Credit Default Risk Prediction – John Arnold

# Introduction

According to a 2012 paper published by Statistics Canada, 66.4% of Canadians are liable for some form of debt, with a mean debt per borrower being 114,400 dollars1. Non-mortgage borrowing amounted to 7.1 billion dollars in Canada (Q3 2017)2. Banks and lenders constantly have to evaluate credit risk and lines of credit of borrowers to determine whether they are worthy of a loan.

The main research question is**: What are the main attributes of borrowers that lead to high risk of defaulting on loans?**

The bulk of the work to answer this question will involve dimensionality reduction, exploratory data analysis and feature selection on the 140 columns to narrow them down to a subset of impactful attributes in predicting the class attribute **loan status**. The coding will be done in **R**.

The second part in answering this question will involve testing several classification techniques such as logistic regression and decision tree algorithms on the data to develop an effective prediction model. Finally, a probability will be assigned to all borrowers currently paying off loans to be validated in the future.

GitHub link: <https://github.com/jwarnold/Capstone>

# Literature Review

# 1 - Developing Prediction Model of Loan Risk in Banks Using Data Mining

This study from University Khartoum built a model from banking sector data to predict the future status of loans. The study was concerned to first defining what credit risk really is, then formulating a model to predict it. Three algorithms were used on structured data using the Weka tool to implement and test the models. They tested J48, BayesNet and Naïve Bayes models, with the J48 decision tree algorithm performing the best.

# 2 – Data Mining for loan evaluation at ABN AMRO: a case study

This study from the University of Twente in the Netherlands is very similar to the research question of this project. It involved using historical data of banking clients to predict their repayment behaviour. They first determined which characteristics of clients made them more or less likely to repay a loan from the bank. Positive attributes on loan repayment included an age greater than 45, employment length longer than 15 years and home ownership. They then tested classification models of Linear Discriminant Analysis and C4.5 decision trees to distinguish defaulters from non-defaulters. The C4.5 decision tree used less attributes than linear discriminant analysis but performed marginally worse (2%).

# 3 – Assessing Credit Default Using Logistic Regression and Multiple Discriminant Analysis

This study was concerned with the default risk of companies in south eastern Europe, specifically in Bosnia and Herzegovina. It investigated bankruptcy prediction within the banking market using logistic regression and multiple discriminant analysis. The model class variable was whether a company was 0: Healthy or 1: Defaulting. What they found was that, return on assets (ROI) - how effectively the bank’s assets were generating revenue was the most statistically significant attribute. Logistic regression had better predictive ability than that of multiple discriminant analysis

# 4 – Credit Risk Analysis using Machine and Deep Learning models

This study examined businesses applying for loans from a bank ranging from small to enterprise level. Each business was classified into a binary category as either default or non-default. Its datasets were characterized by general metrics that signify the health of an organization such as financial statements, balance sheets, income statements and cash-flow statements. They used 7 different models such as logistic regression, random forest, gradient boosting and variations of neural networks with ROC and AUC criteria as accuracy measures. They found the tree based models performed far better and were far more stable than the neural network models. With AUC on the random forest and gradient boosting to be 0.99.

# 5 – Default Prediction of a Swedish Mortgage Portfolio using Logistic Regression

This study focused on predicting default of customers taking out mortgage loans. The concern of the study was that when banks give out loans, a limited number of factors are taken into account such as assets, debt and income. This study set out to factor in bank card transactions, direct debits and other customer specific factors into models in determining whether to provide a loan. They developed a logistic regression model to factor in several customer attributes such as their municipality, their fund, the number of direct debits they make on their account, and sum of card payments. The model they developed had a strong predictive ability due to its AUC of 0.93 and a low classification error of <0.05.

The significant attributes they found were the sum of the borrower’s assets, if they had a sub borrower, their income among others.

# 6 – Logistic Regression Analysis Of Predictors Of Loan Defaults by Customers of Non-Traditional Banks in Ghana

This Ghanaian study focused on the default risk of borrowers with microfinance loans. Microfinance loans being define as “small financial transactions using non-standard methodologies like character-based lending, group guarantee and short term repeated loans”. They developed a binomial logistic regression model fitted to 548 customers who were granted credit. Predictor variables used in this study included age, gender, marital status, dependents, purpose of loan and number of years at residence among others. Their study found that marital status, dependents, type of collateral, duration and loan type were significant factors in predicting whether a borrower will default or not on their loan. The logistic regression model that they developed had a successful default prediction rate of 86.67% on borrowers with the significant attributes in mind.

