pay back and the lender to get their money back on time. If your loan to income percentage is greater than 0.305 and you either mortgage or own a house, then you are more likely to not default and not a risk. This makes sense because people who have houses or pay mortgage pay more because of their home ownership, so they have larger loans than their income. Another interpretation is if your loan to income percentage is less than 0.305 and you have a loan grade of A, B, or C, then you are more likely to be not a risk depending on your income and loan to income percentage. Another interpretation is that if you have a loan grade of anything below C, and your intent for the loan is debt consolidation or medical-related, then you are more likely to be a risk.

The accuracy of the model is 92.22% while the AUC value is 0.8237. While the model has very high accuracy, it's AUC value is lower than the Ridge Regression model and the LDA model. This means that the model makes correct predictions for most of the observations but does not do as good of a job as the 2 previous models in distinguishing between the positive (did not default) and negative (default) classes. We calculate the true positive rate (TPR) which is 99.56%, the false positive rate (FPR) which is 34.83%. The FPR is high considering the cost of the misclassification of people who would default into people who would not default.

IV. Conclusion (1 - 2 paragraphs)

After comparing the accuracy and AUC values of each respective model, the Ridge Regression model has the highest AUC value, while the Classification Tree model has the highest accuracy. Meaning, a classification tree is a good model to use to make correct predictions, but not to classify positive and negative classes. For example, if the model classifies a risky person with a high loan amount who is more likely to default on a loan into the non-default class, then that would be a loss to the lender. Thus, this model would be good if the majority of the clients are low-risk or have low loans. The ridge regression model would perform better in distinguishing positive and false classes given its high AUC value. This model would be the safer option when classifying risk of defaulting.

