Machine Learning Classification of Credit Card Default Payments

Amir Ghaderi – 500794236

March 27th, 2017

ME8135 – Data Visualization

Ryerson University

Introduction

Financial institutions are actively seeking new and innovative ways to tap into the world of data science. There is a huge market centered around extracting valuable information from data that will help drive business decisions. Credit card companies are one of the major players in this space. One of the major issue that these companies are faced with, is payment defaulting. Defaulting in the context of credit card payments, refers to when an individual fails to make their credit card payment. Credit card defaulting is a major issue for credit card companies because it can result in a loss of profit. Therefore, these companies are financially motivated to search for patterns in their datasets that can potentially prevent a payment default. By way of this project I am aiming to provide answers to two questions that credit card companies would be interested in. The first question is which variables are the strongest predictors of credit card default payments. The second question is which machine learning algorithm provides the highest accuracy when classifying unlabeled points.

Dataset

The dataset that I used for the completion of this project is a credit card dataset that comes from the UCI Machine Learning Repository and can be obtained from www.Kaggle.com. The dataset is titled: “Default of Credit Card Clients Dataset: Default Payments of Credit Card Clients in Taiwan from 2005”. The unedited dataset contains 30000 records, 25 features, and 2 classes (1= default, 0 = not default). The features include: {Id, Limit, sex, education, marriage, age, Payment1:6, Bill_amount1:6, Pay_amount1:6, Class}.

A complete description of the dataset can be found here:

https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

Technologies

The technologies used for this project include R and D3. R was used for data processing and D3 was used for data visualization. There were several R packages that were used for various purposes, see list below:

- library(class): The "class" library was used for the development of the k-nearest neighbor classification model.
- library(e1071): The "e1071" library was used for the development of the support vector machine classification model.
- library(ipred): The "ipred" library was used for the development of the rpart decision tree classification model.
- library(randomForest): The "randomForest" library was used for the development of the random forest decision tree classification model.
- library(C50): The "C50" library was used for the development of the C50 decision tree classification model.
- library(tree): The "tree" library was used for the development of the tree decision tree classification model.
- library(neuralnet): The "neuralnet" library was used for the development of the neural network classification model.
- library(corrplot): The "corrplot" library was used to create a correlation matrix in R prior to creating one in D3.
- library(ggplot2): The "ggplot2" library was used to create visualizations in R prior to creating the D3 visualizations.

Challenges

No matter the project, there will always be associated challenges and hurdles. Throughout this project, I faced several types of problems. The first challenge that I was faced with was attempting to pass my data through various machine learning classifiers in R. Each of the machine learning classification algorithm requires data to be in a specific format. For example, decision trees require all features to be factors while, neural networks just require the class to be a factor. I was able to overcome this challenge by looking up the documentation for each of the classification functions and transforming the data appropriately. In addition, some of the machine learning algorithm (i.e. Neural Networks or SVM) are extremely computationally heavy and take a significant amount of time to compute. Thus, I was limited to only passing a subset of my features into these classification models. I overcame this issue by examining my correlation matrix and selecting the most valuable subset of features. The second challenge that I was faced with was creating visualizations in D3. Due to my unfamiliarity with HTML, CSS, JavaScript, and D3 this was probably the most difficult part of the project for me. In order to overcome this challenge I needed to spend a significant amount of time researching JavaScript functions and D3 syntax.

Preprocessing

For this project, all of the data processing and data manipulation was conducted in R. The downloaded csv file was first imported into R using the read.csv () function. After the data was loaded into R, the class column was then converted into a factor type. This was important because many of the machine learning classification models require the class to be a factor in order to function properly. Next, all of the features were normalized using a custom

normalization function into a range from 0(min) to 1(max). Feature normalization is important because it puts all of the features on the same scale. This is particularly important when using classification algorithms that use distance functions (i.e. Euclidean). Once feature normalization was completed, I split the data into training (80%) and testing (20%) sets. By splitting the data into training and testing sets, this allows me to test the accuracy of my models. Next, I randomized the records of my dataset in order to reduce bias in my classifiers. I found that if I pass my data into a classification algorithm when my class column is sorted, the classification accuracy gets diminished. Next, I ensured that the distribution of classes in my testing set were balanced. Through experimentation I found that a balanced testing set (equal distribution of classes) trains my classification models with less bias. Finally, I removed some of the non-correlated features from my dataset because some of the classification algorithms are very computationally heavy.

