

**“CREDIT DISCRIMINATION AND DEFAULT EXTRAPOLATION”**

**SUBMITTED TOWARDS PARTIAL FULLFILMENT OF THE CRITERIA FOR AWARD OF PGPBA BY GLIM**

**REPORT BY - Capstone Group 3**

**MEMBERS -**

**Asha Murmu (BAHFEB18005)**

**Devlina Chai (BAHFEB18044)**

**Nabasish Bhattacharjee (BAHFEB18016)**

**Naveen Kumar Sambangi (BAHFEB18017)**

**Ramya Boodidha (BAHFEB18023)**

**BATCH: PGPBABI-FEB2018**

**MENTOR - Amareswar N**

****

**ABSTRACT**

**In today’s world, lending and financial institutions are going contemporary and digital at a much faster rate, generating enormous amounts of complex data every day. Data analytical tools have the ability to analyze complex and large data sets with greater accuracy and efficiency, and therefore many financial institutions are trying to adopt data analytics into their environment. Lending institutions can achieve competitive advantage by leveraging the data analytics which can play a vital role in decision-making (based on actionable insights from customer data) and in formulating strategies (to enhance marketing, influence consumer behavior and maximize its revenue opportunities).**

**There are various areas in which data analytics techniques can be used in financial sectors like customer segmentation and profitability, predicting payment default, marketing, credit analysis, ranking investments, fraudulent transactions, optimizing stock portfolios, cash management and forecasting operations, most profitable Credit Card Customers and Cross Selling.**

**Among various types of credit that lending institutions provides to its customers, loan occupy a major portion of them and hence they are obvious source of risk for them.**

**Lending Club is one such lending institute, based out of USA. Our project aims to analyze its customer data to solve the following business problems:**

**ASSESS THE RISK INVOLVED IN GRANTING A LOAN TO A NEW CUSTOMER: Risk can include complete or partial loss of principal, loss of interest, and disruption of cash flow. Hence, there is a necessity to analyze and assess the risk of the borrowers. By making use of data analytical models, they can save time, money, and resources to anticipate the risk involved.**

**CUSTOMER RETENTION: Customer retention needs more focus today in order to reduce customer attrition. For this to happen, Loyal customers (low risk borrowers obtained from customer profiling) can be rewarded by providing relevant offers (like upselling, cross selling and other retail offers) based on the insights derived from credit history.**

**FAST RECOVERY FROM NON-PERFORMING ASSETS: Data analytics enables lending institutes to perform customer profiling which helps banks in taking personalized follow-up recovery actions resulting in increased, fast recovery and reduced collection costs.**

**OPINION OF PEOPLE ABOUT LENDING CLUB: Today, social media is increasingly becoming the platform of communication for every means. We can effectively utilize this by carefully listening and monitoring consumers by leveraging Sentiment Analysis. It helps to identify customer needs, preferences & opinions on current products and services. Lending Club’s reputation can be also be monitored.**

**FORECAST CREDIT DEMAND: forecasting the amount needed by the Lending Club to meet the loan demand for the upcoming year helps in liquidity management, financial planning, identification of any shortfall of monetary balance in advance**

**KEY WORDS**

**Finance and risk analytics**

**Banking Sector**

**Lending Club**

**Loan Default Extrapolation**

**Credit Discrimination**

**Credit Risk**

**Customer Retention**

**Customer Profiling**

**Recovery**

**Non-performing assets**

**Promotional Offers**

**Loan Status**

**Loan Demand**

**Financial Planning**

**Grade/CIBIL score**

**Interest Rate**

**Tweets**

**Social Media**

R-studio

Logistic Regression

Random Forest

Support Vector Machine

K-Means Clustering

Hierarchical Clustering

Sentiment Analysis

Text Mining

Classification and Regression Tree

Accuracy

Area under the curve

Sensitivity

Specificity

Confusion Matrix

KS-coefficient

McFadden R-square

**Time Series**

**Autoregressive Integrated Moving Average**

**Mean Absolute Percentage Error**

SMOTE

**ACKNOWLEDGEMENTS**

“We are particularly grateful for our mentor N.Amareswaran for his valuable guidance provided during the planning and development of this project.”

