MGSC 661 - Final Project

Shivangi Soni

Vivek Saahil

December 16th, 2020

Multivariate Statistics for Machine Learning

Master of Management in Analytics
McGill University

Table of Contents

1 Introduction

Peer to peer lending, also referred to as P2P lending is the practice of lending to individuals and businesses through an online platform that matches lenders with borrowers. P2P lending enables individuals to obtain loans while avoiding financial institutions as the intermediaries. One of the leading P2P lending platforms is Lending Club, which has now lent over $ 45 billion to more than 3 million customers. The way the platform operates is that borrowers can create loan listings on the website by filling in an application that would ask them for details regarding themselves and the loan requirement, such as reason of the loan, their annual income, credit history etc. Lending Club lists the loan requests within 24 hours for borrowers who apply and meet their credit policy so that the interested investors can start committing to the investment. Lending Club follows a verification process on every borrower and once all the required information is completed and verified, loans are marked as approved. During this verification process, investors can fund portions of the loans and once the loans are approved, the borrower is issued the loan, provided that there was sufficient investor commitment. Loan amounts range from $10,000 to $40,000 and investors can invest as little as $25 per loan. The company makes money through origination and service fees from borrowers and investors, respectively. Borrowers pay a one-time origination fee of 1.11% to 5% of the total loan amount while the investors pay a service fee of 1% of each payment received from a borrower [1].

The major reason that the P2P industry is successful is due to its flexibility to give instant lending to people in need. However, it does comes up with a major risk, where a borrower might not be able to repay and the investors might lose their money. The major goal of this report is to predict if a borrower will be able to repay the loan on the basis of the details provided in their application. This is done by utilizing the historical loan data from Lending Club to generate a classification model that would predict whether a loan would be fully paid back or charged off (loan is deemed unlikely to be paid back according to the original terms). This model can be leveraged by Lending Club to make data-driven decisions on whether a borrower should be lent money by investors or not. To achieve this goal, different classification models were built by using different supervised machine learning algorithms, such as Logistic Regression, Random Forest, and Gradient Boosting Machine (GBM) and the model with the highest accuracy is recommended for making decisions. Additionally, an unsupervised machine algorithm, K-means clustering is also performed to determine different segments of borrowers to gather insights and help the organization understand its user base (borrowers) better to expand the business. This report outlines the rationale and methodology that was adopted to build a classification model and generate different clusters, and also the insights that were gained from the results.

2 Data Description

2.1 Data Pre-processing

In order to build the classification model and perform clustering, the data was obtained from Kaggle[2], which derived the data from Lending Club's original website. The dataset contained information of about 2.3 million accepted loans from 2007 to 2018. The first step was to clean the data to ensure that it can be used for analysis. The cleaning process included only filtering out the data which had the loan status of either paid fully or charged off. Since it is a classification model, a new categorical variable called "paid" was created, which had values of either 1 or 0 where 1 signified the loans that were fully paid and 0 signified the ones that were charged off. Following this, the original variable for loan status was dropped. Variables with more than 25% of N/A values and the ones with just a single value (unary variables) were dropped. Variables such as url, loan ID, employment title, and zip code, which were deemed insignificant for the performance of the model were also dropped. Since this dataset contained over 300,000 observations, a subset of 20,000 observations was taken for this project.

For the purpose of the classification model, it was observed that there were some variables in the dataset that would not be available at the time of the loan application and will be obtained only after a loan is funded. Such variables were removed from the dataset since they would have resulted in an inaccurate model. Some of these variables include issue date, post charge off collection fee, payments received to date for total amount funded, months since most recent installment accounts opened etc. After this analysis, the dataset was left with a mixture of 26 numerical and categorical variables, which were used throughout the analysis and are listed in Table 1 of Appendix. All the categorical variables were converted into factor variables in R so that the program recognized them as categorical variables. Some of the variables were modified as described below:

Employment length: The employment length is a categorical variable, which had around 12 categories that ranged from having employment experience of less than one year to being employed for greater than 10 years. This variable was decided to be binned together under four dummy variables: employed for one to three years, employed for four to six years, employed for seven to ten and employed for greater than ten years where employed for less than one year variable was treated as a reference. This approach was adopted as the employment experience is usually provided in range of 1-3 years.

Address State: This variable consisted of different U.S state. Since there were over 50 categories of states, the states were grouped together according to regions: northeast, southeast, southwest and west where states in the mid-west region were taken as reference.

Year of earliest credit line: This variable included the date when the borrower's earliest reported credit line was opened. For the ease of modelling, the duration of earliest credit line was considered to be more appropriate and thus, another variable called "year since earliest credit line" was created. The variable was essentially the difference of days between 1^{st} Jan 2019 and the earliest credit date as the dataset ranges till the end of 2018.

Fico Range: Since both the upper and lower boundary ranges of borrower's FICO at loan origination were provided, another variable called fico range was created, which was the average of these two values.

