**Credit Card Default**

Bryan Butto

March 1, 2022

**Executive Summary**

Credit card default can be extremely costly to lenders. It is very important to be able to determine which customers are more likely to default on their credit cards, and those who are less likely to default on their credit cards. This can save lenders valuable time and money. A lender does not want to deny customers a credit card who are likely to not default, and a lender does not want to accept customers who are likely to default. The goal of this project is to be able to predict whether a customer will default on their credit card. The conclusions of this project will be able to take data and accurately predict whether a customer will default on their credit card using specific variables or observations.

**Introduction**

The purpose of this project is to determine whether someone will default on their credit card. The data being used for this project is the Default dataset from the ISLR library in R. This data will be used to predict whether a customer will default on their credit card using the variables student, balance, and income.

**Problem Identification**

The research question posed by this project is will a customer default on their credit card. Does someone being a student have an impact on whether they will default on their credit card?

Does the average balance someone has on their credit card affect whether they will default?

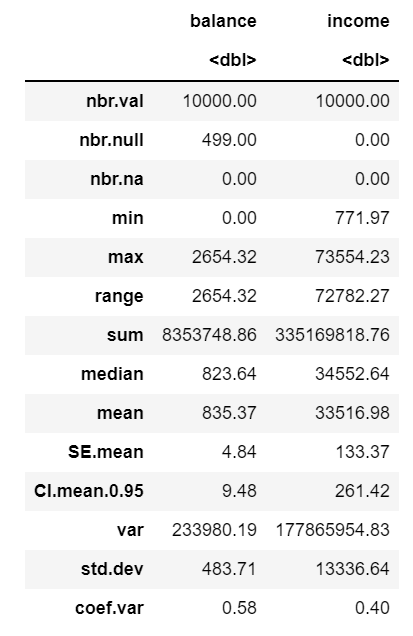
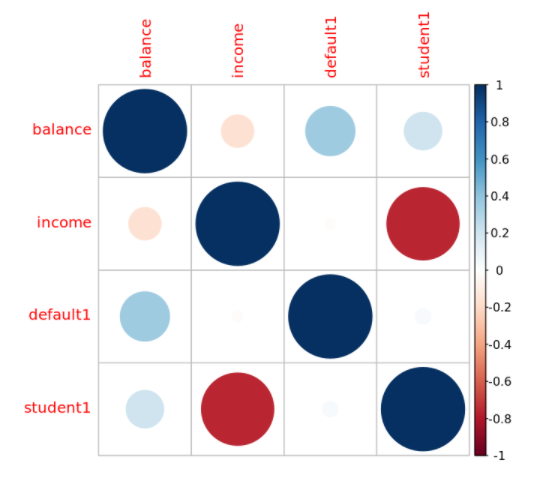
Does customer income impact whether they will default on their credit card? The null hypothesis is the variables student, balance, and income have no effect on whether someone will default on their credit card. The alternative hypothesis is one or more of the variables student, balance, and income will be a predictor of whether someone will default on their credit card.

**Data Preprocessing**

The Default dataset is composed of 10,000 observations of credit card data. The response variable is default which is a categorical variable with two factors, which are “yes” or “no”. There are 3 predictor variables which are student, balance, and income. Student is also a categorical variable with two factors, which are “yes” and “no”. Balance is a numerical variable and it is the average balance that the customer has remaining on their credit card after making their monthly payment. Income is a numerical variable and it is the customer’s income.