# Dataset

The dataset that will be used in this project is 2016 historical and current loan data from the website LendingClub.com. LendingClub puts together borrowers and investors who are willing to purchase the debt with a promise of a return through the interest rate on the loan. Borrowers take a loan for a number reasons such as home improvement, debt consolidation, car finance etc. The loans range from 1,000$ to 40,000$ for personal use. LendingClub does not provide loans to borrowers with a credit score below 600. **The dataset and corresponding data dictionary can be found in Reference [1].**

The 3 most common employment titles were teacher, manager and owner **(Figure 2a, pg. 9).** Most of the loans in this dataset came from Californian and Texas **(Figure 2b, pg. 9)**, interestingly the state of Iowa had no borrowers, possibly due to state-specific laws on loaning money.

Most variable types in this dataset are numeric in nature. They range from percentages, to number of months, to tangible values such as loan amount, installment amount, credit card balance etc. In general most of the categorical variables were removed with the exception of the class variable, term of loan (36/60 months), home ownership type and purpose of loan. In general, most borrowers took out a loan to consolidate their debt, (**Figure 2c, pg. 10).** The categorical variables that were removed either had too many classes such as ZIP code, residing U.S. state or offered no predictive value by nature of the dataset.

Mean loan amount and interest rate are 15,508$ and 12.99% respectively, with both distributions being skewed right, see **Figures 1a and 1b, pg.8.** The majority of borrowers in this dataset had been at their jobs for longer than 10 years, see **Figure 1c, pg. 8.** The class variable was balanced with 52,364 individuals having fully paid off their loans and 22,207 being classified as defaulters. One stipulation about this figure is that individuals who were classified as being late, in grace period or on a payment plan were classified as defaulters.

Of the 142 columns in this dataset 60 have been immediately removed. There are 3 general clusters that were removed from the dataset. The first was the set of columns that are filled after the fact that a borrower becomes delinquent on their loans. The columns involve collection fees and post-default payment plans. The reasoning is that since these are filled after the fact of defaulting, they offer no information as to predicting whether they will default. The second cluster of columns that were removed involved joint borrowers. The vast majority of rows are individual borrowers (>95%). The third cluster of columns that were removed are due to various data-cleaning techniques such as low variance filter, high correlation and large amounts of missing values.

The main attributes of interest in the exploratory analysis involved loan amount, interest rate, credit grade, home ownership type, annual income, past delinquency, open credit lines and amounts. These attributes are the ones found to be most pertinent by past studies and literature on the subject. One main stipulation of this research project is that is ***not***representative of all borrowers. By nature of the websites loan criteria, it is only representative of borrowers who have a credit score of above 600.

# Approach

## Step 1: Problem Identification and Data Pull

With the goal of this project to find out what attributes borrowers have to lead them to defaulting, the data is pulled from LendingClub.com with data on 130,000 borrowers with relevant financial attributes.

## Step 2: Exploratory Data Analysis and Cleaning

This step will involve rigorous cleaning of the data by removing as many columns as possible from the large pool of 142 without specialized techniques. This will then be followed by EDA to understand the structure of the dataset as well as accompanying descriptive statistics.

## Step 3: Dimensionality Reduction

This step will involve the usage of common specialized techniques to further reduce the attributes of the dataset such as low variance filtering, high correlation filtering, random forest feature selection and high NA occurrence filtering

## Step 4: Modelling

Several classification models will be implemented with the one returning the highest test accuracy to be chosen for future analysis. Initial models will include logistic regression, C5.0 decision tree, and random forest modelling. For the logistic regression model, the probability cut-off value will be 0.5 to separate defaulters from borrowers who pay their loans back in full.

## Step 5: Testing

All of the models will be tested on the 130,000 rows using a 70/30 random sample train/test split.

## Step 6: Model Evaluation

Each of the models will be tested using Confusion matrices on true positive, false positive, true negative and false negative metrics as well as ROC curves and their corresponding AUC values.

## Step 7: Discussion and Conclusion

The findings and accuracies from the three models will be discussed to report which attributes of borrowers indicate a high loan default risk, how similar the models classified borrowers and what findings will be valuable in the future when providing loans to potential borrowers.