2.2 Feature Selection

Feature selection is a very important aspect of modelling that is used for removing correlated variables, biases and unwanted noise. To create an accurate model, the predictors were checked for multi-collinearity. As can be seen in Figure 1 in Appendix, installments and loan amount were highly correlated to each other, which is reasonable considering if the loan originates, the loan amount decides the monthly payment owed by the borrower. Additionally, grade and interest rate were highly correlated to each other, which is reasonable as well considering the Lending Club assigns a grade on the basis of different ranges of interest rate on the loan. Hence, installment and grade were dropped. Following this, feature analysis was conducted using both Random Forest and Boruta package in R. As can be observed in Figure 2 in Appendix, Boruta showed that variables including employment length of four to six years, employment length of seven to ten years, states in southeast region, states in southwest region, and states in west region had negligible importance. These results were further confirmed by using Random Forest. As can be seen in Figure 3 in Appendix, dropping these variables would not lead to a significant reduction in the model's accuracy. Hence, these five variables were dropped.

2.3 Distribution of Variables

Prior to building the model, the box plots were created for all the numerical predictors to analyze the distribution of all predictors. The histograms were not created for categorical variables since most of these variables only had two categories. The distribution of the numerical predictors that were included in the model is summarized below:

Loan Amount: On an average, the people have asked for loan amount of $14,735 with 75% of the applications asking for loan amount of $20,000.

Annual Income: The average annual income of the borrowers is $78,470 with around 75% of the borrowers earning around $94,000.

Debt to income ratio: The average debt to income ratio of the borrowers is 18.72 with around 75% of the borrowers having the ratio as 24.48.

Number of open credit lines in the borrower's credit file: The average number of open credit lines in the borrower's credit file is around 12.

Number of derogatory public records: On an average, the borrowers had zero derogatory public records.

Total credit revolving balance: The average total credit revolving balance is $ 16,358 for the borrowers.

Revolving line utilization rate: This is the amount of credit the borrower is using relative to all available revolving credit, which on an average is 50.94%.

Total number of credit lines: The borrowers have an average of 26 credit lines in their files at the time of application.

Number of satisfactory bankcard accounts: On an average,the borrowers have around 5 satisfactory bank accounts.

Years since earliest credit line was opened: On an average, the earliest reported credit line of borrowers is around 18 years, calculated till January 1^{st} 2019.

Fico Range: The average range of the borrower's FICO at loan origination was around 692 with around 75% having fico range of around 712.

Number of public record bankruptcies: On an average, the borrowers had no public record bankruptcies.

3 Model Selection and Methodology

3.1 Model Issues

Prior to building the model, it is essential to check for model issues. Since the issue of multi-collinearity was addressed during the feature engineering process, another important issue to check was for outliers. To detect outliers, the scatter plots of all numerical variables were plotted and three major outliers were obtained. Two observations with very high amount income yet very low loan amount were discarded and one observation with the debt to income ratio of 999 income was discarded. These observations can be seen in the scatter plots found in Figure 4 and 5 in Appendix.

3.2 Classification Model Building

Once the model issues were addressed, the significant features obtained from Boruta, which were further confirmed through the variable importance plot from random forest, were used to build the classification model. The classification model aimed to predict if the loan will be full paid or charged off, hence, the target variable was a categorical variable. Three different techniques were used to build the classification model and the methodology adopted to generate these models is as follows:

Random Forest: Random forest was used to generate the model as it is one of the most powerful machine learning algorithm and due to its ability to eliminate bias. The number of trees were selected to be 500, so that the number is big enough to ensure that every input row gets predicted at least a few times. Since random forest includes building multiple bootstrapped samples,the model was not cross-validated to check its predictive power. The out of bag error of this model was obtained to be 20.13%, which means that the accuracy of the model is 79.87%.

Gradient Boosting Machine: GBM was also used to build the classification model. Since GBMs are weak learners, they can tend to minimize all error, which can result in overfitting. Hence, cross validation was used to check the out of sample performance of the GBM model. The data was split into training and test, with training being 67% of the original dataset by using createDataPartition from caret package in R. This feature was used over so that a stratified random split of the data is generated. The GBM was then performed on the trained dataset. The number of trees were chosen to be 10,000 and the number of internal nodes were chosen to be 3 to ensure that the model builds many simple trees. The model was run on the test datset and the out of sample error rate was determined to be 23.16%, hence the accuracy of the model is 76.84%.

Logistic Regression:The logistic regression was performed on the training dataset by using the glm() package with family "binomial". The model was then tested on the test dataset and the error rate was obtained to be 20.76%, which means that model has an accuracy of 79.24%.

3.3 Clustering

The features used for clustering included the top ten important numerical features from the Random Forest model, but since the feature "term" was not numeric, only nine variables were used. These nine variables were: interest rate, revolving balance, annual income, fico range, the number of open and total credit lines in the borrower's credit file, debt to income (dti) ratio, loan amount and the revolving line utilization rate (the amount of credit the borrower is using). The variables were then scaled and the original variables were dropped. The optimal number of clusters 'k', was then computed using the elbow method, which is a heuristic method used to determine the optimal number of clusters. Since the curve in figure Figure 6 in appendix shows a diminishing return at k = 6, the optimal number of clusters was determined to be 6. The average silhouette score for 6 centers was determined to be around 0.42, which is quite satisfactory.

4 Results

4.1 Model Results

As discussed in the section above, it can seen that Random Forest and Logistic regression gave very similar results in terms of accuracy. The accuracy from random forest model and logistic regression was obtained to be 79.87% and 79.24%, respectively. Although there is negligible difference in the results, Random Forest was chosen as the best model due to its robustness to overfitting and outliers. Additionally, the ability of Random Forest to handle multi-collinearity is its another advantage when compared to logistic regression. Based on the confusion matrix for the Random Forest model, it can be concluded that the model mis-classified 4026 observations out of 19,997 total observations. Out of these 4026 mis-classified observations, 3734 observations were False Positive i.e. that the model predicted that the loan would be charged off whereas it actually was paid back. 292 observations were False Negative i.e. the model predicted that the loan would be paid back, but it was actually charged-off.