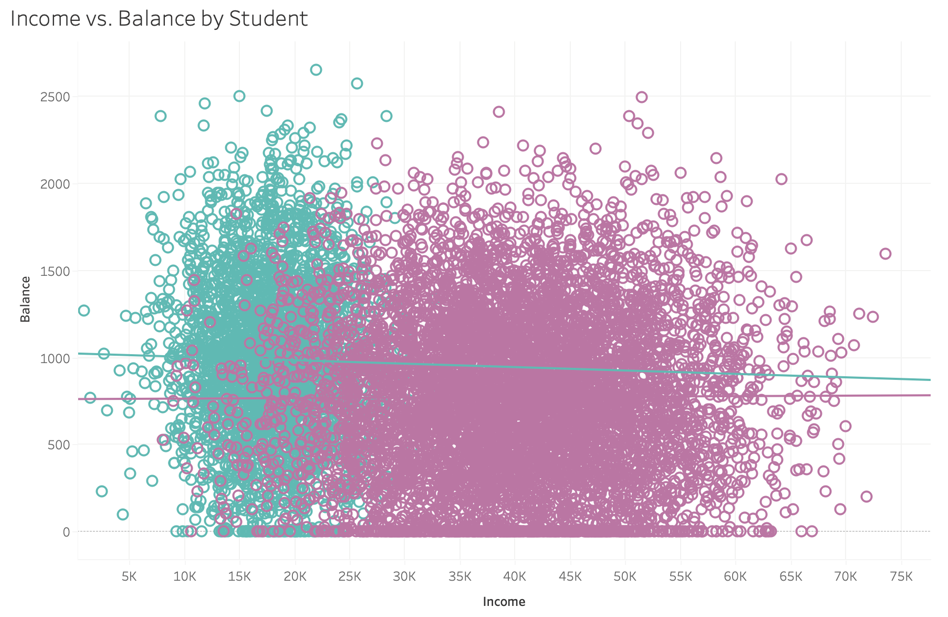
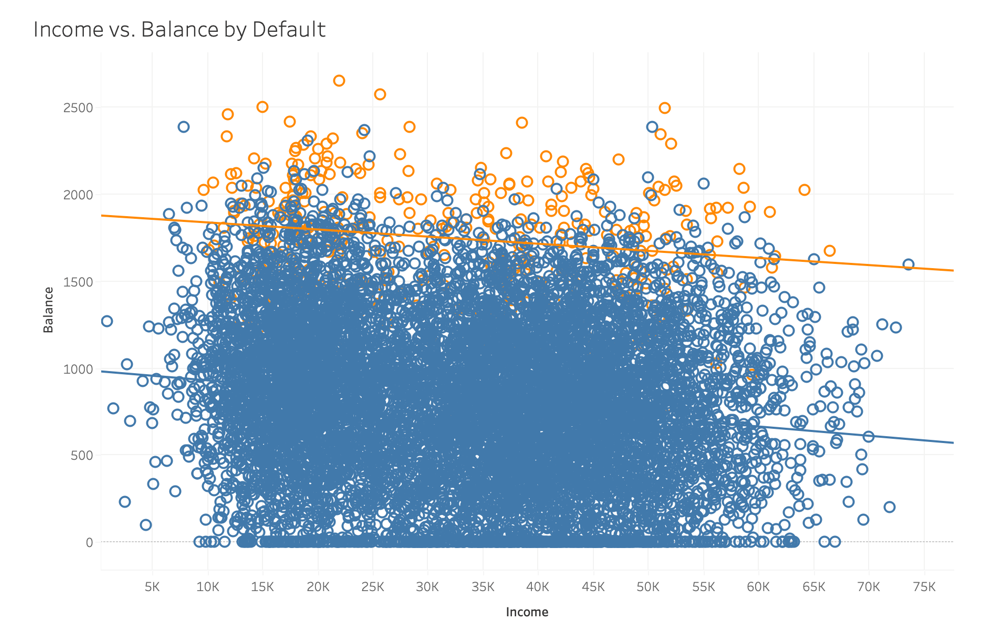
**Exploratory Data Analysis**

The researcher began by looking at Exploratory Data Analysis. For the default variable, there were 9,667 observations of customers that did not default on their credit card and 333 observations of customers that did default on their credit card. For the student variable, there were 7,056 observations of customers that are not students and 2,944 observations of customers that are students. The balance variable has a median value of 823.64, a mean value of 835.37, and a standard deviation of 483.71. The income variable has a median value of 34,552.64, a mean value of 33,516.98, and a standard deviation of 13,336.64. The full list of descriptive statistics for the balance and income variables are shown in figure 1. The researcher then created a dummy variable for default and student to code “no” as 0 and “yes” as 1 and then looked at the correlation between the variables. There was almost no correlation between default and income, and default and student. There was a strong negative correlation between income and student. Students tended to have lower income. The correlation plot is shown in figure 2.