Visualizations

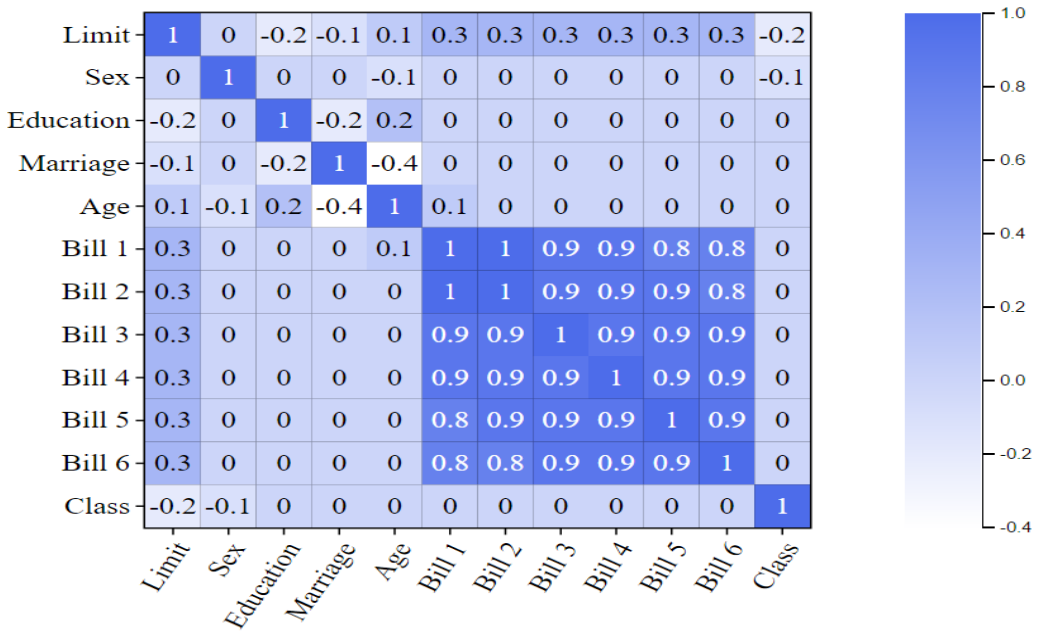


Figure 1: The first visualization is a correlation matrix that was constructed using D3. This visualization provides critical information regarding correlations that exist in the dataset. The most significant piece of information that can be extracted from this matrix is that only Limit and Sex are correlated with Class. However, it is important to note that these correlations are very weak.