For clustering, the optimal number of clusters was found to be 6, based on the elbow plot. The silhouette score is a metric used to calculate the fit of a clustering technique. Its value ranges from -1 to 1, where 1 means that the clusters are clearly distinguished and well apart, 0 means clusters are indifferent or the distance between clusters is not significant and -1 means that the clusters are assigned in the wrong way. The average silhouette score for our model was observed to be 0.42, which is reasonable for a real-life dataset since it indicates that the between cluster variation is decent and the clusters are well apart. The results of the clustering can be seen in figure 7 in Appendix.

4.2 Managerial Implications

In the context of Lending Club, if we have False Positive values, then the model might hamper chances of some individuals to get loans whereas if we have False Negative values, the model predicts that the money would be paid back but the loan is actually charged-off and the investor loses his money. Clearly, in this case False Negative is much more risky as compared to False Positive. Based on the results of confusion matrix for the Random Forest model, we can say that although our model has an accuracy of 79.87%, our model is practically much more feasible and less risky, since the number of False Negative observations are far less as compared to False Positive values. An accurate classification model can help Lending club to promote its company and attract more investors. Additionally, it can be concluded that the top five most-important predictors that determine whether a loan would be paid back or charged off are: the interest rate, the number of terms, the revolving balance, fico range (credit score) and the annual income as can be seen in the figure 3 in Appendix. Since, Lending Club charges borrowers as well, it is vital for the company to understand its consumer (borrowers) base, hence clustering was also performed. The insights generated after analyzing the result of clustering are as follows:

- **Cluster 1:** These customers tend to have the lowest revolving balance (i.e. they pay on time and clear their dues timely), have low annual income and have low credit score (fico range). They also tend to have the lowest number of open and total credit lines (i.e. number of sources where a person can take credit from), take the minimum loan amounts and have average revolving utilization (i.e. utilization of credit available). The best feature of this population is that they have lowest debt to income ratio and it is highly likely that they will pay back the small loan amounts they take. The population in this cluster represents 30.5% of the total observations as can be seen in figure 8 in Appendix and can be labelled as **Economically Weaker Section - with good personal finance management**. Thus, such customers should be encouraged to take small amount loans as the risk is minimal and the chances of timely repayment are high.

- **Cluster 2:** These customers tend to have the highest number of total and open credit lines, have an average revolving balance and an average annual income. Though they have average credit score (fico range) and low revolving utilization, these customers have high debt to income ratio and thus, they pay average interest rates. The population in this cluster represents 15.5% of the total observations and can be labelled as **Pre-Burdened average customer - People who are already in debt and are trying to pay off their previous loans**. Thus, lending high amount loans to these customers is not advised.

- **Cluster 3:** These customers tend to have the highest annual income, the highest revolving balances and have high number of total and open credit lines. They also have high credit scores (fico range), have average debt to income ratio and average revolving utilization. However, these customers tend to take high amount loans and pay very less interest to the Lending Club. The population in this cluster represents less than 1% of the total observations and can be labelled as **High income professionals - People who have high salaries, but still have average debt to income ratio**. Thus, it is recommended to increase the interest rates for these customers as they are just untapped resources and the Lending Club can generate more revenue from them.

- **Cluster 4:** These customers tend to have the highest credit scores (fico range), have low revolving balance (i.e. they pay on time) and have average income. They tend to have average open and total credit lines, have low debt to income ratio and have lowest revolving utilization. They also take lower amount loans and pay the least interest. The population in this cluster represents around 14.5% of the total observations and can be labelled as **Medium income professionals - with good personal finance management**. Thus, it is recommended to provide more loan options to attract these customers as the risk is minimal and chances of repayment are high.

- **Cluster 5:** These customers have high revolving balance and high annual income. They tend to take the highest amount of loans and have average credit score (fico range), due to which they pay high interests. They also tend to have average open and total credit lines and have the highest revolving utilization. They also tend to have the lowest debt to income ratio. The population in this cluster represents around 14.5% of the total observations and can be labelled as **High income professionals - with good personal finance management**. These are the most attractive customers, since they possess minimum risk and are the most likely to take loans of higher amounts.

- **Cluster 6:** These customers tend to have the highest debt to income ratio, have highest revolving utilization and low total and open credit lines available, due to which the pay highest interest rates. They tend to take low loan amounts, have the lowest credit scores and earn the least, in terms of annual income. The population in this cluster represents around 24% of the total observations and can be labelled as **Economically Weaker Section - who lack good personal finance management**. Thus, it is advised to proceed with caution while giving loans to this population as the chances of the loan being paid off are bleak. Lending high amounts is not recommended.

5 Conclusion

The main objective of this project was to build an accurate classification model that can classify if a borrower will fully repay the loan or loan will be charged off at the time of the loan application. Additionally, the other major objective was to understand the different segments of borrowers who have already used Lending Club to determine ways to attract more people to use the platform. The insights gained from the clustering along with the predictions that can be generated from the classification model can be of great value to Lending Club to attract more borrowers and investors.