# Results

**Important Attribute Findings**

Of all the three models used in this project the three models resulted in similar selections of important features in predicting loan default risk. The C5.0 classification tree and the random forest model had more overlap then with the logistic regression model. This is because of the similar calculations on information gain and entropy that is used within the C5.0 and random forest algorithms. Of the top 10 variables that were selected by the C5.0 classification tree and random forest model, interest rate, average current balance, annual income and grade were where the overlap between the models occurred.

The logistic regression model found a number of attributes **(Table 1a, pg. 8)** that were similarly found in past research on this topic **[2]**. Examples of significant attributes in the logistic regression model include interest rate, home ownership type, debt-to-income ratio, purpose of loan and employment length.There was also overlap in important features with the other decision tree based models.

In determining the importance of attributes when predicting whether a borrower is likely to default or not on a loan, all three of the tested models reported back attributes that corroborates past literature on the topic. The attributes that were selected as the most important by all three models were, interest rate on the loan, annual income, credit grade. These attributes tie in very closely with past literature on the topic.

**Classification Accuracies**

When running the models on the training set and testing set comprised of the remaining 75,000 borrowers (after current payers removed), all three models returned similar accuracy values based on TP/FP/TN/FN confusion matrices as well as AUC measures.

The logistic regression model performed the best, (although only marginally) with an AUC of 0.7560 **(Figure 3a, pg.12)** and true positive/true negative accuracy of 0.7238. The random forest model saw an AUC of 0.7362 **(Figure 3b, pg. 12)** and TP/TN accuracy of 0.7199. The C5.0 decision tree performed marginally worse than the other two models with a TP/TN accuracy of 0.7088. See **Tables 2a through 2c, pg.13** for TP/TN accuracies

Interestingly, the logistic regression and random forest models underestimated the amount of borrowers who would default and the C5.0 decision tree grossly overestimated the amount of defaulters. I.e. the C5.0 decision tree predicted 6,696 people would default when in reality only 1,559 people defaulted in the test set. In the context of a business case, overestimating the amount of defaulters would be seen as worse because that would result in certain loss of revenue. Conversely, if they overestimated the amount of borrowers who would pay their loans back in full, there would still be a possibility of recouping the funds.

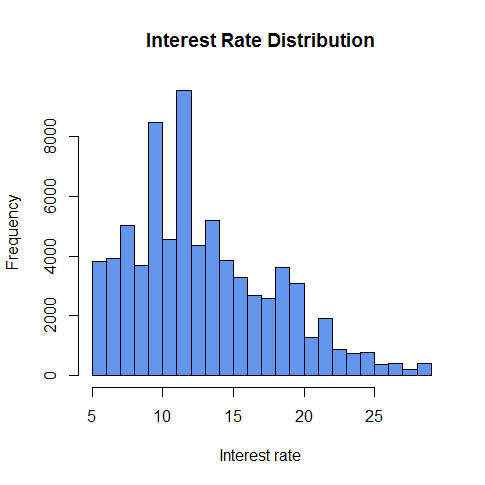
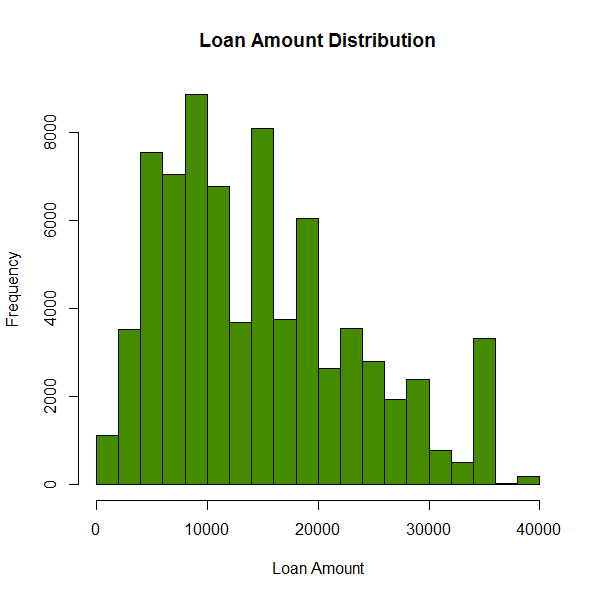
# Conclusions