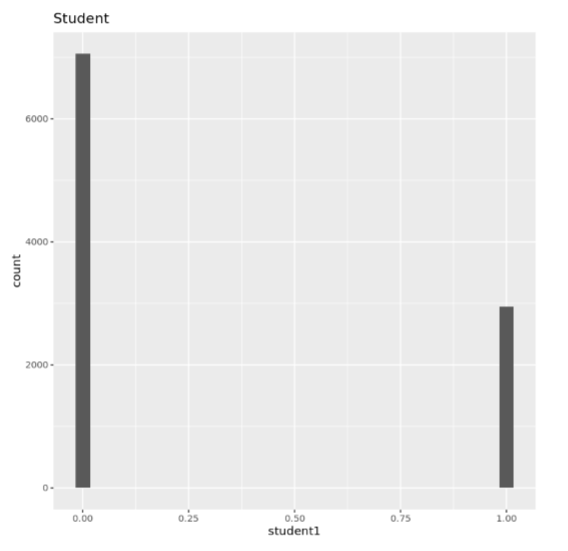
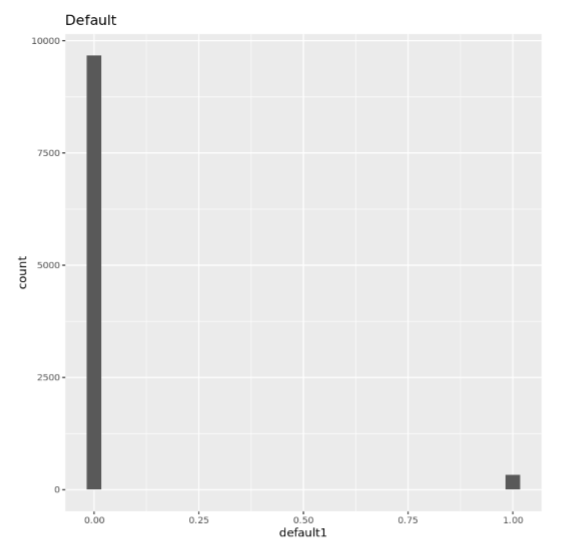
**Figure 1:** Descriptive statistics **Figure 2:** Correlation Plot

The researcher then built scatter plots in Tableau on Income versus Balance and colored them by Default (figure 3) and Student (figure 4). The scatter plot in figure 3 is showing that a customer is more likely to default on their credit card with a higher balance. The scatter plot in figure 4 is showing that students tend to have a lower income than non-students.