“We would like to express our great appreciation to Dr.Srabashi Basu for her valuable and constructive feedback given to improvise the synopsis and project report”

“We would also like to extend our thanks to Great Lakes, which provided a platform to get an exposure to real world data analytics”

“We wish to thank our program manager Anwesh for his timely help whenever needed”

“And finally, last but not the least, we would like to extend our thanks to our friends and family for their constant support”

**CERTIFICATE OF ORIGINALITY AND AUTHENTICITY**

**This is to certify that the project titled “CREDIT DISCRIMINATION & DEFAULT EXTRAPOLATION” is an**

**original work of ours and is being submitted in partial fulfillment for the award of the Post-Graduation**

**(in Business Analytics and Business Intelligence)** by Great Lakes Institute of Management**.**

**We would also like to acknowledge that except for the data, entire content of this project**

**is free from any kind of plagiarism, all the codes and results are a true projection of our endeavor.**

**DATE: 3 - FEB - 2019**

**SIGNATURES:**

**Asha Murmu**

**Devlina Chai**

**Nabasish Bhattacharjee**

**Naveen Kumar Sambangi**

**Ramya Boodidha**

**CERTIFICATE OF COMPLETION**

**This is to certify that the members of the “PGP-BABI-HYD Capstone Project Group3” namely,**

**Asha Murmu, Devlina Chai, Nabasish Bhattacharjee, Naveen Kumar Sambangi, Ramya Boodidha, under**

**my guidance has successfully completed the project titled “CREDIT DISCRIMINATION & DEFAULT**

**EXTRAPOLATION”.**

**I would also like to acknowledge that this project is a reflection of the genuine efforts of my**

**mentees. It was indeed a very good interaction and experience to guide them!!**

**DATE: 3 - FEB - 2019**

**SIGNATURE:**

**Amareswaran.N**

**TABLE OF CONTENTS**

1. **CHAPTER 1 – INTRODUCTION**

PROJECT IDEA

NEED OF THE STUDY

PROJECT OBJECTIVES

DATA SOURCES

STATISTICAL TOOLS & TECHNIQUES

ABOUT LENDING CLUB & ITS BUSINESS MODEL

CHALLENGES FACED BY LENDING CLUB

LIMITATIONS OF THE PROJECT

1. **CHAPTER 2 – LITERATURE REVIEW**
2. **CHAPTER 3 – DATA CLEANING AND EXPLORATORY DATA ANALYSIS**

UNIVARIATE AND BI-VARIATE ANALYSIS

MULTI-VARIATE ANALYSIS

OUTLIER DETECTION AND HANDLING OUTLIERS

IMPUTATION OF MISSING DATA

1. **CHAPTER 4 – PREDICTIVE MODELS**

LOGISTIC REGRESSION MODEL FOR EXISTING CUSTOMERS

LOGISTIC REGRESSION MODEL FOR NEW CUSTOMER

SUPPORT VECTOR MACHINE MODEL FOR NEW CUSTOMER

RANDOM FOREST MODEL FOR NEW CUSTOMER

1. **CHAPTER 5 – CLUSTERING FOR CUSTOMER PROFILING**
2. **CHAPTER 6 – TIME SERIES FORECAST ON CREDIT DEMAND**
3. **CHAPTER 7 – SENTIMENT ANALYSIS OF TWITTER DATA OF LENDING CLUB**
4. **CHAPTER 8 – RECOMMENDATIONS AND CONCLUSIONS**
5. **BIBLIOGRAPHY**
6. **ANNEXURE**