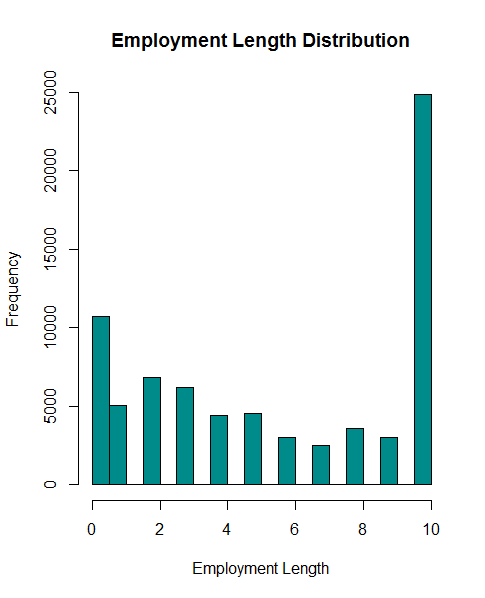
At the beginning of this project, its aim was to find out which attributes that borrowers have that would lead them defaulting. Research on this topic concluded that the following attributes are the most significant in predicting whether a borrower will default or not, namely employment length, home ownership, borrowers assets (average current balance), income and debt-to-income ratio [2].

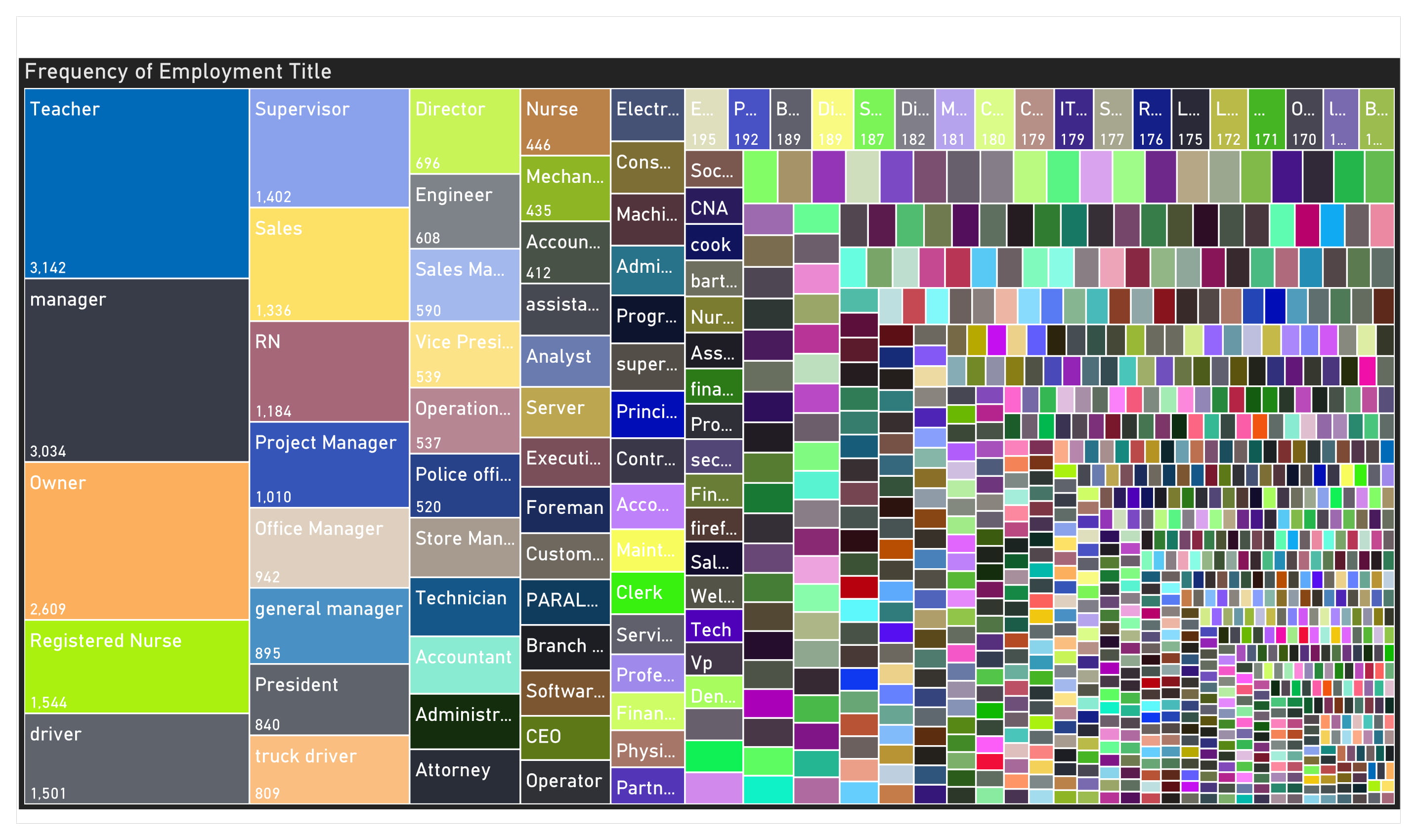
This project found similar attributes in determining whether a borrower would default or not. The models all achieved modest accuracies of roughly 72% in predicting our class variable of Loan Status. The usefulness of the study mainly serves this student as there is ample research and work done on this topic every day in the financial sectors of the economy and academia. With far more context and precision. The main insights to draw from the project is if someone were to come to you asking for a loan, one should verify all of the significant attributes found in this study and ask the relevant questions before trusting them with a loan.