V. Appendix

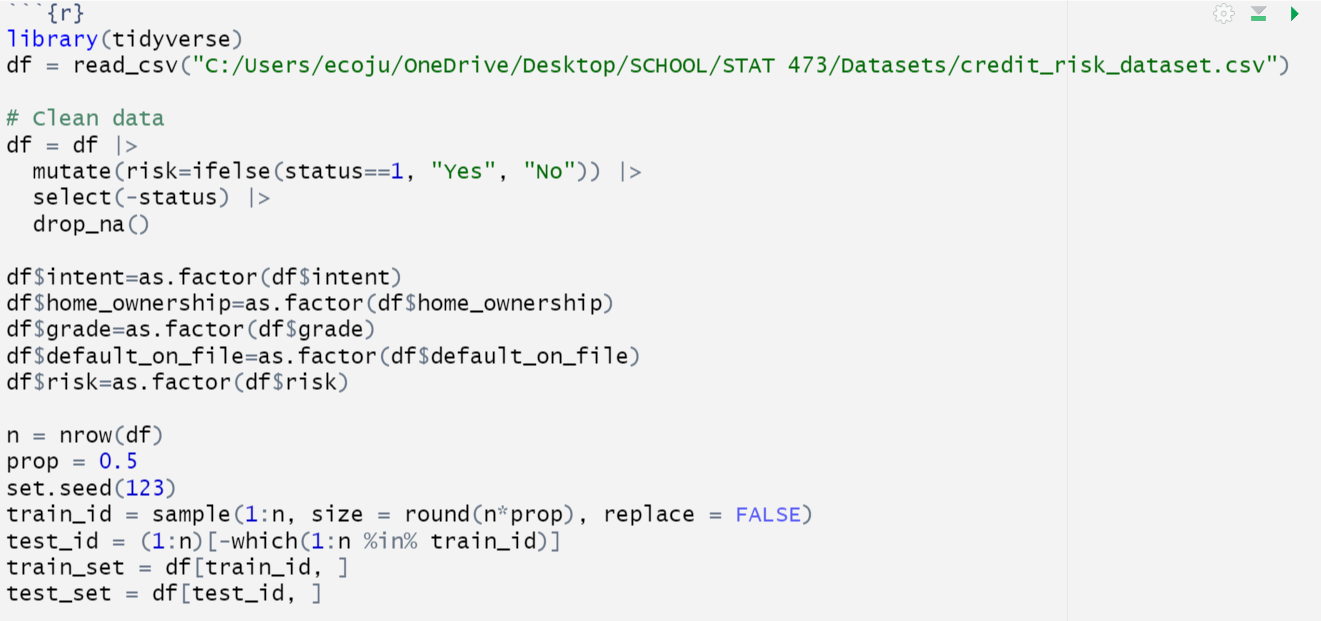


Figure 1: R code to clean data

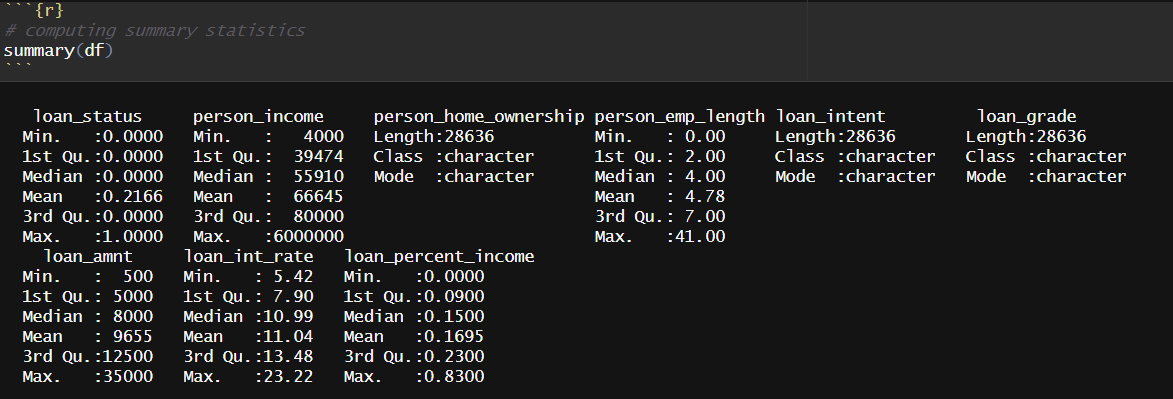


Figure 2: R code and output to compute summary statistics

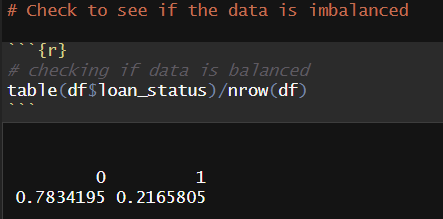


Figure 3: R code and output to determine if the data is imbalanced

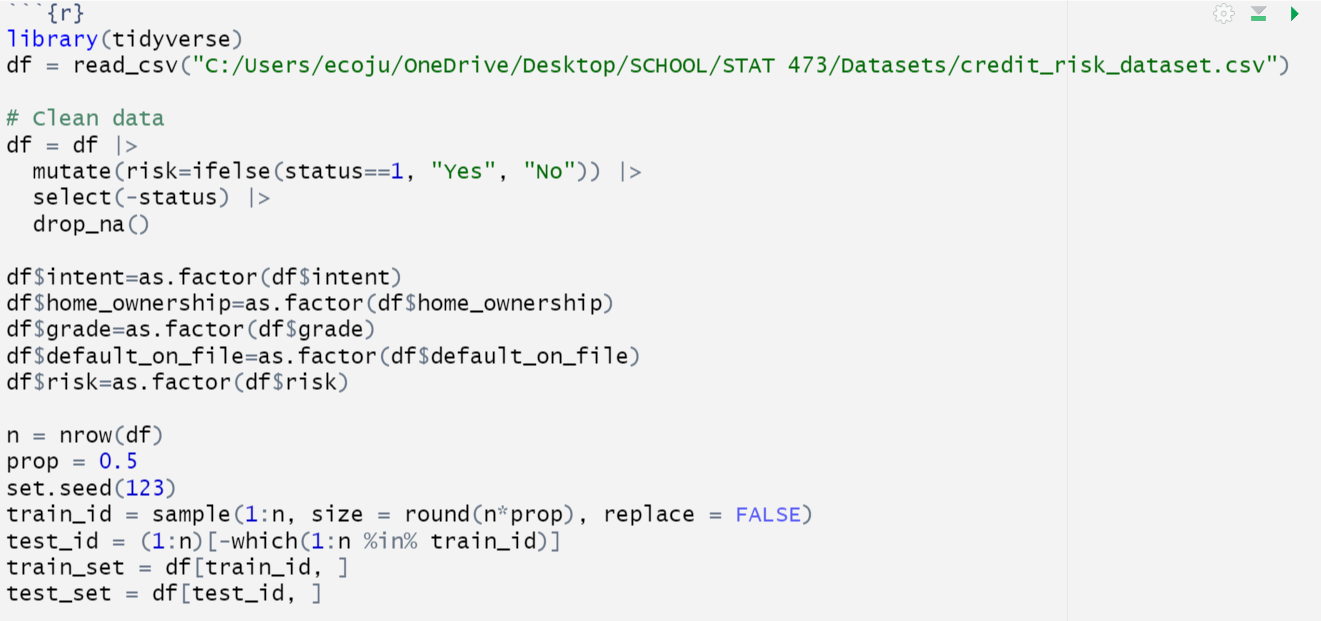


Figure 4: R code to split the data into test set (50%) and training set (50%)