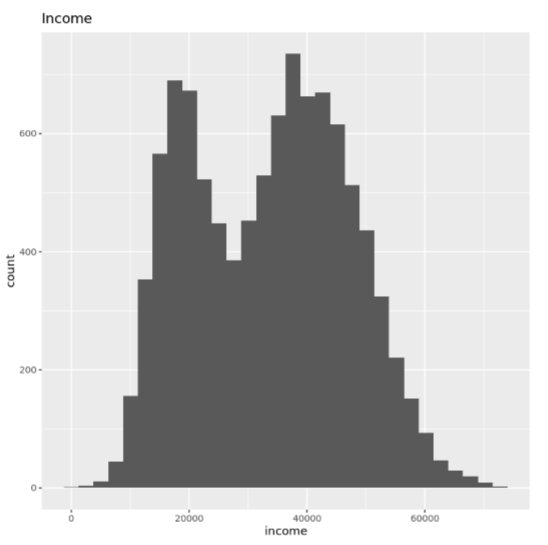
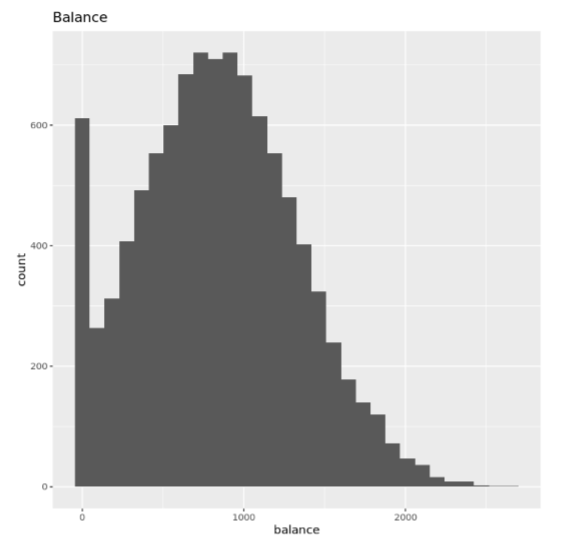


**Figure 3:** Income vs. Balance by Default **Figure 4:** Income vs. Balance by Student

The researcher then created histograms in R for each variable. These are shown in figures 5-8. Figure 5 is a histogram of the default variable which is a count of how many customers defaulted on their credit card. Figure 6 is a histogram of the student variable which is a count of how many customers are students or not. Figure 7 is a histogram of the balance variable. The histogram is showing the data for balance is skewed right and there are also a high number of customers that have zero balance. Figure 8 is a histogram of the income variable. Income appears to have two tiers of peaks where it peaks just below $20,000 and dips off, and then peaks again just below $40,000.

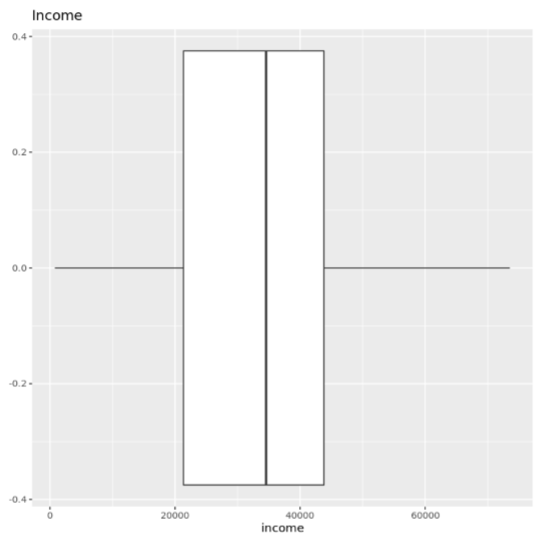
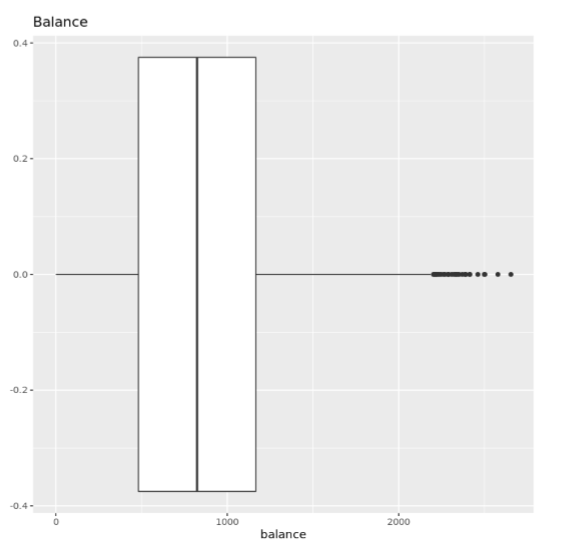