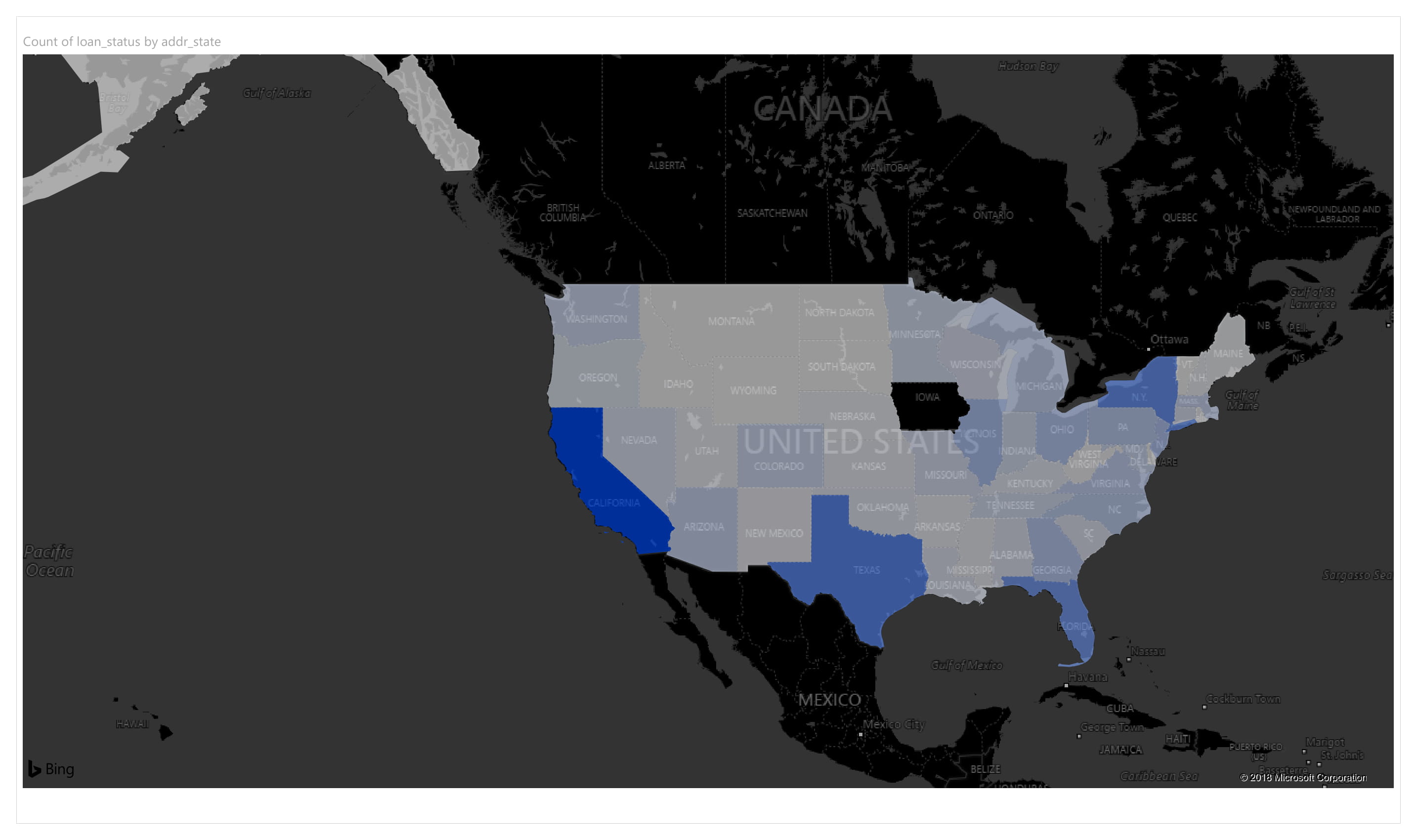
**Descriptive Statistics**

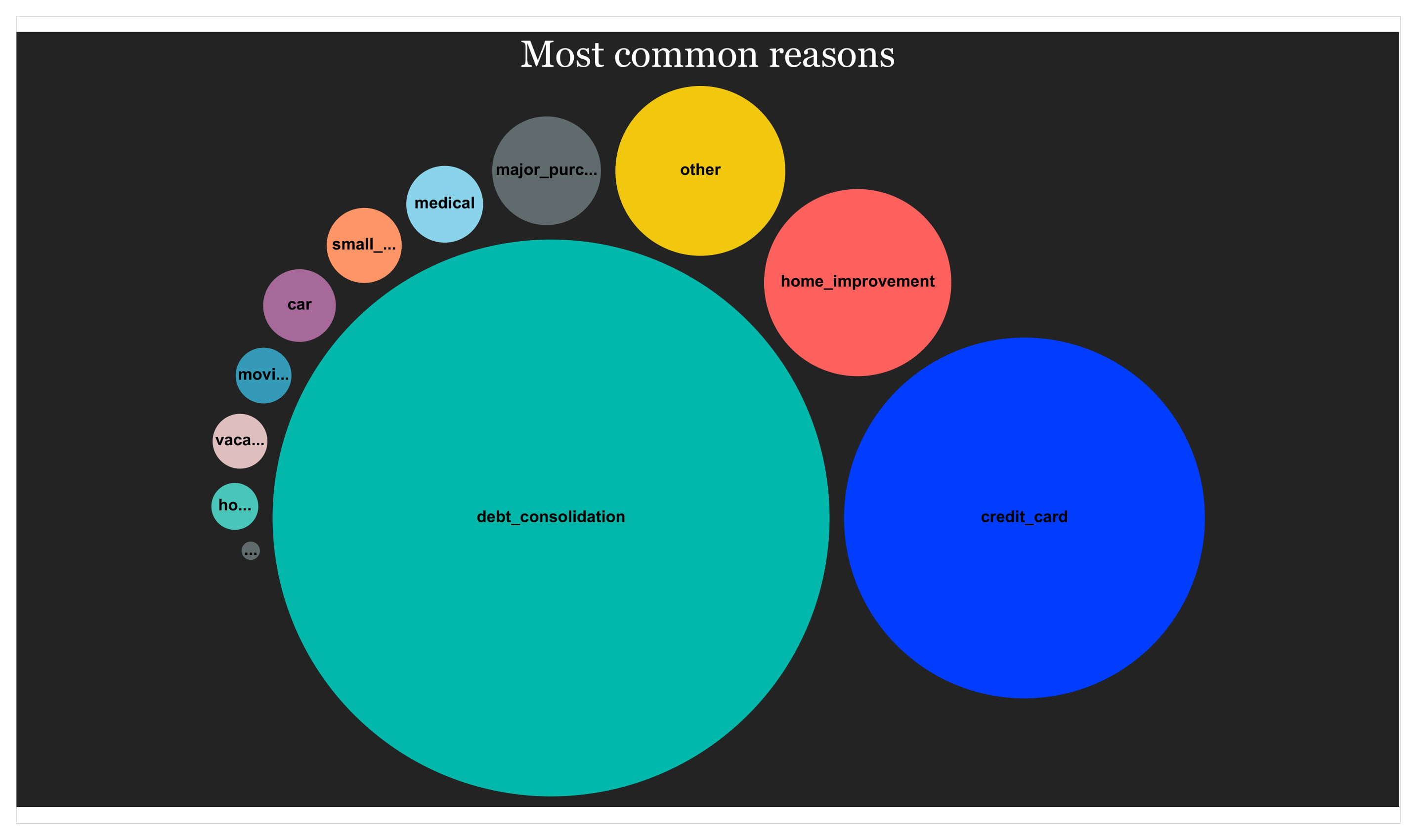
**Figure 1a Figure 1b**

****

**Figure 1c**

**Figure 2a – Frequency by Employment title**

**Figure 2b– Heat map of number of loans by U.S. State**

**Figure 2c – Reasons for taking out loan**

**Table 1a – Logistic Regression Significant Variables**

|  |  |
| --- | --- |
| **Variable** | **p-value** |
| Term – 60 months | 0.0000378 |
| Interest rate | 0.00019 |
| Debt-to-Income Ratio | 2.22x10^-11 |
| Total revolving credit high limit | 0.0000359 |
| Grade | 0.0000103 |
| Average current balance | 0.0297 |
| Annual Income | 0.008379 |
| Purpose: Medical, vacation and Small Business | <0.00005 |