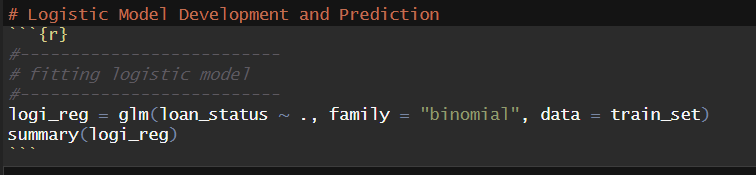
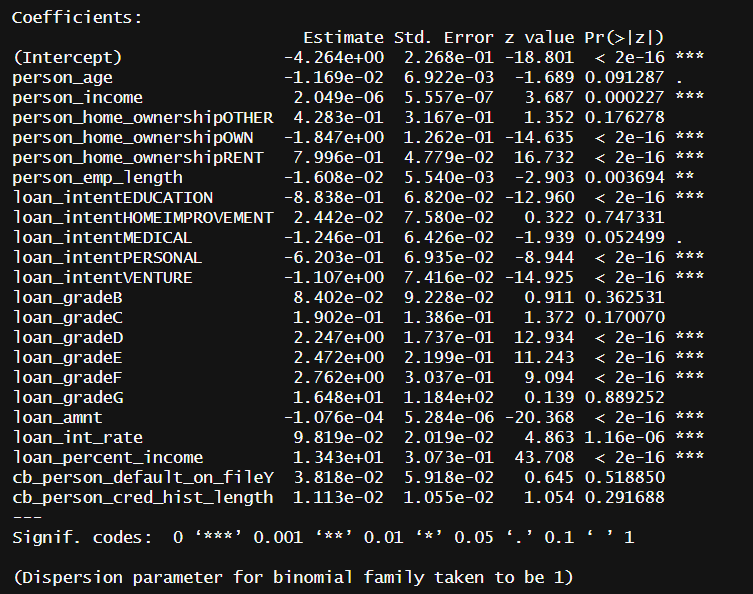
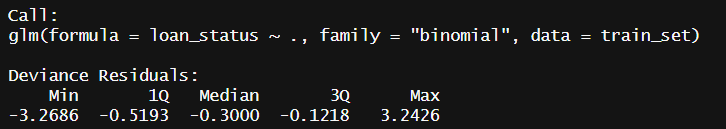