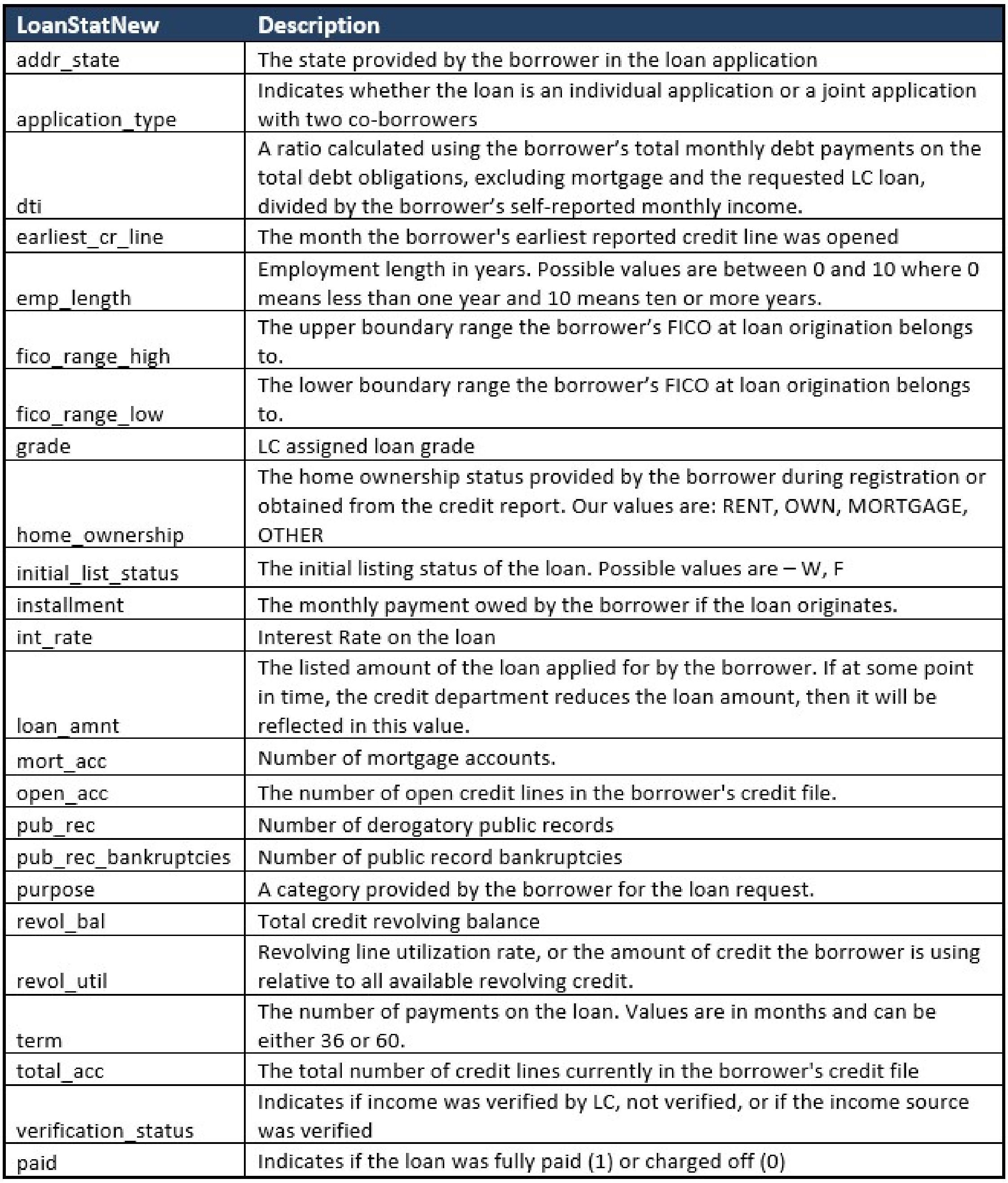
6 References

[1] FAQ. (n.d.). Retrieved December 14, 2020, from https://help.lendingclub.com/hc/en-us/sections/203810547-How-It-Works

[2] Nathan George. (2019, April 10). All Lending Club loan data. Retrieved December 1, 2020, from https://www.kaggle.com/wordsforthewise/lending-club

7 Appendix

Table 1: Data Dictionary for the variables used in the analysis



LoanStatNew	Description
addr_state	The state provided by the borrower in the loan application
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
fico_range_high	The upper boundary range the borrower’s FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower’s FICO at loan origination belongs to.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
initial_list_status	The initial listing status of the loan. Possible values are – W, F
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
mort_acc	Number of mortgage accounts.
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
total_acc	The total number of credit lines currently in the borrower's credit file
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
paid	Indicates if the loan was fully paid (1) or charged off (0)