**LIST OF TABLES AND FIGURES**

TABLE 3.1: VARIABLE NAME AND DATA TYPE

FIGURE 3.2: PERCENTAGE OF MISSING VALUES

FIGURE 3.3: CORRELATION PLOT OF ALL VARIABLES

TABLE 4.1: VARIABLE IMPORTANCE BASED ON Z-VALUES

TABLE 4.2: PERFORMANCE OF LR MODEL FOR NEW CUSTOMER

TABLE 4.3: PERFORMANCE OF SVM FOR NEW CUSTOMER

FIGURE 4.4: VARIABLE IMPORTANCE FROM RF

TABLE 4.5: PERFORMANCE OF RF FOR NEW CUSTOMER

TABLE 5.1: CUSTOMER PROFILING

TABLE 5.2: ACTIONABLE INSIGHTS FROM CLUSTERING

FIGURE 6.1: AVERAGE LOAN AMOUNT FORECAST

FIGURE 6.2: FORECAST OF NUMBER OF LOAN APPLICATIONS

TABLE 6.3: FORECAST SUMMARY FOR AVERAGE LOAN AMOUNT & NUMBER OF LOAN APPLICATIONS

FIGURE 7.1: WORD CLOUD OF LENDING CLUB TWEETS

FIGURE 7.2: CLASSFICATION OF LENDING CLUB TWEETS BASED ON EMOTIONS

FIGURE 7.3: CART MODEL OUPUT FOR TWEETS OF LENDING CLUB

**ABBREVIATIONS**

**LC - Lending Club**

**EDA - Exploratory Data Analysis**

**CORR - Correlation**

**MICE - Multivariate Imputation Via Chained Equations**

**PMM - Predictive Mean Matching**

**LR - Logistic Regression**

**AUC - Area Under Curve**

**ROC - Receiver Operating Characteristic curve**

**GLM - Generalized Linear Model**

**TPR - True Positive Rate**

**FPR - False Positive Rate**

**SVM - Support Vector Machine**

**RF - Random Forest**

**CART - Classification and Regression Tree**

**TS - Time Series**

**ARIMA - Autoregressive Integrated Moving Average**

**MAPE - Mean Absolute Percentage Error**

**SMOTE - Synthetic Minority Oversampling TEchnique**

**NPA - Non-Performing Asset**

**NRC- National Research Council Canada**

**EXECUTIVE SUMMARY**

**PROJECT FLOW DIAGRAM:**

****

**This document attempts to explain the overall default prediction for “Lending Club”, an US based financial institution Headquartered in San Francisco. Lending Club operate fully online without any branch locations, which allows to keep operating costs low and focus more resources on their customers. Lending Club provides banking system into a frictionless, transparent, and highly efficient digital experience. Along with “Default prediction”, this document also explains about future credit demand for next 18 months. Nevertheless, attempt has been made to study public perception about “Lending Club”.**

**As part of detailed study below objectives are covered.**

* **Predict the level of default risk for existing customers based on historical records of past customers. For this we have performed exploratory data analysis & checked for missing values. We have also looked for correlation between different variables & dropped variables based on missing value & variable importance.**
* **Predict the level of credit worthiness for new customers based on historical records of existing customer data. We applied different models such as logistic, SVM (Support Vector Machine) & Random Forest. Among these 3 models SVM gave the highest accuracy.**
* **In our project we have clustered the customer base of the lending club into high risk, medium risk, low risk customer so that they can be targeted accordingly for different offer or follow up program, for benefits of lending club.**
* **In our project we have applied time series to forecast the Credit demand for next one and half year using historical data. This analysis will help the lending club to take necessary actions to meet future credit demands.**
* **We have captured the sentiments about lending club from tweets of different people around the world. It will help lending club to gauge what is their brand perception, how positively or negatively people are viewing Lending club. This sentiment analysis will help lending club to take necessary measure for improving their overall performance.**

**This document will help the “Lending Club” to bridge the gap & fine tune their performance to meet growing credit demand among its borrowers & also give a good return to its Lenders.**

**CHAPTER – 1**

**INTRODUCTION**

**PROJECT IDEA - CREDIT DISCRIMINATION AND DEFAULT EXTRAPOLATION**

**Using loan data of numerous existing customers of Lending Club, in this project we want to build a Default Extrapolation model, which can help Lending Club to predict the risk (of defaulting) involved in granting a loan to a NEW CUSTOMER and take necessary precautionary actions. Credit Discrimination (clustering/profiling the existing customers) helps to reward low risk customers by providing relevant offers and to take personalized fast recovery actions on NPAs.**

**NEED OF THE STUDY:**

**• NECESSITY TO ASSESS THE RISK OF THE BORROWERS: In simple terms credit(loan) risk refers to the potential for loss due to failure of a borrower to make a payment when it is due. The risk is mainly for the lending institution and it can include complete or partial loss of principal, loss of interest, and disruption of cash flow. Having said that, there is a necessity to analyze and assess the risk of the borrowers. By making use of data analytical models, lending institutions can save their time, money, and resources to anticipate the risk involved.**