Figure 5: R code to fit a logistic model with the training set



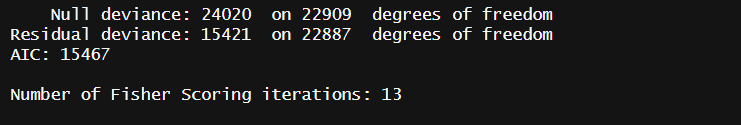


Figure 6: Output of logistic model

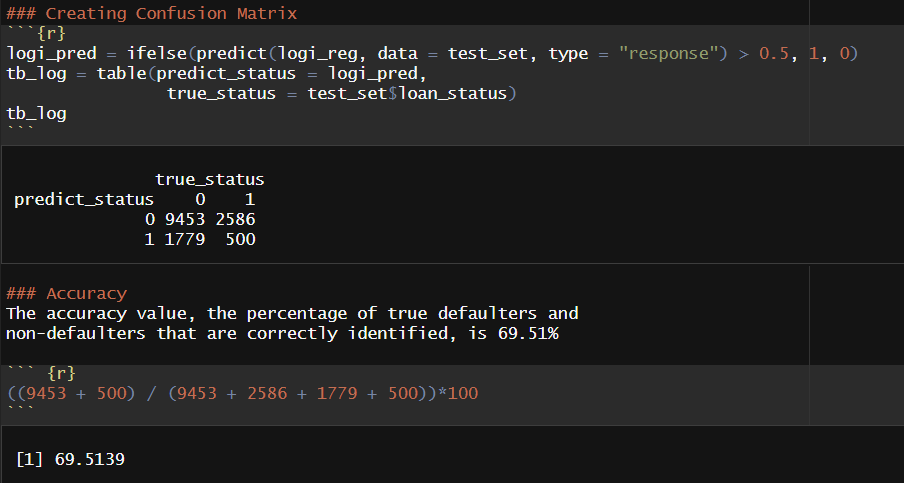


Figure 7: R code and output for computing confusion matrix and accuracy

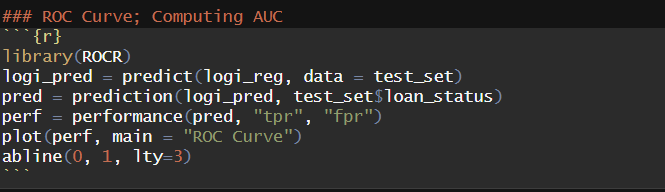


Figure 8: R code to create ROC Curve

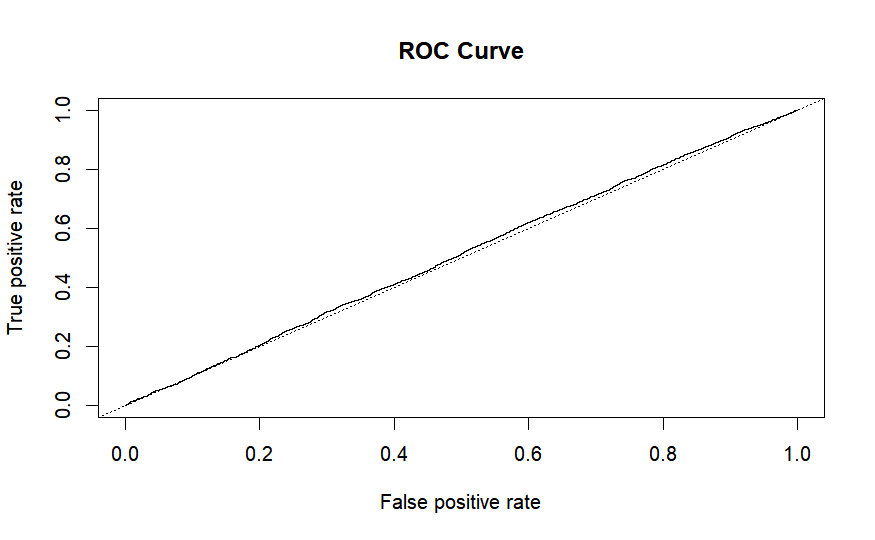


Figure 9: ROC Curve

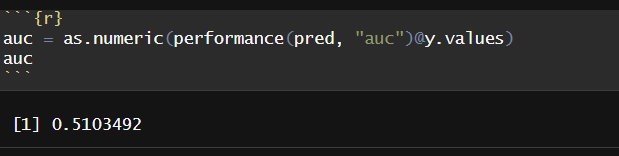


Figure 10: R code and output of area under the curve (AUC) value

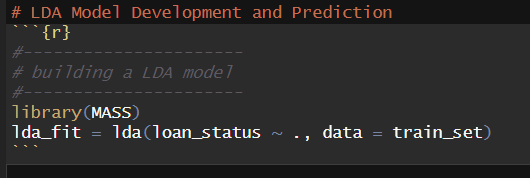
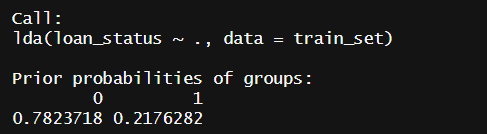
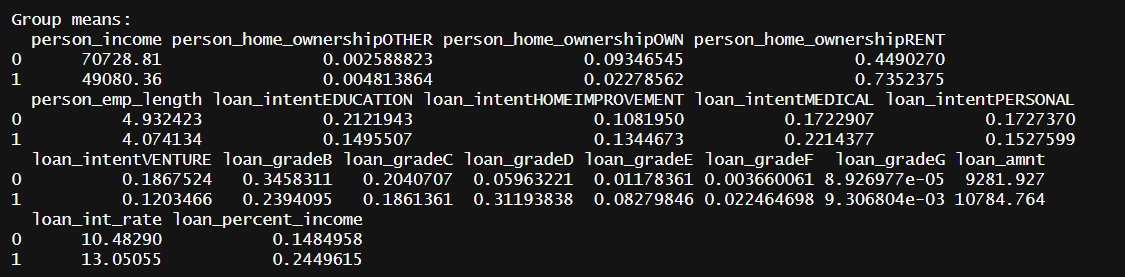


Figure 11: R code to fit the training set to a linear discriminant analysis model