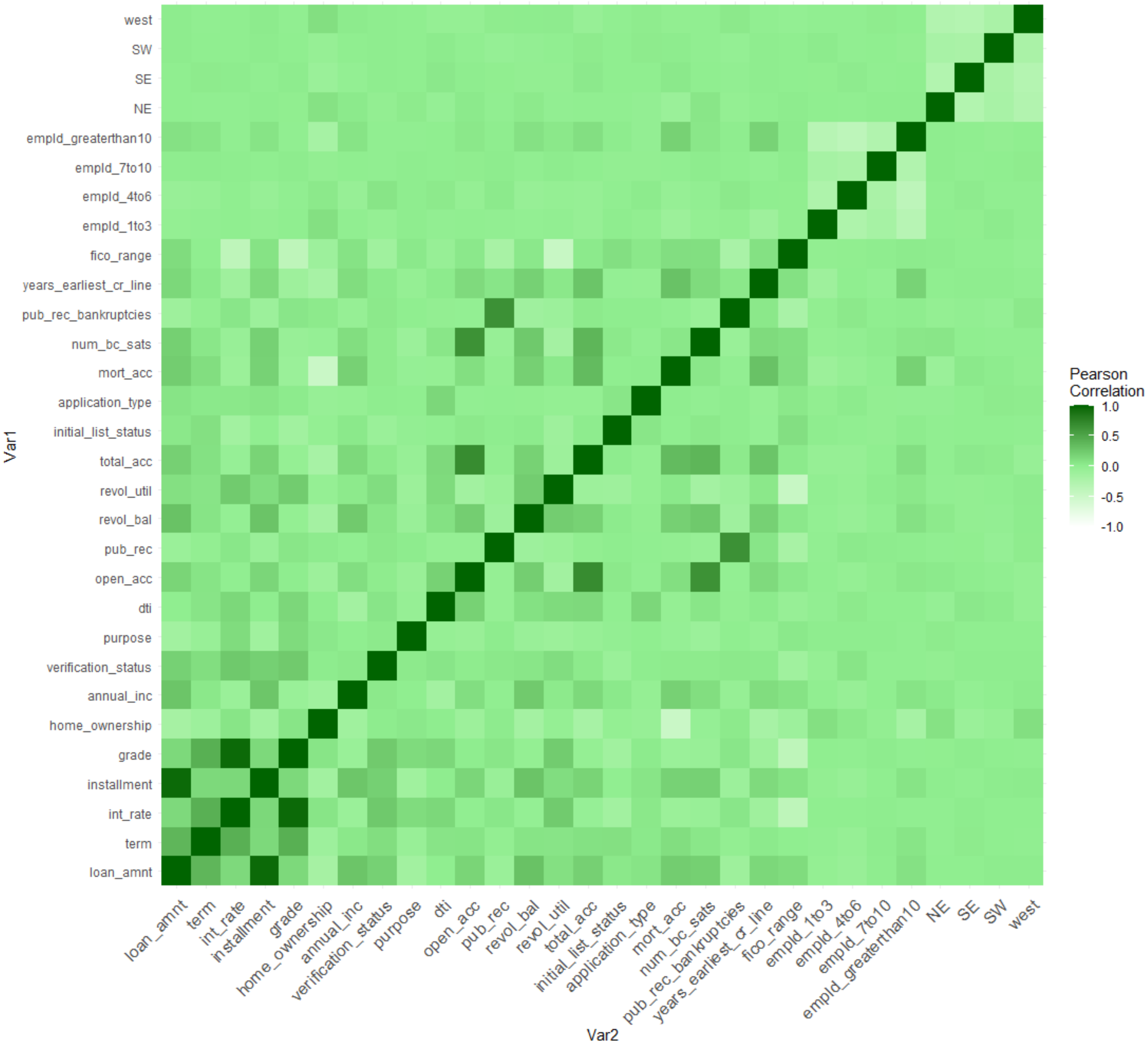


Figure 1: Correlation Matrix of all predictors

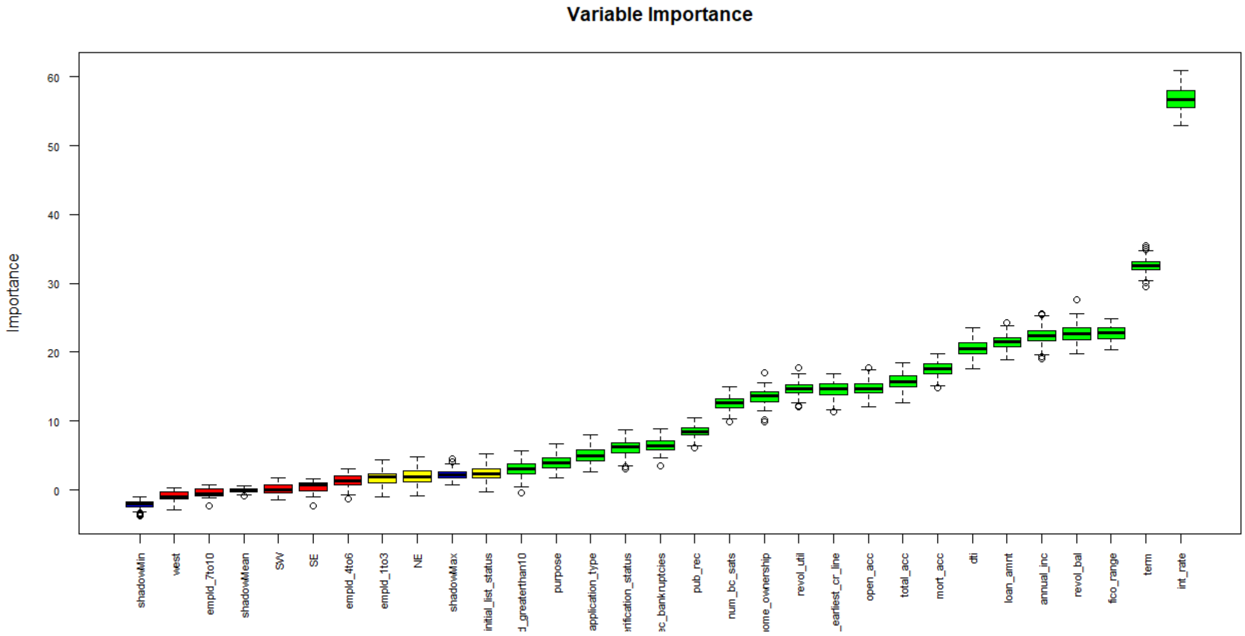


Figure 2: Significance of Predictors using Boruta