**• CUSTOMER RETENTION (REWARDING LOW RISK BORROWERS): Customer retention is another area which needs more focus today in order to reduce customer attrition. For this to happen, Loyal customers (low risk borrowers obtained from customer profiling) can be rewarded by providing relevant offers (like upselling, cross selling and other retail offers) based on the insights derived from credit history.**

**• NEED FOR FAST RECOVERY AND COLLECTION FROM NPAs (Non-Performing Assets): In most of the lending institutions, conventional process of collection and recovery lacks personalization and focused approach resulting in accumulated non-performing assets, increased cost of collection and dissatisfied customers (they sometimes chase even the customers who are regular in their payments). Data analytics enables lending institutes to perform customer profiling which helps them in taking personalized follow-up recovery actions resulting in increased, fast recovery and reduced collection costs.**

**OBJECTIVES:**

**1) MAIN OBJECTIVE – DEFAULT EXTRAPOLATION:**

**In our project we are using data analytics to build a predictive model which is developed from historical records of credit loans containing customers’ financial, demographic, psychographic, geographic information, etc. From the past credit information, predictive models can learn patterns of different loan default/delinquency ratios, and can be used to predict risk level (ex: "High risk", "Medium risk", "Low risk") involved in granting a loan to a NEW CUSTOMER.**

**2) CREDIT DISCRIMINATION (CUSTOMER PROFILING/CLUSTERING):**

**With the existing customers data, customer profiling can be done using clustering techniques. Resulting clusters can be used to explore the following possibilities:**

**• CUSTOMER RETENTION/ACQUISITION STRATEGY: low risk borrowers cluster can be rewarded with personalized retail, cash back, cross selling and up selling offers. Such kind of actions are a win-win situation for both the customers and Lending Club. These actions also help in retaining the existing customers in long run and also helps in attracting new customers.**

**• PERSONALISED FAST RECOVERY ACTIONS ON NPAs: high risk borrowers cluster can be further sub-clustered based on the delinquency period as follows:**

**Sub-cluster Due amount delayed by (No of Days)**

**NPA 90**

**Repossession 100-120**

**Valuation 120-150**

**Sale 150+**

**NPA (non-performing asset) – customer whose due payment is delayed by 90 days is considered as a non-performing asset.**

**Repossession – Lending Club will have an authority to take the possession of the property of the customers whose due amount payment is delayed by 100-120 days.**

**Valuation – customers who delayed the payment by 120-150 days fall under this. Lending Club in the process of recovery of the due amount, has the right to evaluate the value of the property of the customer.**

**Sale – if customers delayed the payment of due amount by more than 150 days then Lending Club has the right to put the property of the customer on sale, to recover the due amount.**

**Based on above mentioned customer profiling, Lending Club can take necessary personalized actions on high-risk borrowers depending on the sub-cluster they belong to. This helps in fast recovery, increases the revenue, decreases the collection costs too.**

**3) SENTIMENT ANALYSIS OF TWITTER DATA OF LENDING CLUB:**

**Today, social media covers a huge part of everyone’s life. They are increasingly becoming the platform of communication for every means. Lending institutions can effectively utilize this by carefully listening and monitoring consumers by leveraging Sentiment Analysis. It helps the Lending Club in the following ways:**

**• Identification of customer needs, preferences and opinions on the current products and services of the bank**

**• Lending Club’s reputation can be monitored**

**• Analysis of the results of marketing campaigns (if any)**

**• It can help to identify opportunities for up-selling, reduce customer churn, increase customer acquisition, improve customer retention and handle customer grievances.**

**4) FORECAST LOAN DEMAND (IN TERMS OF MONEY):**

**From customer data- “loan amount granted” and “issue date of the loan” fields can be used to forecast the amount needed by the Lending Club to meet the loan demand for the upcoming year. This forecast helps Lending Club in following ways:**

* **They can intimate the forecasted demand to their lenders and other institutions, so that they can be prepare**
* **Helps in liquidity management and financial planning**
* **Helps to identify any shortfall of monetary balance in advance**

**DATA SOURCES:**

* **https://www.kaggle.com/wendykan/lending-club-loan-data**
* **Tweets of Lending Club on twitter**