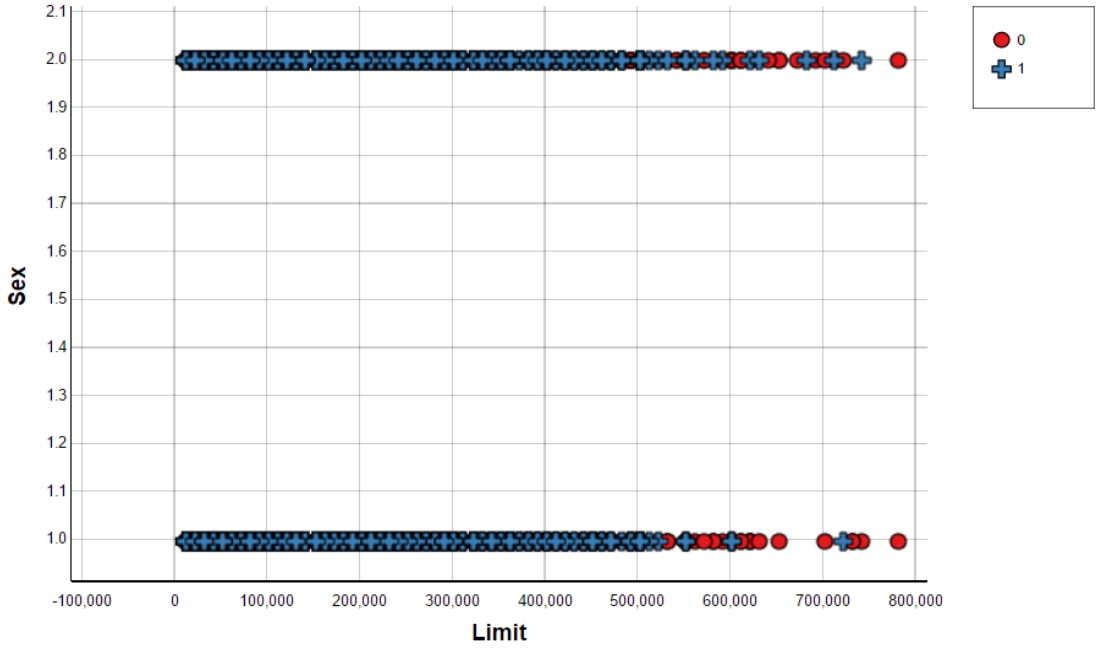


Figure 2: The second visualization is an interactive scatter plot constructed using D3. This visualization plots Limit against Sex in an attempt to predict class. The interactive component of this plot is the dynamic x and y axis's. This allows the user to zoom into dense areas of the plot and gain valuable information.

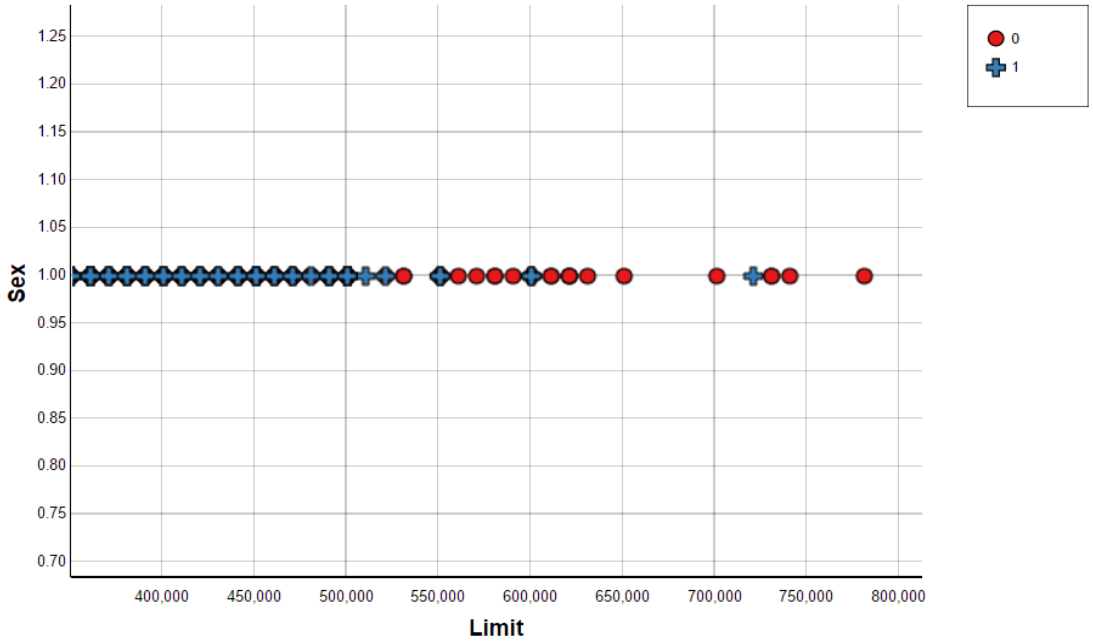


Figure 2: The above visualization is a zoomed in version of the interactive scatter plot. This visualization makes it clear that there exists a cluster of class 0, where sex = male and limit is greater than 550,000.