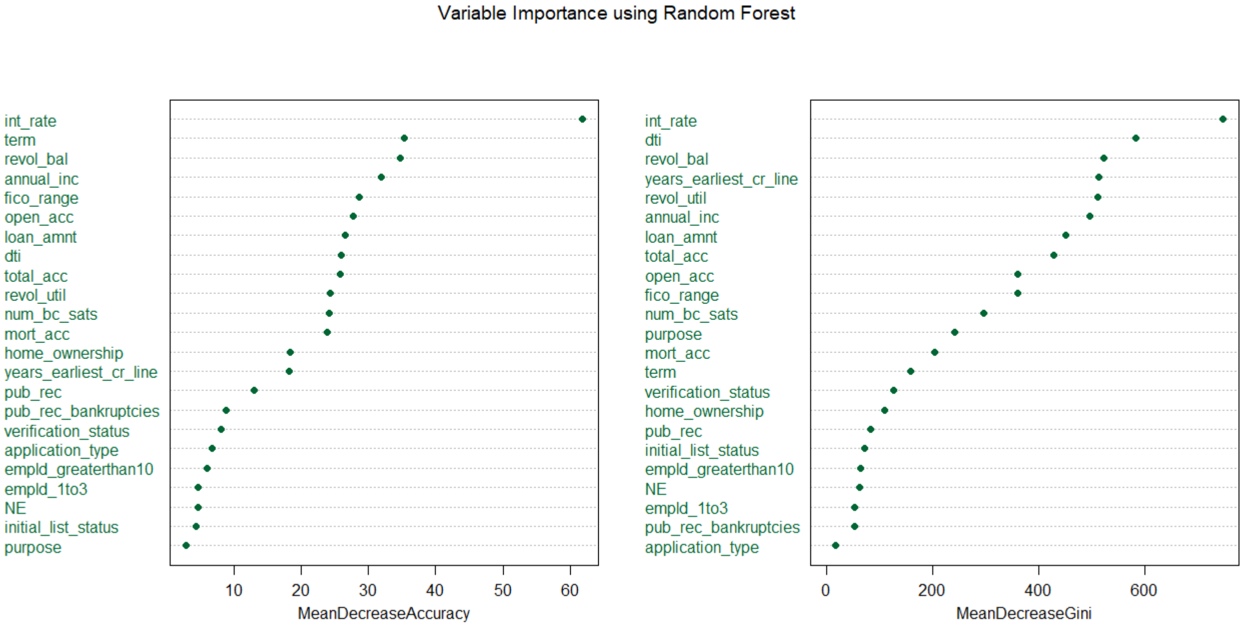


Figure 3: Variable Importance using Random Forest

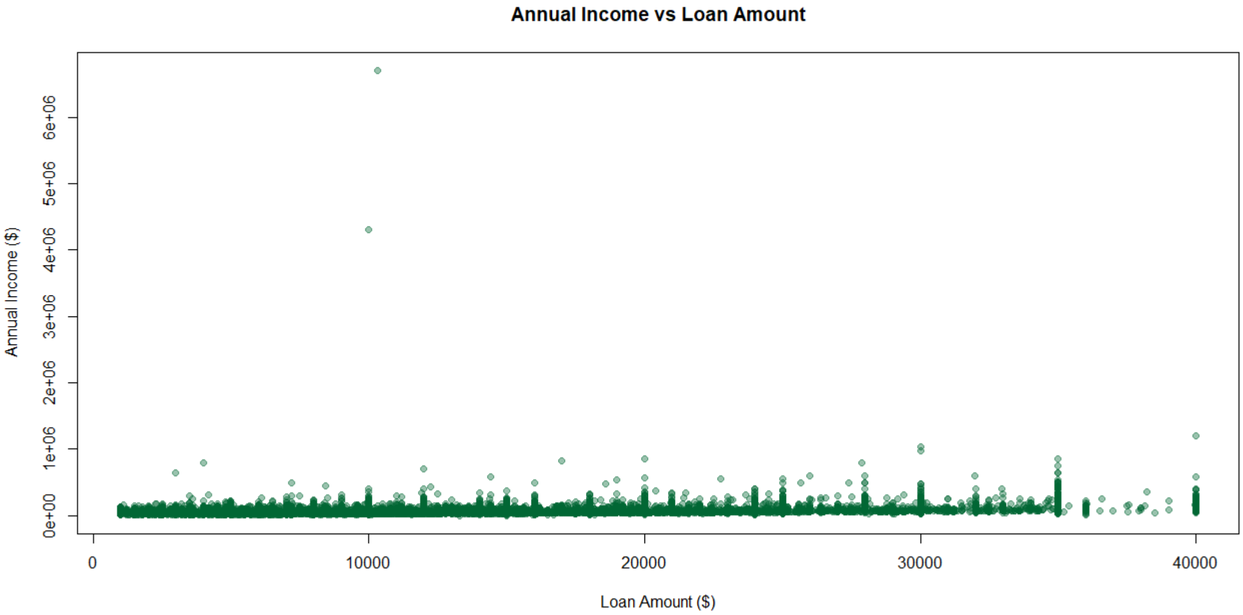


Figure 4: Annual Income vs Loan Amount plot to detect outliers