**STATISTICAL TOOLS AND TECHNIQUES:**

* **Box plots for outlier detection**
* **Correlation plots to identify the correlation among the variables**
* **Univariate and Bi-variate Analysis**
* **Imputation using MICE**
* **Logistic Regression**
* **Random Forest**
* **Support Vector Machine**
* **K-Means Clustering**
* **Hierarchial Clustering**
* **Text Mining**
* **Classification and Regression Tree**
* **ARIMA algorithm**
* **R studio**
* **Confusion Matrix**
* **AUC-ROC**

**ABOUT LENDING CLUB:**

**Lending Club is a US based and world’s largest peer-to-peer lending company, headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. Lending Club is the world's largest peer-to-peer lending platform. The company claims that $15.98 billion in loans had been originated through its platform up to December 31, 2015.**

**BUSINESS MODEL OF LENDING CLUB:**

**It helps matching borrowers with lenders who will be willing to fund their loans. Investors can search the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee. Lending Club also makes traditional direct to consumer loans, including automobile refinance transactions, through WebBank, an FDIC-insured, state-chartered industrial bank. Sometimes, loans are not funded by investors but are assigned to other financial institutions.**

**CHALLENGES/ PROBLEMS FACED BY LENDING CLUB:**

**CHALLENGE 1:** Lending Club is attracting people who don't intend to pay back loans. As awareness of alt lenders spreads, they're more likely to attract people with bad debt habits. People often take loans from marketplace lenders to pay off credit card debt, but 46% of people who borrowed from marketplace lenders ended up with at least 10% more in credit card debt after getting the loan, according to an Experian analysis cited by the Wall Street Journal. For unsecured personal loans made by all lenders this figure is 30%.

It probably doesn't have to do with the credit environment. While there are some signs of economic turmoil, charge offs are falling for banks and increasing for Lending Club.

**RESOLUTION 1:** To solve this challenge we have clustered lending clubs existing customer base of 8 lakhs customer to predict high risk customer customers, then again, we have sub classed customers who have very high risk and can result in NPA's. Further Profiling of the customers will help Lending club classify the future customers based on risk and make efficient credit decision accordingly.

Also, we have applied Logistic regression to identify the effect of different demographic, psychographic and financial variables on the Probability of default of customers to measure their riskiness and accordingly determine their creditworthiness. So Lending Club can credit loan to a customer on variable interest rate according to their credit worthiness avoid defaults and losses of the lenders.

**CHALLENGE 2:** Prediction of future demand of Credits so that adequate buyers are available for Borrowers

**RESOLUTION 2:** through time series analysis we have forecasted the monthly average of future credit demand for one and half year in advance so that accordingly Lending Club cab tune their policies to attract good lenders who will be willing to lend their money for profit on lending club.

**LIMITATIONS OF THE PROJECT:**

* There can be cases of default because of lack of integrity of the customers, who despite possessing capability to clear the loan will default. Our model cannot predict such kind of defaults.
* Predictive Model is built on the assumption that the customer data of Lending Club is subject to stable macro-economic conditions. Therefore, defaults caused due to unforeseen events like recession, geo-political crisis, natural calamities are not in the scope of prediction of our model.
* Sentiment analysis on twitter data of Lending Club can give only a high-level understanding of public opinion on Lending Club and therefore analysis at micro level is not in the scope of text mining model.
* Model does not include a variable to capture the number of dependents on annual income of the borrower. In other words, financial condition of the borrower is not accurately captured.

**CHAPTER - 2**

**LITERATURE REVIEW**

Predictive modelling for Financial Institutions:

1. https://www.aba.com/Products/Endorsed/Documents/Predictive-and-Prescriptive-Analytics-in-1) Banking\_PERF-18106-001E.PDF

Loan default predictive models:

1. <http://valleyinternational.net/index.php/theijsshi/article/download/1338/1345/>
2. <https://www.researchgate.net/publication/267864165_Loan_Default_Prediction_on_Large_Imbalanced_Data_Using_Random_Forests>
3. <ftp://ftp.repec.org/opt/ReDIF/RePEc/ami/articles/11_4_5.pdf>
4. <https://hrcak.srce.hr/file/197349>

**Clustering techniques for profiling:**

1. <https://www.omicsonline.org/open-access/profiling-of-high-risk-profiles-of-clients-in-order-to-prevent-moneylaundering-and-terrorism-jfa-1000102.pdf>