**Figure 5:** Histogram of Default **Figure 6:** Histogram of Student



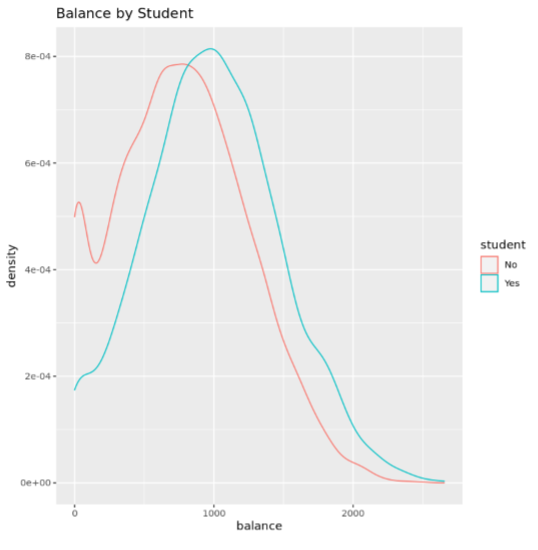
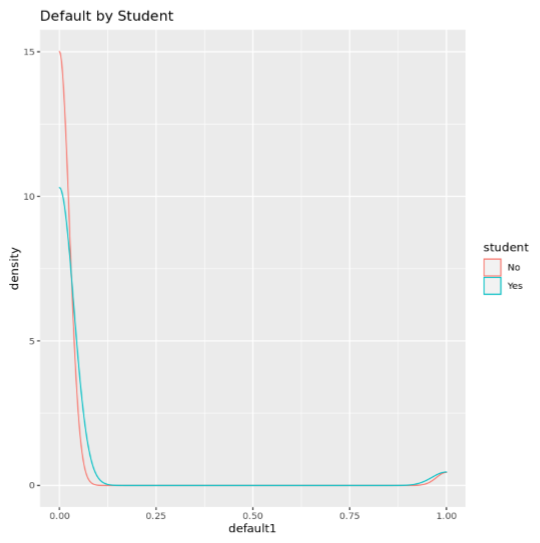
**Figure 7:** Histogram of Balance **Figure 8:** Histogram of Income

The researcher then created boxplots in R for each variable. These are shown in figures 9-10. Boxplots for the default and student variables are not helpful as they are categorical variables with only two factors. Figure 9 is a boxplot of the balance variable. The boxplot is showing a median value for balance around $800-$850. There are some higher values past $2,000 that make the data skewed right. Figure 10 is a boxplot of the income variable. The boxplot is showing the median value for income a little below $35,000.

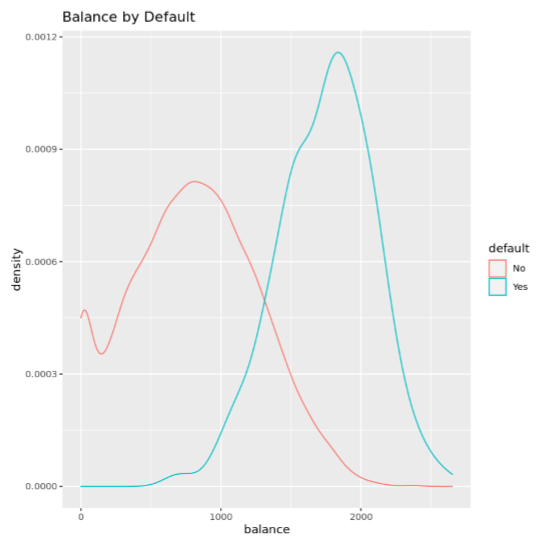
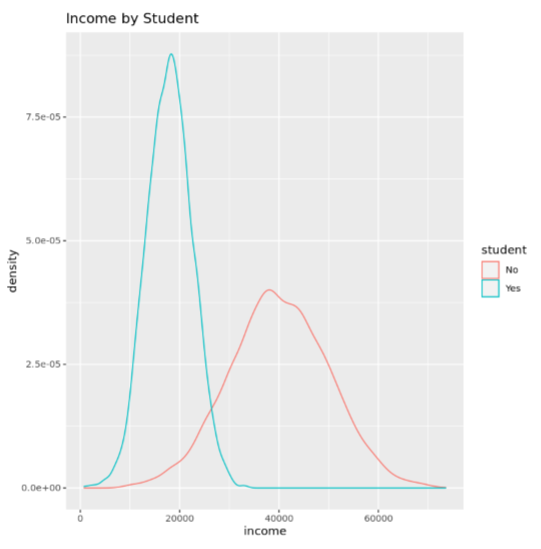