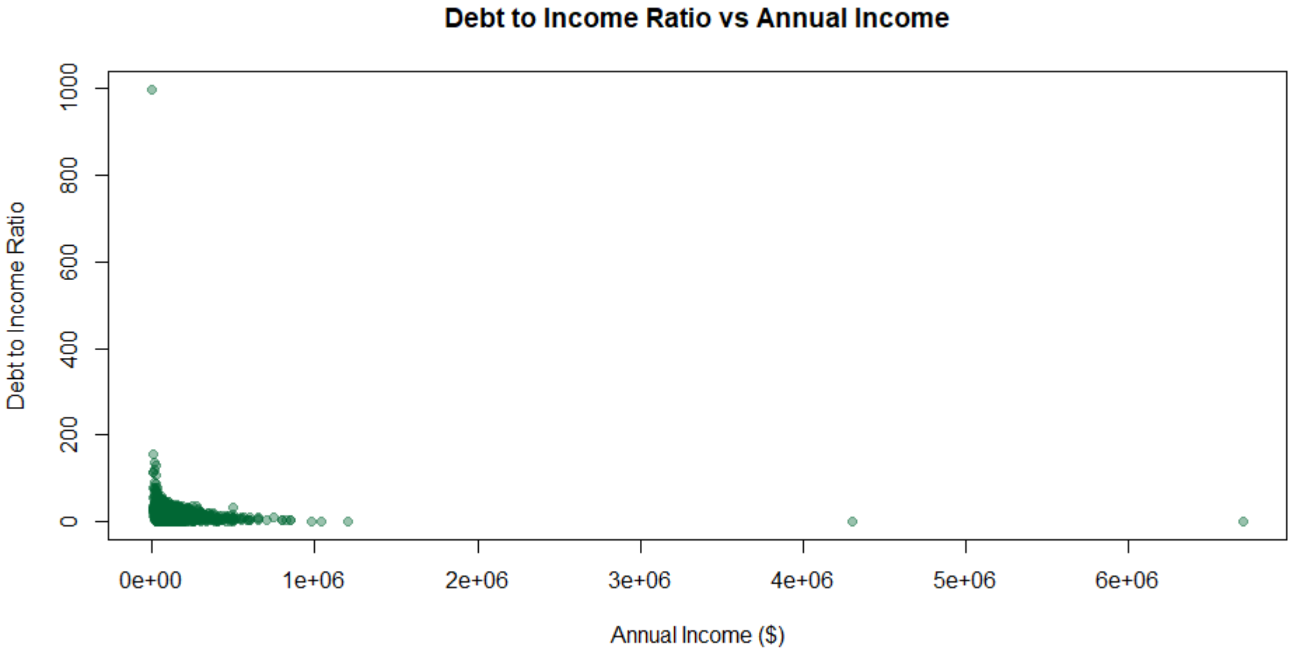


Figure 5: Debt to Income ratio vs Loan Amount plot to detect outliers

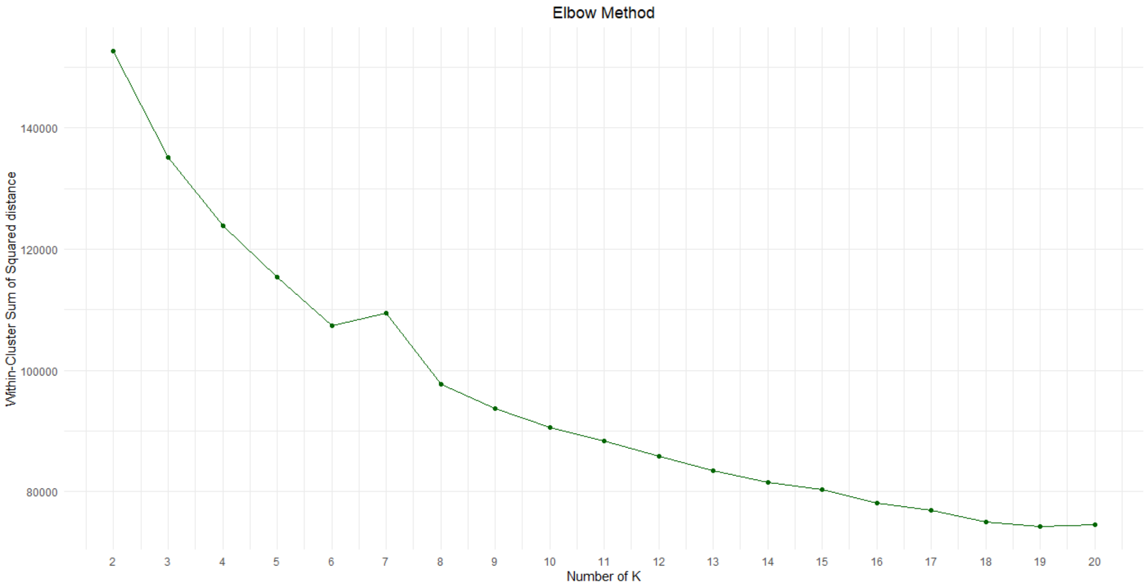


Figure 6: Elbow Method to obtain optimal number of clusters (k)