**Time Series Forecast of Bank Financials:**

1. <https://www.researchgate.net/publication/5053120_A_time_series_analysis_of_business_loans_at_large_commercial_banks>
2. <http://www.hrpub.org/download/20171030/UJAF1-12209162.pdf>

Sentiment Analysis of Twitter Data using R:

1. <http://www.iraj.in/journal/journal_file/journal_pdf/12-422-151678753913-17.pdf>

**CHAPTER - 3**

**DATA CLEANING AND EXPLORATORY DATA ANALYSIS**

**IMPORT DATA INTO R AND CHECK STRUCTURE OF THE DATASET:**

The dataset has total 8,87,379 rows & 74 columns. Below are the field details of variables of the dataset:

TABLE 3.1: VARIABLE NAME AND DATA TYPE

|  |  |  |
| --- | --- | --- |
| Sr.No | Field Name | Field Type |
| 1 | id | integer |
| 2 | member\_id | integer |
| 3 | loan\_amnt | numeric |
| 4 | funded\_amnt | numeric |
| 5 | funded\_amnt\_inv | numeric |
| 6 | term | factor |
| 7 | int\_rate | numeric |
| 8 | installment | numeric |
| 9 | grade | factor |
| 10 | sub\_grade | factor |
| 11 | emp\_title | factor |
| 12 | emp\_length | factor |
| 13 | home\_ownership | factor |
| 14 | annual\_inc | numeric |
| 15 | verification\_status | factor |
| 16 | issue\_d | factor |
| 17 | loan\_status | factor |
| 18 | pymnt\_plan | factor |
| 19 | url | factor |
| 20 | desc | factor |
| 21 | purpose | factor |
| 22 | title | factor |
| 23 | zip\_code | factor |
| 24 | addr\_state | factor |

|  |  |  |
| --- | --- | --- |
| Sr.No | Field Name | Field Type |
| 25 | dti | numeric |
| 26 | delinq\_2yrs | numeric |
| 27 | earliest\_cr\_line | factor |
| 28 | inq\_last\_6mths | numeric |
| 29 | mths\_since\_last\_delinq | numeric |
| 30 | mths\_since\_last\_record | numeric |
| 31 | open\_acc | numeric |
| 32 | pub\_rec | numeric |
| 33 | revol\_bal | numeric |
| 34 | revol\_util | numeric |
| 35 | total\_acc | numeric |
| 36 | initial\_list\_status | factor |
| 37 | out\_prncp | numeric |
| 38 | out\_prncp\_inv | numeric |
| 39 | total\_pymnt | numeric |
| 40 | total\_pymnt\_inv | numeric |
| 41 | total\_rec\_prncp | numeric |
| 42 | total\_rec\_int | numeric |
| 43 | total\_rec\_late\_fee | numeric |
| 44 | recoveries | numeric |
| 45 | collection\_recovery\_fee | numeric |
| 46 | last\_pymnt\_d | factor |
| 47 | last\_pymnt\_amnt | numeric |
| 48 | next\_pymnt\_d | factor |
| 49 | last\_credit\_pull\_d | factor |
| 50 | collections\_12\_mths\_ex\_med | numeric |
| 51 | mths\_since\_last\_major\_derog | numeric |
| 52 | policy\_code | numeric |

|  |  |  |
| --- | --- | --- |
| Sr.No | Field Name | Field Type |
| 53 | application\_type | factor |
| 54 | annual\_inc\_joint | numeric |
| 55 | dti\_joint | numeric |
| 56 | verification\_status\_joint | factor |
| 57 | acc\_now\_delinq | numeric |
| 58 | tot\_coll\_amt | numeric |
| 59 | tot\_cur\_bal | numeric |
| 60 | open\_acc\_6m | numeric |
| 61 | open\_il\_6m | numeric |
| 62 | open\_il\_12m | numeric |
| 63 | open\_il\_24m | numeric |
| 64 | mths\_since\_rcnt\_il | numeric |
| 65 | total\_bal\_il | numeric |
| 66 | il\_util | numeric |
| 67 | open\_rv\_12m | numeric |
| 68 | open\_rv\_24m | numeric |
| 69 | max\_bal\_bc | numeric |
| 70 | all\_util | numeric |
| 71 | total\_rev\_hi\_lim | numeric |
| 72 | inq\_fi | numeric |
| 73 | total\_cu\_tl | numeric |
| 74 | inq\_last\_12m | numeric |