**Figure 9:** Boxplot of Balance **Figure 10:** Boxplot of Income

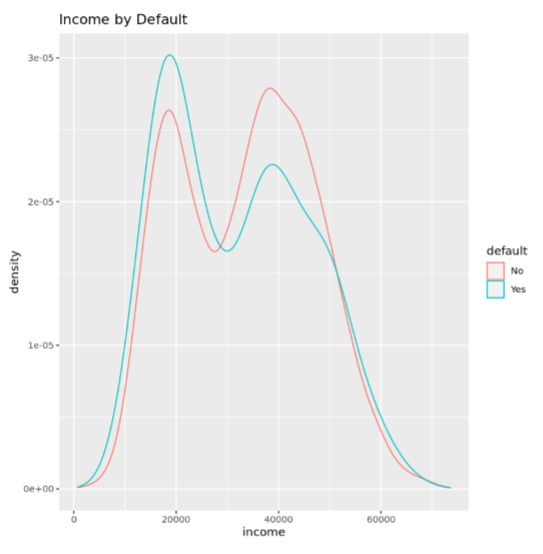
The researcher then created five density plots in R to look at differences of different variables by default or student. Figure 11 is a density plot of default, color coded by student. There does not appear to be too much of a difference in whether a customer defaults on their credit card by if they are a student or not. Figure 12 is a density plot of balance, color coded by student. Looking at the density plot, students appear to have a higher average balance than non-students. Figure 13 is a density plot of income, color coded student. This density plot is showing that students appear to have much lower income than non-students. Figure 14 is a density plot of balance, color coded by default. This density plot is showing that customers that have a higher balance are more likely to default on the credit card. Figure 15 is a density plot of income, color coded by default. This density plot is not showing too much of a difference between income levels and whether or not a customer defaults on their credit card.



**Figure 11:** Default by Student **Figure 12:** Balance by Student



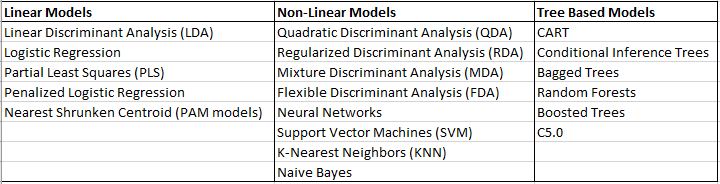
**Figure 13:** Income by Student **Figure 14:** Balance by Default



**Figure 15:** Income by Default

**Analytics Methods**

The researcher used three different types of analytics methods to look at the data. These were linear models, non-linear models, and tree based models. There were five linear model methods that were used which were Linear Discriminant Analysis (LDA), Logistic Regression, Partial Least Squares (PLS), Penalized Logistic Regression, and Nearest Shrunken Centroid (PAM models). There were eight non-linear model methods that were used which were Quadratic Discriminant Analysis (QDA), Regularized Discriminant Analysis (RDA), Mixture Discriminant Analysis (MDA), Flexible Discriminant Analysis (FDA), Neural Networks, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naive Bayes. There were six tree based model methods that were used which were CART, Conditional Inference Trees, Bagged Trees, Random Forests, Boosted Trees, and C5.0. Please see Table 1 for the analytics models and their types. There were nineteen models in total.