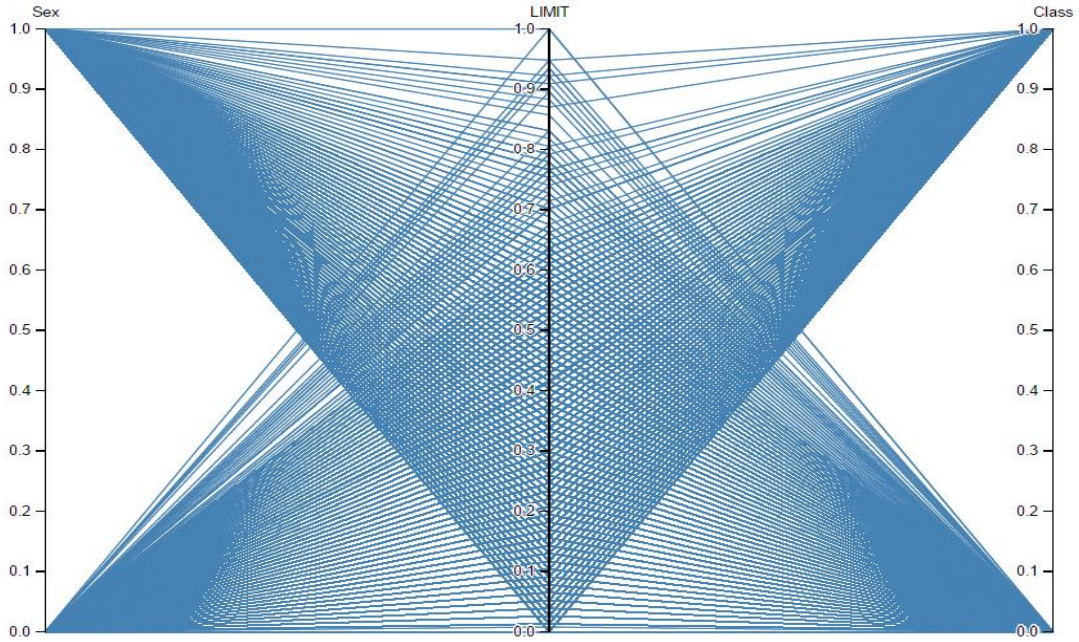


Figure 3: The third visualization is an interactive parallel coordinates plot constructed using D3. At first glance this visualization may look useless because of the amount of data being shown. However, the interactive component allows users to limit the number of coordinates being shown.