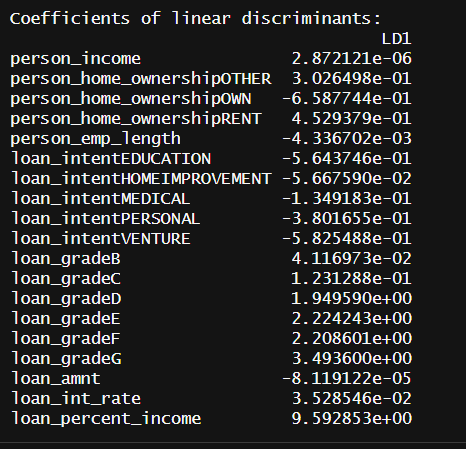


Figure 12: Output of LDA model

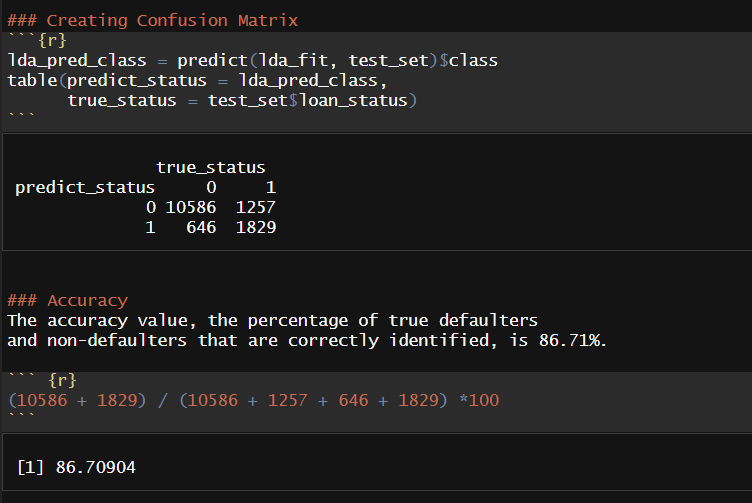


Figure 13: Confusion matrix and accuracy value of LDA model

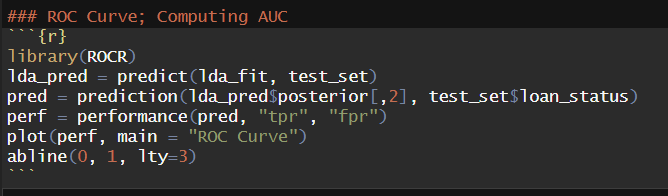


Figure 14: R Code of ROC Curve

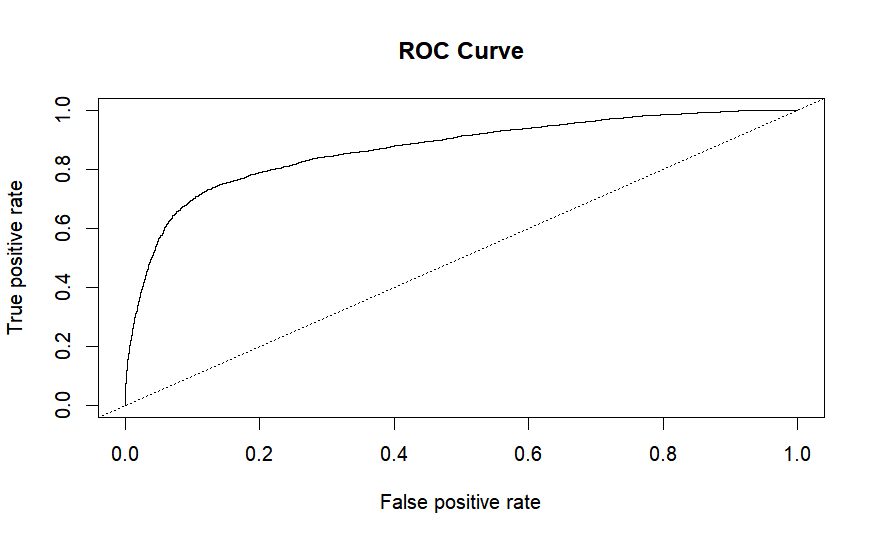


Figure 15: ROC curve of LDA model

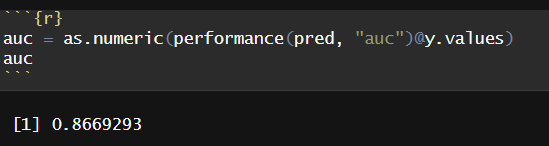


Figure 16: R code and output to compute area under the curve of LDA model