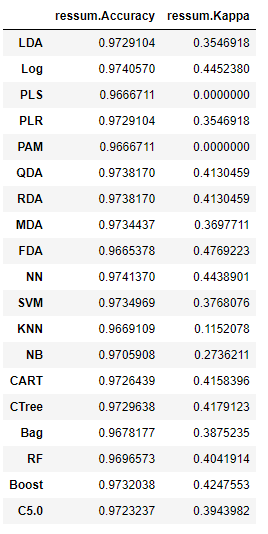


**Table 1:** Analytics Models

**Validation and Testing**

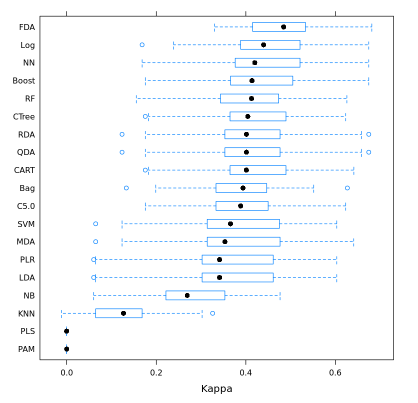
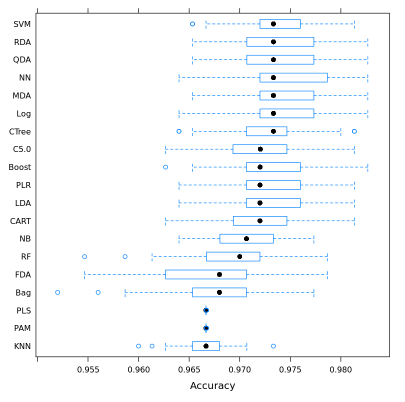
Before the researcher started building the models, the data was broken into a training set and a testing set. Three quarters of the data was used for the training set, and one quarter of the data was used for the testing set. The researcher then ran the nineteen models using the training set. Each model was set up with the default variable as the response variable and the student, balance, and income variables as the predictor variables. For each model, a confusion matrix was built to determine the accuracy of the model’s ability to predict whether a customer would default on their credit card. The metrics used to determine the success of the model were “Accuracy” and “Kappa”. “Accuracy” was used as the main determinant of the model’s success.

The researcher then gathered all the results of the nineteen models to compare them with the training set. Table 2 shows the accuracy and kappa results for each model on the training set.



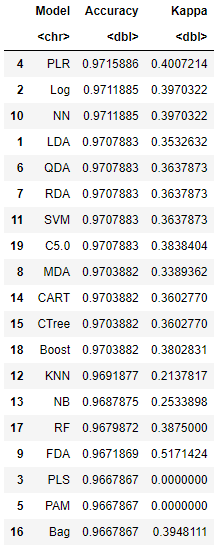
**Table 2:** Accuracy and Kappa training set results

The researcher also set up box plots to visualize the results of the accuracy and kappa metrics. Accuracy is shown in Figure 16 and kappa is shown in Figure 17.



**Figure 16:** Accuracy of training set **Figure 17:** Kappa of training set

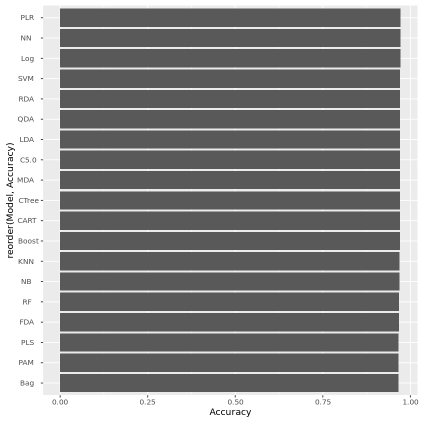
After the training set was analyzed with the nineteen models, the researcher used the testing set to predict whether a customer would default on their credit card based on the results of the models built with the training set. The researcher set up predictions for each of the nineteen models. Table 3 shows the accuracy and kappa for each of the nineteen models ordered by accuracy, as accuracy was determined by the researcher to be the most important metric.