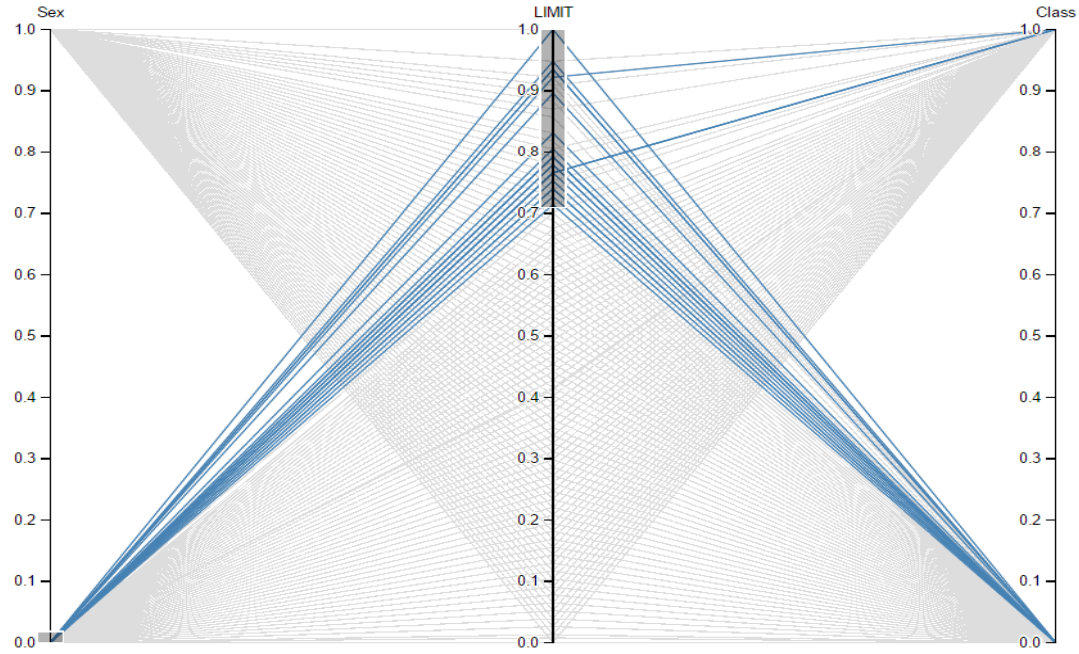


Figure 3: This figure is an limited version of the parallel coordinates plot above. As you can clearly see there exists a cluster of coordinates that map to class 0, when sex = male and limit is in the top 30%.

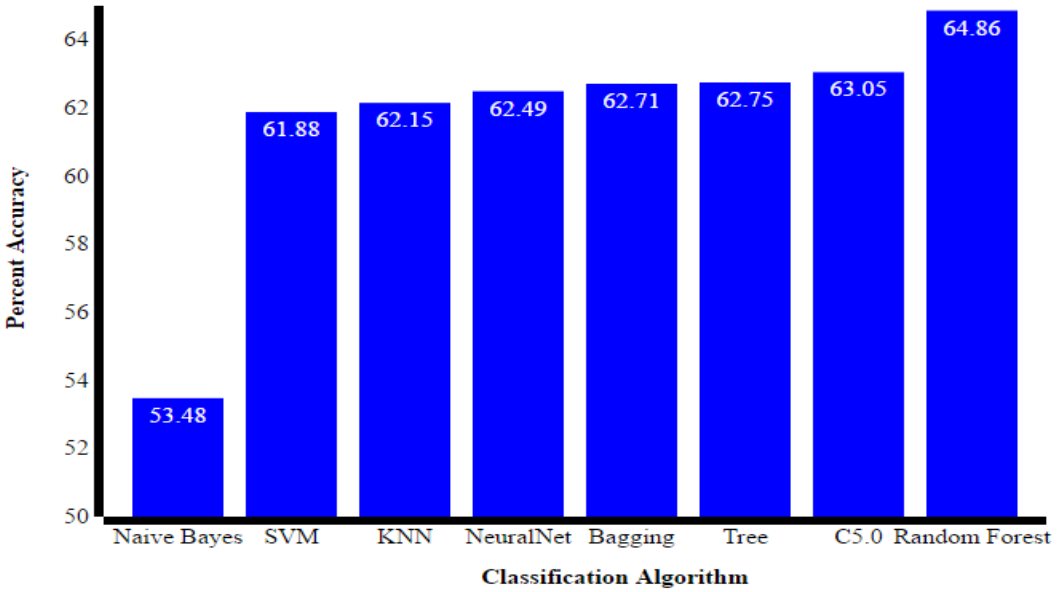


Figure 4: The forth visualization is a bar plot that was constructed using D3. This plot represents the percent accuracy for each of the classifiers tested during this project. As you can see decision trees consistently performed better than any other machine learning classifier. In addition, among the decision tree algorithms, Random Forest produced the highest accuracy.