**STRUCTURE OF DATASET: Dataset has total 2 integer variables, 23 factor variables & 49 numeric variables**

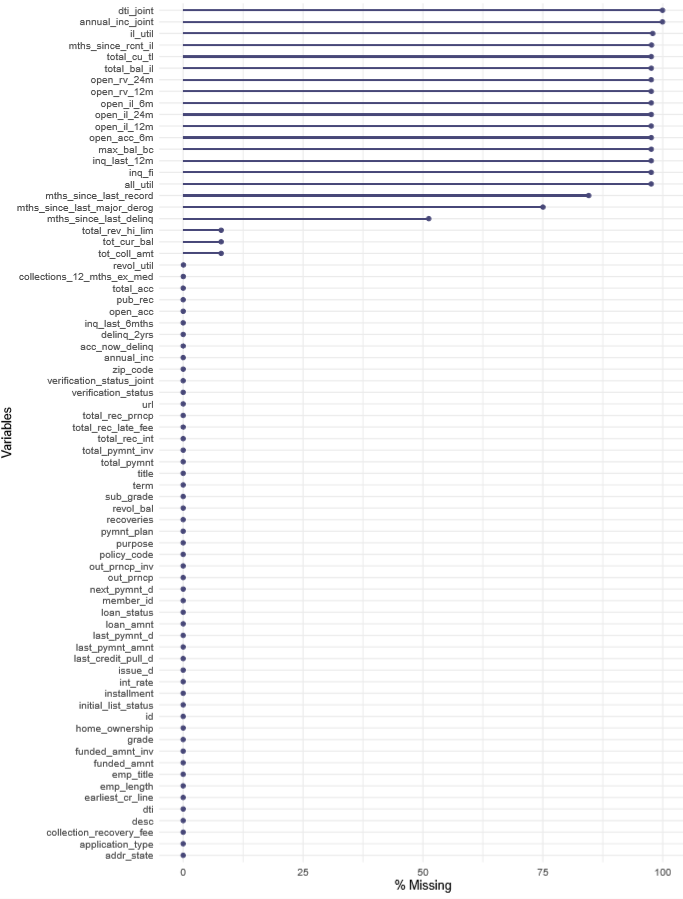
**FINDING:**

|  |  |
| --- | --- |
| Type of Variables | Count |
| Integer | 2 |
| Factor | 23 |
| Numeric | 49 |

**CHECK FOR MISSING VALUES:**

Check if there are any missing values for all the variables and below is the plot for missing value parentage:

FIGURE 3.2: PERCENTAGE OF MISSING VALUES



**FINDINGS FROM MISSING VALUE CHART:**

1. Variables having missing value > 52 %: Total 18 Variables falls under this category.
2. Variables having missing value between 0 & 52 % [ >0 & < 52]: Total 18 Variables falls under this category.
3. Variables having missing value equal to 0 %: Total 38 Variables falls under this category.

**EXPLORATORY DATA ANALYSIS and DATA CLEANING:**

Now we will perform Explanatory data Analysis (EDA) on all 74 variables. Following are the steps for EDA:

1. **UNI-VARIATE AND BI-VARIATE ANALYSIS ON ALL VARIABLES:**
2. Perform EDA on those variables which has missing value percentage > 52 %
3. Perform EDA on variables which has missing value percentage < 52%
4. Perform EDA on those variables which DO NOT have missing values
5. **CHECK FOR ZERO VARIANCE**
6. **OUTLIER DETECTION- Boxplots on the variables**
7. **MULTIVARIATE ANALYSIS – Correlation Plot**

**NOTE ON DATA CLEANING: based on the observations from every step of EDA, wherever required some variables are dropped.**

**A) UNI-VARIATE AND BI-VARIATE ANALYSIS ON ALL VARIABLES:**

**1) Perform EDA on those variables which has missing value percentage > 52 %**

**Below is the attached EDA document (of 18 variables falling under this category):**



**FINDINGS:** From the graphs of all **18 variables** we could see that the number of records is very less (when frequency is observed in univariate graphs) indicating nearly **97% of the missing values**. Hence these variables are not included in further process of modelling.

**DROPPING VARIABLES:** For better function of the model we have kept the missing percentage cutoff at 52% and hence we are DROPPING these variables for which missing value percentage is > 52%.