Figure 17: R code of ridge regression model

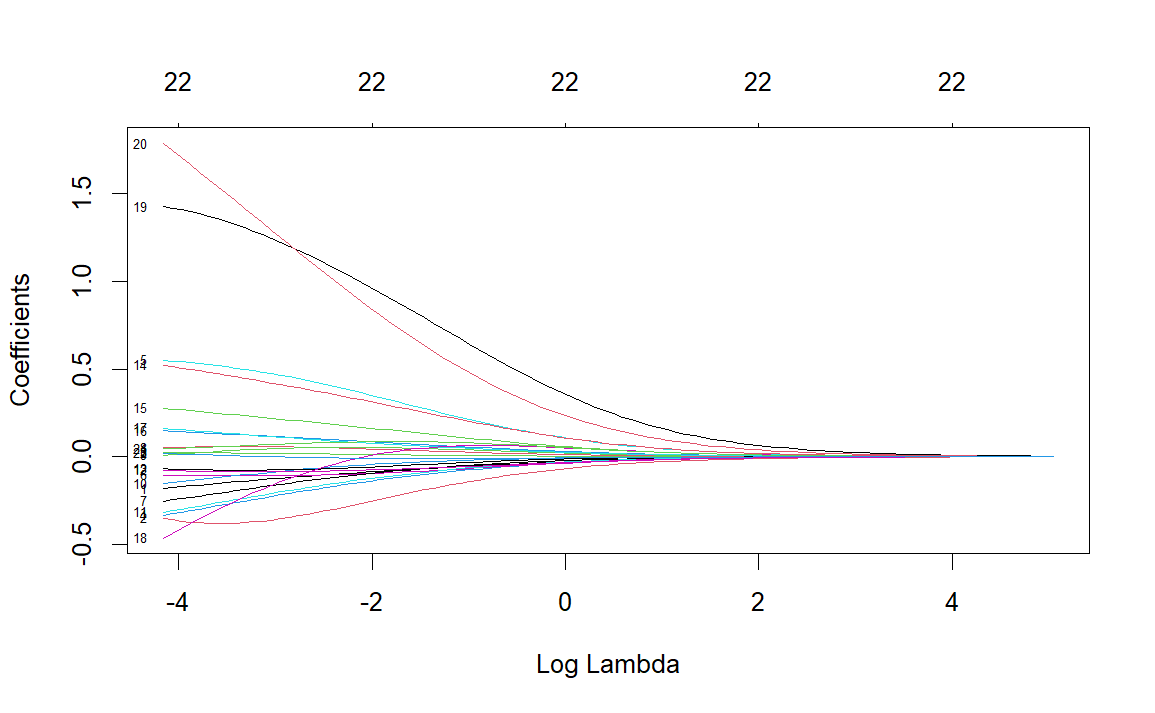


Figure 18: Plot of Lambda and Coefficients

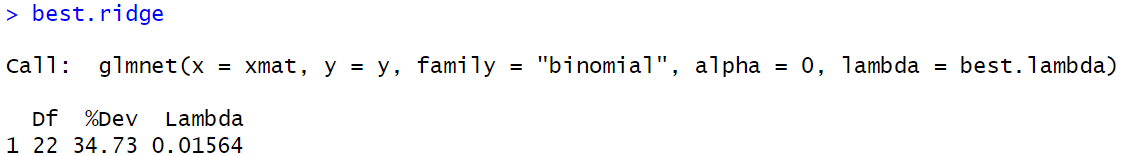


Figure 19: Ridge Regression fit with best lambda

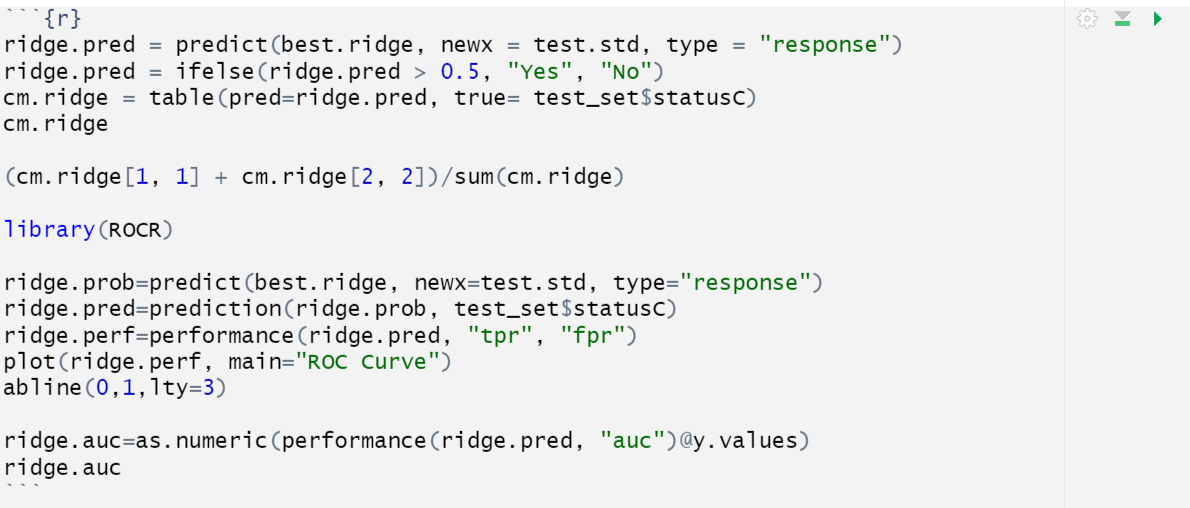


Figure 20: R code to calculate accuracy, ROC, and AUC of ridge regression model

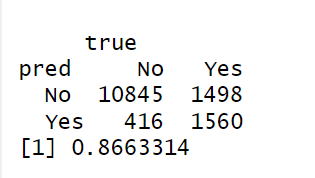


Figure 21: Confusion matrix and accuracy of ridge regression model