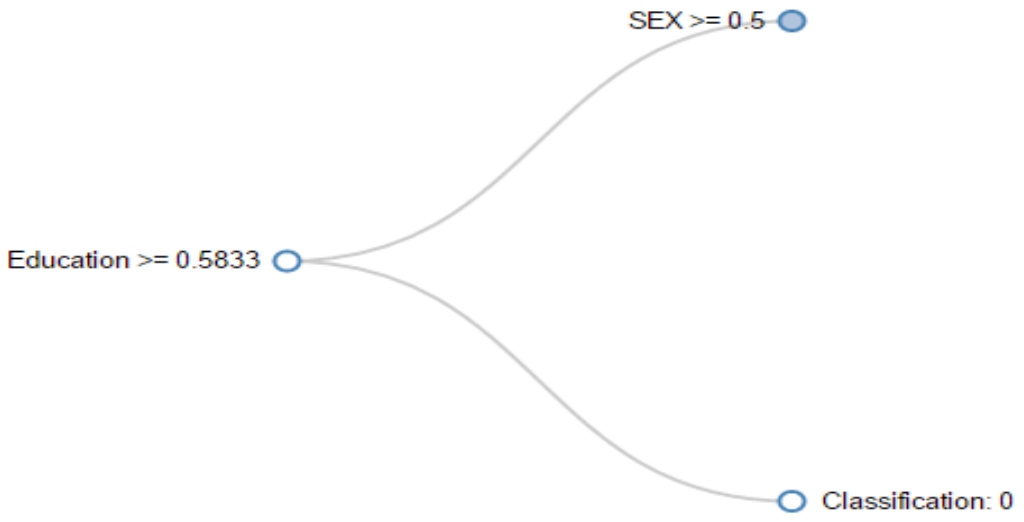
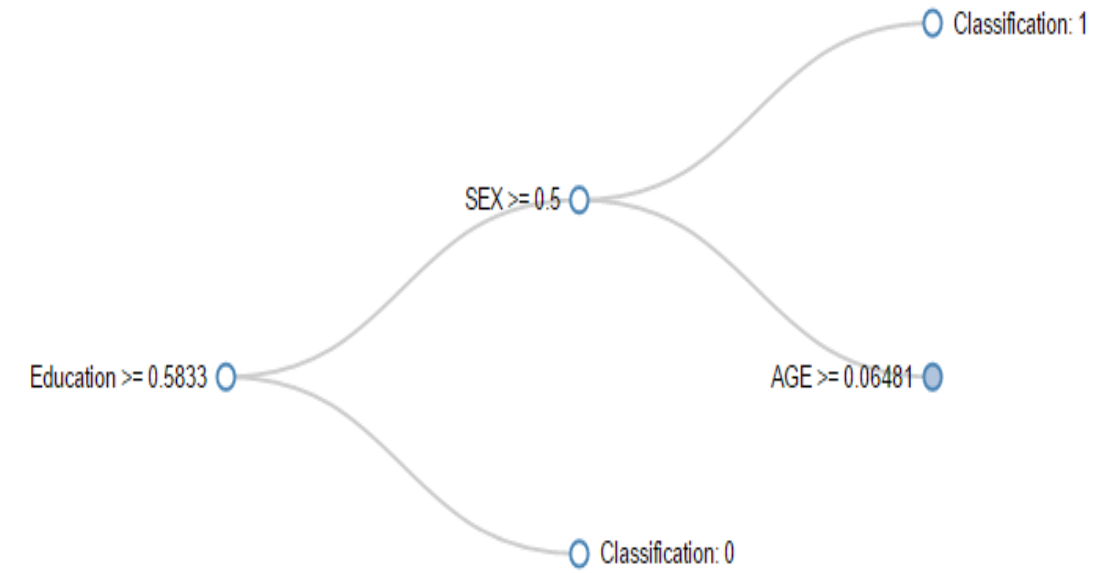
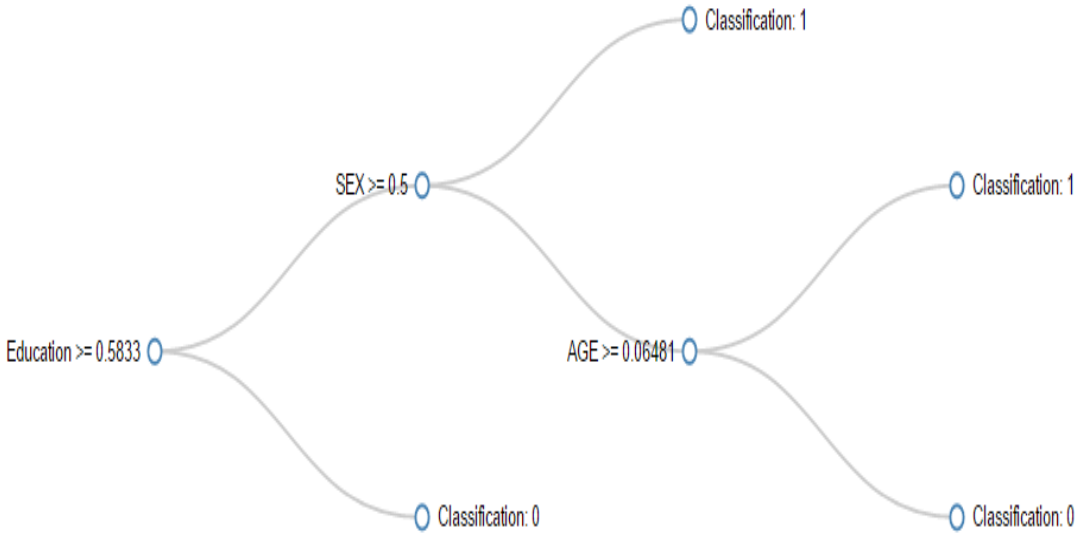