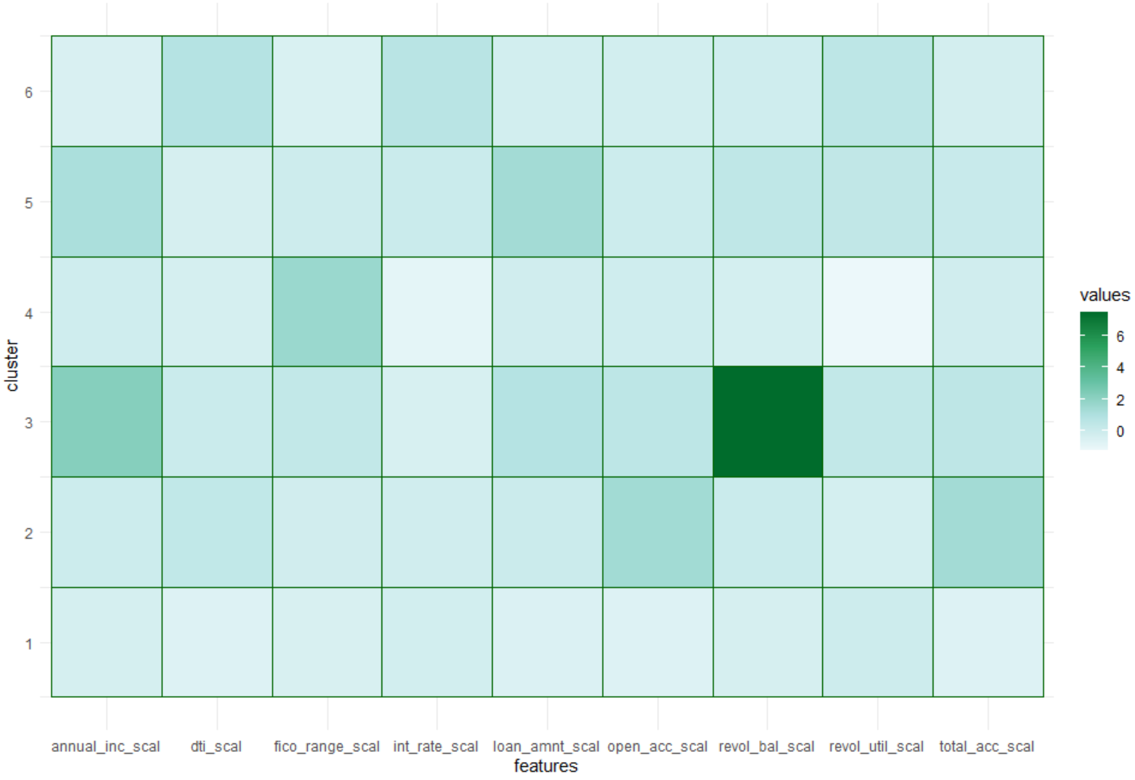


Figure 7: Heatmap showing cluster centroids for different features

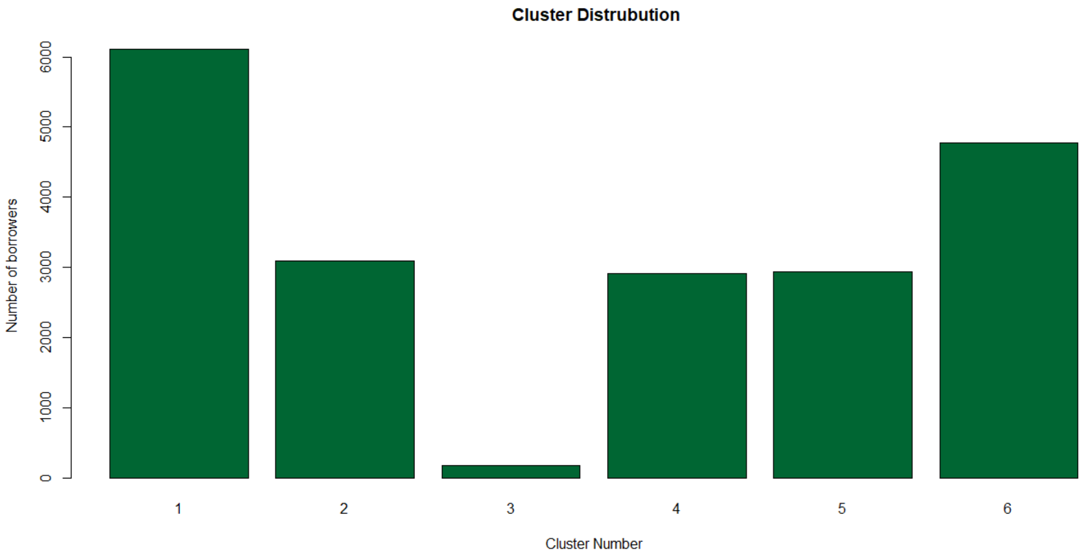


Figure 8: Distribution of different clusters