| Home Ownership: Rent | <2x10^-16 |
| Employment Length | 3.40x10^-16 |
| Mortgage Accounts | 0.000469 |

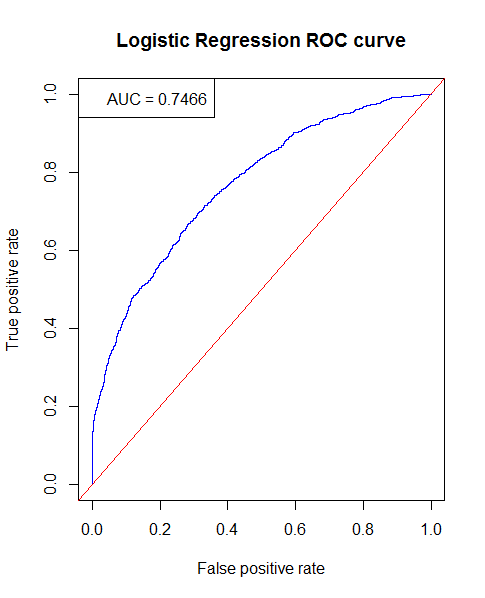
**Table 1b– Random Forest Significant Variables**

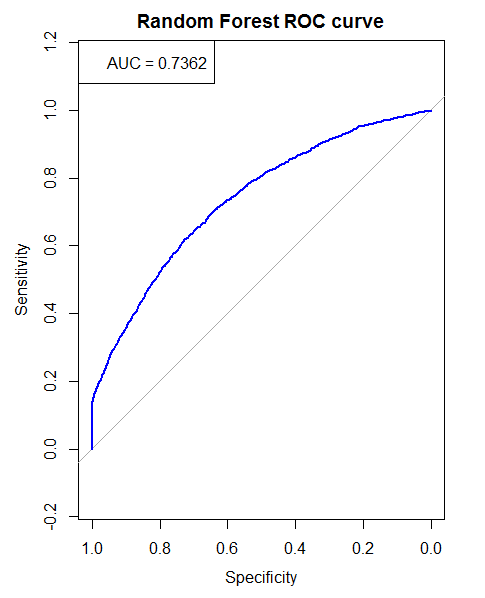
Mean Decrease Gini as the decrease in accuracy if one of the variables are removed across entire forest of 100 trees

|  |  |
| --- | --- |
| **Variable** | **Mean Decrease Gini** |
| Interest Rate | 1284.70 |
| Debt-to-Income Ratio | 820.706 |
| Grade | 734.4122 |
| Months since oldest revolving account | 695.038 |
| Installment | 674.122 |
| Average current balance | 673.74 |
| Total Bankcard Limit | 667.331 |
| Total High Credit Limit | 667.05 |
| Total Revolving High Limit | 665.986 |
| Annual Income | 645.0272 |

**Table 1c– C5.0 Variables Classified by Attribute Usage / Hierarchy Levels**

|  |  |
| --- | --- |
| **Variable** | **Attribute Usage** |
| Interest Rate | 100% |
| Average Current Balance | 46.67% |
| Term | 25.95% |
| Number of Revolving Accounts | 22.96% |
| Number of Mortgage Accounts | 22.77% |
| Purpose | 21.37% |
| Home ownership | 16.76 |
| Grade | 12.02% |
| Annual Income | 10.49% |

**Figure 3a – Logistic Regression ROC Curve**

**Figure 3b – Random Forest ROC Curve**

**Table 2a – Logistic Regression Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted** | | |
| **Accuracy: 0.7238** | | **Paid** | **Default** |
| **Actual** | **Paid** | 14477 | 1199 |
| **Default** | 4980 | 1716 |