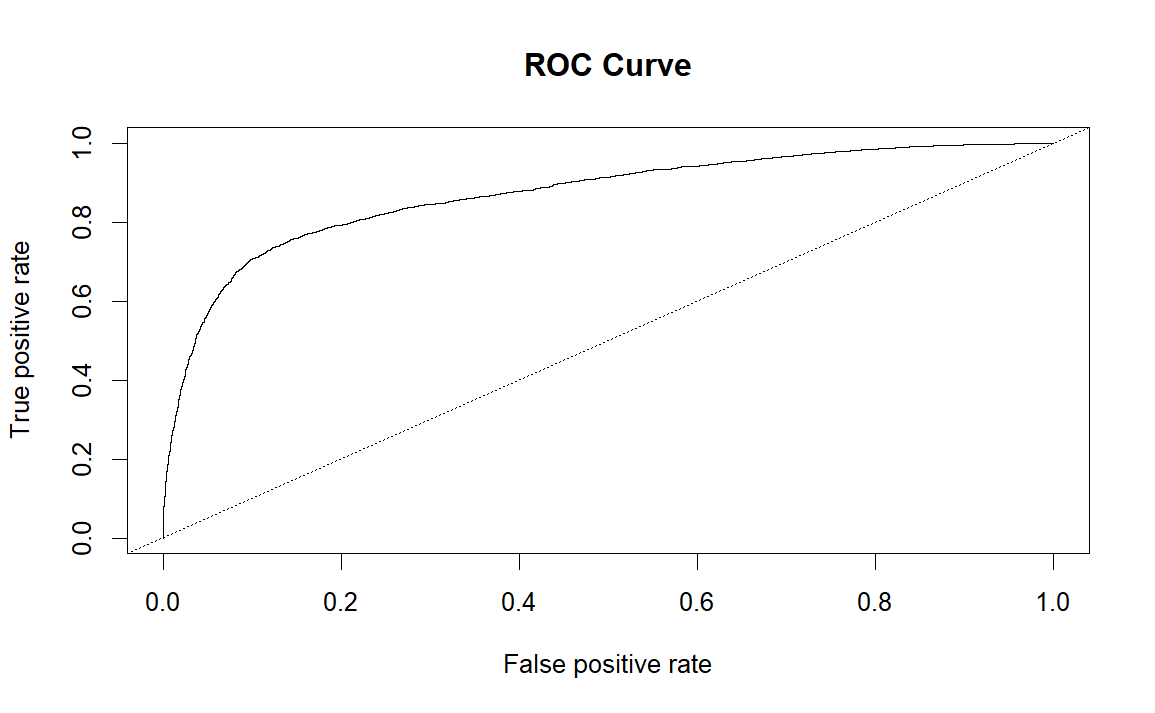


Figure 22: ROC Curve of ridge regression



Figure 23: AUC value of ridge regression



Figure 24: R code to fit classification tree

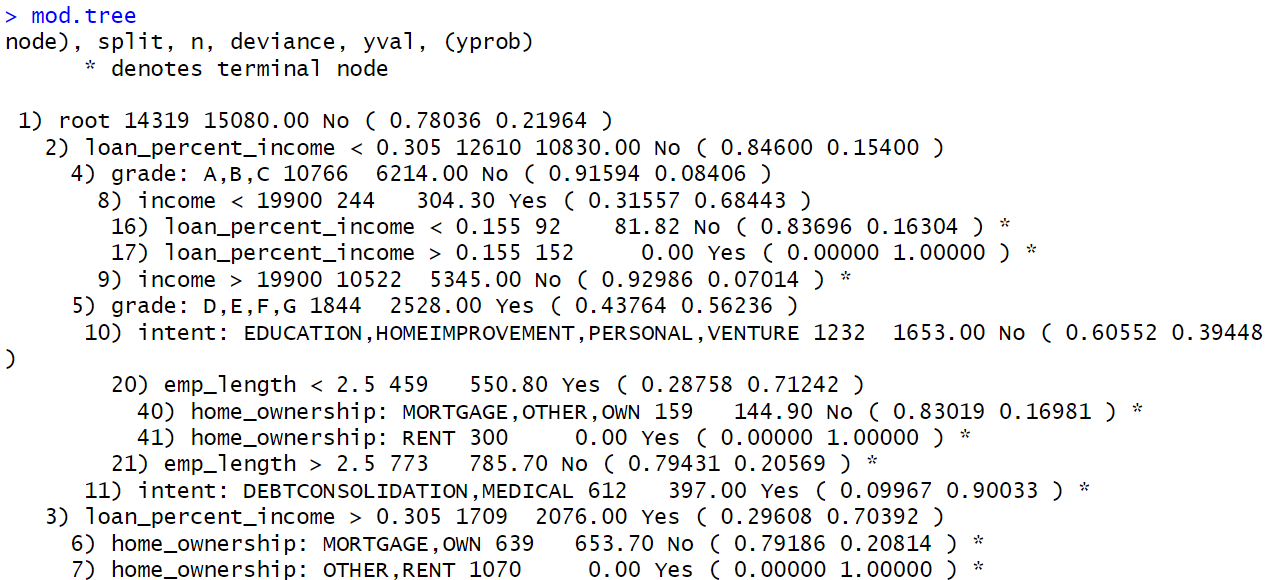


Figure 25: Classification tree output

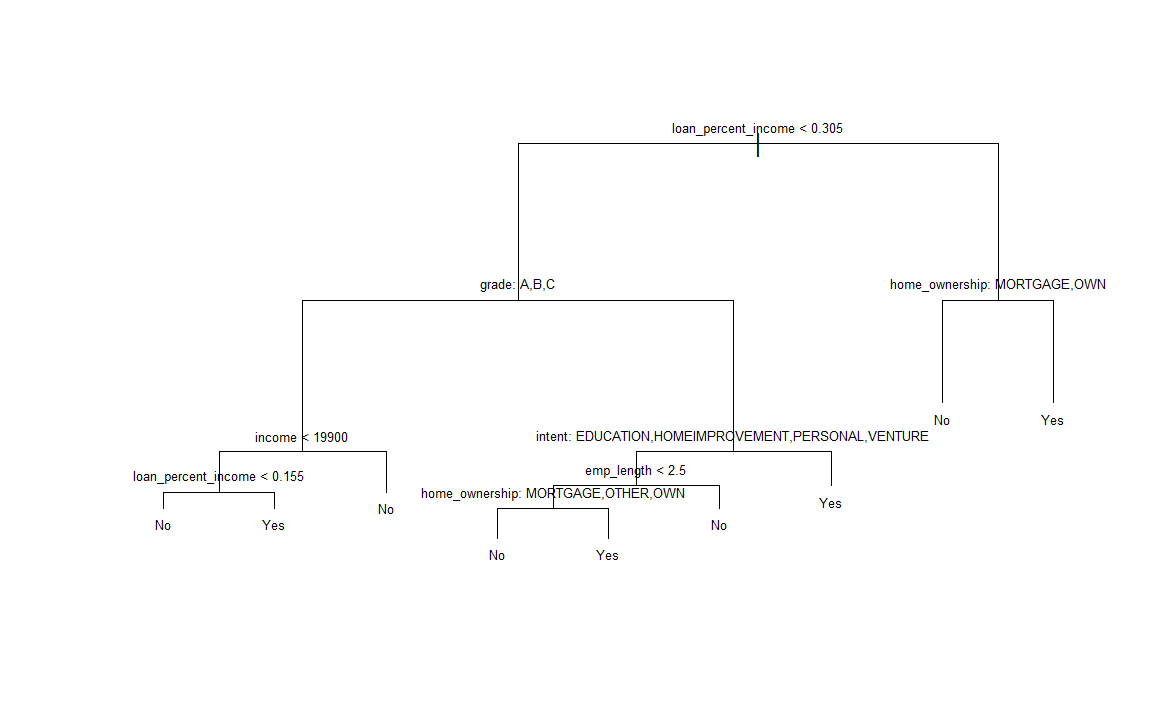


Figure 26: Visual plot of tree



Figure 27: R code for confusion matrix, accuracy, ROC Curve, and AUC value