**Table 3:** Metrics of testing set ordered by accuracy

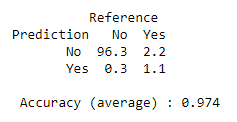
**Analysis of Results**

Analyzing the results, the accuracy of the nineteen models ranged from 96.68% for the least accurate model to 97.16% for the most accurate model. The accuracy range for all nineteen of the models was very close as the difference between the highest and lowest was less than a half of a percent. Figure 18 is a bar chart of the accuracy of each model and it shows just how close the accuracy range was for all nineteen models as you can barely tell the difference. The top three most accurate models were Penalized Logistic Regression, Logistic Regression, and Neural Networks. Penalized Logistic Regression and Logistic Regression were both linear models, and Neural Networks was a non-linear model. The three least accurate models were Bagged Trees, Nearest Shrunken Centroid (PAM models), and Partial Least Squares (PLS). Bagged Trees was a tree based model, and Nearest Shrunken Centroid (PAM models) and Partial Least Squares (PLS) were both linear models. It was interesting that there were two linear models in both the top three and bottom three performing models. Of the five linear models, three of them performed in the top four, while the other two performed in the bottom three. The non-linear models performed around the middle of the pack. The tree-based models performed around the middle and bottom of the pack.



**Figure 18:** Accuracy bar chart

The most accurate model was Penalized Logistic Regression with an accuracy of 97.16% on the testing set. A confusion matrix was built on the training set which achieved an accuracy of 97.4%. This is shown in Figure 19. The confusion matrix accurately predicted 96.3 out of 97.5 for customers that do not default on their credit card, and accurately predicted 1.1 out of 1.4 for customers that do default on their credit card. This is saying that the Penalized Logistic Regression model will accurately predict a customer that does not default on their credit card 98.7% of the time, and will accurately predict a customer that does default on their credit card 78.6% of the time. This is showing that it is much harder to predict a customer that will default on their credit card than a customer that will not default on their credit card. This is usually the case when predicting a class with fever observations, as the dataset had 9,667 out of 10,000 observations of customers who did not default on their credit card, as opposed to 333 out of 10,000 observations of customers who did default on their credit card. That being said, the Penalized Logistic Regression model is still an accurate model.



**Figure 19:** Penalized Logistic Regression confusion matrix on training set

**Discussion and Conclusions**

The purpose of the project was to determine whether a customer will default on their credit card using the Default dataset in the ISLR library in R. The predictor variables were student, balance, and income. The models built to predict credit card default all were able to predict whether a customer will default on their credit card with an accuracy between 96.68% and 97.16%. Penalized Logistic Regression was the most accurate model to predict credit card default and it was a linear model. Given the predictor variables of student, balance, and income, Penalized Logistic Regression was able to accurately predict whether a customer will default on their credit card with an accuracy of 97.16%. From our exploratory data analysis, the balance variable was the best predictor of credit card default.

Possible next steps taken for this project would be to continue to improve on some of the models using tuning parameters and extended grid searches. An additional step would be to look at other possible variables that may impact whether a customer will default on their credit card. Some other possible variables could be marital status, gender, education level (not just whether the customer is a student or not), and credit card limit.

**References**

Default dataset. ISLR library. R package

*Classification for credit card default - GitHub Pages*. (n.d.). Retrieved February 26, 2022, from http://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/ClassificationProcessCreditCardDefault.html

Selener. (2019, May 19). *Prediction of credit card default*. Kaggle. Retrieved February 26, 2022, from https://www.kaggle.com/selener/prediction-of-credit-card-default