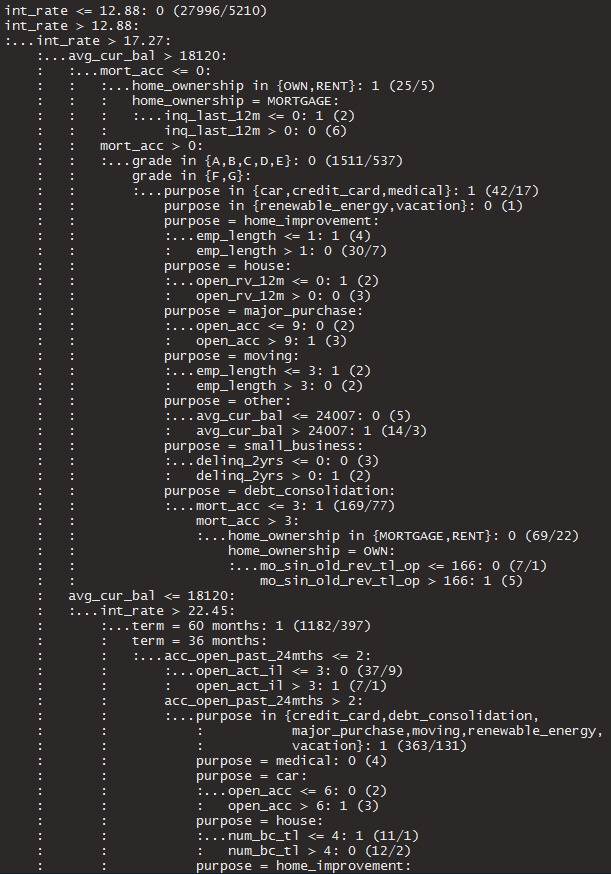
**Table 2b – Random Forest Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted** | | |
| **Accuracy: 0.7199** | | **Paid** | **Default** |
| **Actual** | **Paid** | 14503 | 1173 |
| **Default** | 5092 | 1604 |

**Table 2c – C5.0 Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted** | | |
| **Accuracy: 0.7088** | | **Paid** | **Default** |
| **Actual** | **Paid** | 14300 | 5137 |
| **Default** | 1376 | 1559 |

**Figure 4 – C5.0 Decision Tree in R console**

**** Figure omitted due to high number of nodes

**References**

**Introduction**

1) Chawla, R.K., Uppal, S., Household Debt In Canada, Statistics Canada Catalogue Perspectives on Labour and Income, no. 75-001-X, (2012)

2) Canada, Statistics Canada, National Balance Sheet and financial flow accounts, third quarter 2017, National, (2017), <https://www150.statcan.gc.ca/n1/daily-quotidien/171214/dq171214a-eng.htm?HPA=1>

**Literature Review**

[1] Hamid, A.J., Ahmed, T.M., Developing Prediction Model of Loan Risk in Banks Using Data Mining, Machine Learning and Applications: An International Journal, Vol. 3, No.1, (2016)

[2] Feeders, A.J., le Loux, A.J.F., Data Mining for loan evaluation at ABN AMRO: a casestudy, KDD-95 Proceedings, (1995)

[3] Memi, Demi, Assessing Credit Default Using Logistic Regression and Multiple Discriminant Analysis : Empirical Evidence from Bosnia and Herzegovina, Interdisciplinary Description of Complex Systems vol. 13, No 1, (2015)

[4] Peter Addo, Dominique Guegan, Bertrand Hassani. Credit Risk Analysis using Machine and Deep Learning models. Documents de travail du Centre d’Economie de la Sorbonne 2018.03 - ISSN : 1955- 611X 2018.

[5] Holm, Stephanie, Stegare, Sara, “Default Prediction of a Swedish Mortgage Portfolio using Logistic Regression”, (PhD diss., KTH Royal Institute of Technology, School of Engineering Sciences, (2017)

[6] Agbemava, Edinam, et. al, “Logistic Regression Analysis of Predictors of Loan Defaults By Customers of Non-Traditionnal Banks in Ghana, European Scientific Journal, vol 12. , No 1 (2016)

**Dataset and Data dictionary**

[7]<https://www.lendingclub.com/info/download-data.action>

**GitHub**

**https://github.com/jwarnold/Capstone**