Figure 5: The fifth visualization is an interactive decision tree constructed using D3. The interactive component of this visualization allows the user to click and expand the nodes of the decision tree until a leaf node is reached. This decision tree represents an instance of the random forest algorithm, which produced the highest percent accuracy on the dataset.





Conclusions

In conclusion, the two major questions that I was attempting to answer through this project was (1) which variables are the strongest predictors of default payments and (2) which machine learning algorithm provides the highest accuracy when classifying unlabeled points. Through careful analysis, it was concluded that Limit and Sex were the variables that are the

strongest predictors of default payments. After comparing the accuracy of various machine learning classifiers it was determined that decision trees were the most accurate, in particular the random forest variation. Overall, the project was successful and achieved the desired results.

Lessons Learned

There are several lessons learned and experiences gained throughout the completion of this project. Firstly, I gained valuable skills in HTML, CSS, JavaScript and D3. Over the past couple years D3 has grown to be an industry leader for producing interactive visualizations. Therefore, adding this experience to my portfolio is very beneficial to my future career. Secondly, I gained valuable experience working with and applying machine learning classifications models to a large dataset. Prior, to this project I never had the opportunity to apply machine learning classification algorithms to large datasets. This experience will also bring a lot of value to my portfolio as I one day hope to work with large datasets. Thirdly, I gained valuable skills with data processing and data manipulation in R. Finally, through this project I was introduced to Kaggle and its vast library of datasets. Kaggle will be a valuable resource for my future projects and potentially my Major Research Project (MRP). Overall, I feel that this project brought a significant amount of value to my skill set.

References

Bar Plot - https://www.youtube.com/watch?v=Fjmxh-gnBM0

Correlation Matrix - https://bl.ocks.org/arpitnarechania/caeba2e6579900ea12cb2a4eb157ce74

Scatter Plot - https://bl.ocks.org/starcalibre/f4b8bb0da3b2090c56d79646a338fd81

Parallel Coordinates – https://bl.ocks.org/jasondavies/1341281

Decision Tree – https://bl.ocks.org/ajschumacher/65eda1df2b0dd2cf616f

Confusion Matrix - https://bl.ocks.org/arpitnarechania/dbf03d8ef7fffa446379d59db6354bac