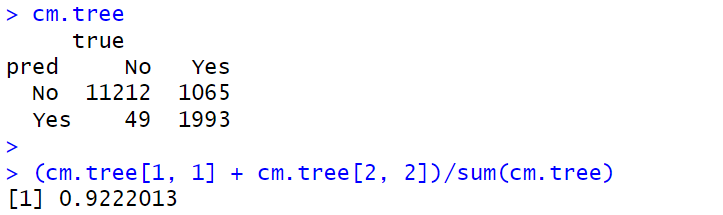


Figure 28: Confusion matrix and accuracy of classification tree

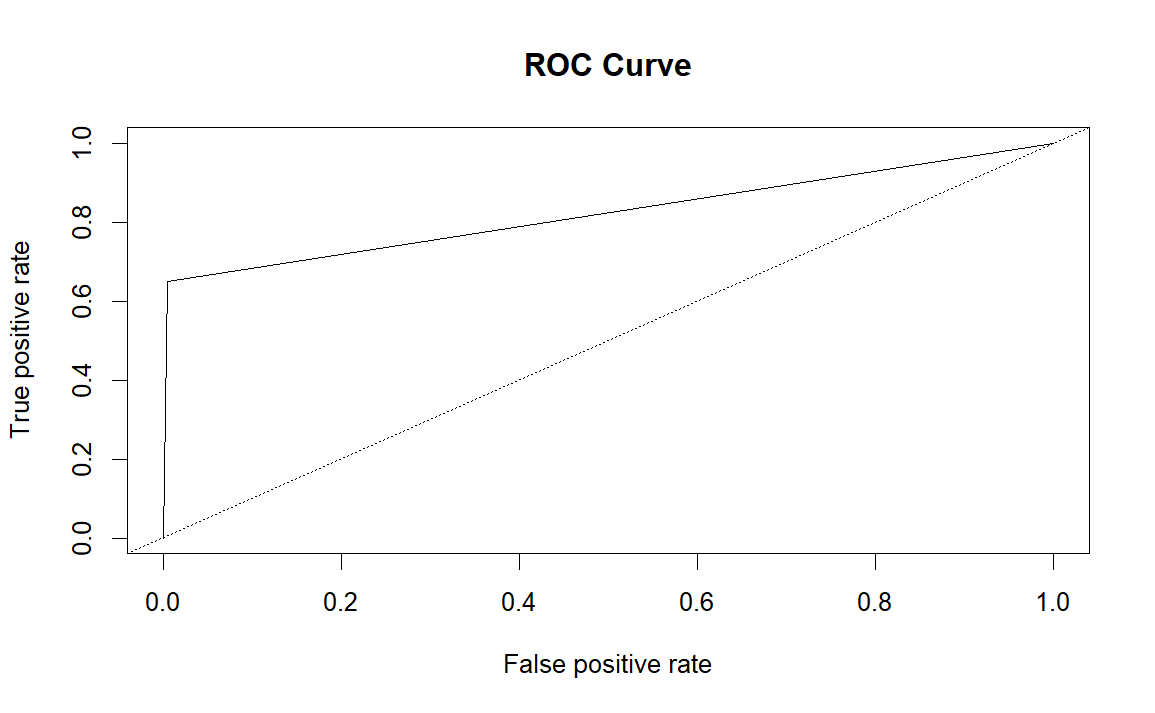


Figure 29: ROC Curve of classification tree

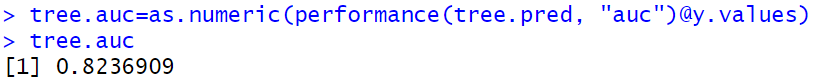


Figure 30: AUC value of classification tree



Figure 31: R code of true positive rate and false positive rate of classification tree

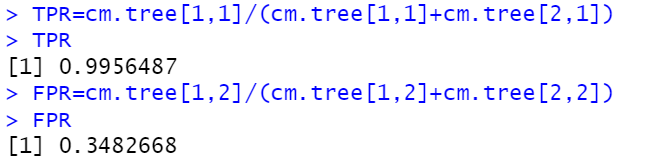


Figure 32: TPR and FPR of classification tree