Below are the 18 dropped variables**:**

all\_util annual\_inc\_joint dti\_joint

il\_util inq\_fi inq\_last\_12m

max\_bal\_bc mths\_since\_last\_major\_derog mths\_since\_last\_record

mths\_since\_rcnt\_il open\_acc\_6m open\_il\_12m

open\_il\_24m open\_il\_6m open\_rv\_12m

open\_rv\_24m total\_bal\_il total\_cu\_tl

After dropping variables having missing values > 52%, we are left with 56 variables.

**2) Perform EDA on variables which has missing value percentage < 52%**

**Below is the attached EDA document (of 18 variables falling under this category):**

****

**MAJOR FINDINGS for those variables which has missing value percentage < 52%:**

From the univariate and bivariate analysis of the variables, following are the inferences drawn, which are used further in model building:

|  |
| --- |
| * From bivariate graphs for variables it can be concluded that for most of the variables “Current” factor level is most prominent: * Payments which are yet to be received by Lending Club like - total collection amount, total current balance is maximum for the “Current” customers. * Delinquency is also high for customers having loan status as “current”, this is evident from fields such as delinq\_2yrs, mths\_since\_last\_delinq, acc\_now\_delinq etc. * Univariate graph of emp\_length shows that maximum people who has applied loan are having employment experience more than 10 years. * Univariate graph of mths\_since\_last\_delinq shows that maximum people who has taken loan has 5-10 open credit lines. * Univariate graph of acc\_now\_delinq shows that maximum people who has taken loan has zero delinquent accounts. |

**3) Perform EDA on variables which DO NOT have missing values**

**Below is the attached EDA document (of 38 variables falling under this category):**

****

**MAJOR FINDINGS for those variables which do not have missing values**

|  |
| --- |
| * The graphs of total\_pymnt, total\_pymnt\_inv, total\_rec\_prncp show that total payment received till date and total received principle is maximum from the “fully paid” and “current” customers. * As the name suggest recoveries, collection\_recovery\_fee will be only for charged-off customers, graphs also indicate the same thing. * The bivariate graph of total\_rec\_late\_fee shows that total received late fees is distributed significantly among all the loan\_status factor levels, indicating that most of the customers who fully paid and who are currently paying also paid some of their instalments late receiving a penalty. * last\_pymnt\_amnt graph shows a “Fully paid” factor level is prominent indicating that the last payment amount received by lending club is maximum from the customers who fully paid the loan * univariate graph of policy \_code, application\_type show that all the customers’ policy code is “1” and loan application type is “INDIVIDUAL”, which results in zero variance for these variables, hence they need not be included in the model building. * Univariate graph of verification\_status\_joint shows that 99% of the customer records have a MISSING VALUE for this variable, hence this variable can be discarded. * Loan\_amnt is slightly right skewed * Funded\_amnt and funded\_amnt\_inv is also slightly right skewed * Most of the loan is termed for 36 months rather than 60 months * We have majority of the loans grade at either B or C, and loan with grade E, F, G is very less * In the dataset we have most of the loans as current around 150k and around 100k as fully paid * Out of the loan issued around 400k have mortgages and 100 k customers have own home. * Payment plan has all has no with 0 standard deviation, so we may remove this column while modelling * Most of the customers around 520k customers have the purpose of Debt consolidation. * Zip code has been masked, last 2 digits is replaced by XX, so we wouldn't use the same in our model instead we will use addr\_state * Most of the lending club customers are from California (120k). * Most of the loans in our dataset has been issued in2015 (400k customers approx.) * Url and Desc column are very descriptive column which can be used for text mining and can be omitted for model building in LR |

**B) CHECK FOR ZERO VARIANCE:**

Now we check for zero variance & based on that we dropped some variables.

**FINDINGS**: application\_type, policy\_code, pymnt\_plan **have zero variance**

In addition to that we have also removed few other variables (Desc, title, url, zip\_code, verification\_status\_joint, dti\_joint, annual\_inc\_joint) which does not have any significance. All records are of type “individual” which means application type “joint” is not present therefore we can drop the variables related to type “joint”.

**DROPPING VARIABLES:** Desc, title, url, zip\_code, pymnt\_plan, policy\_code, application\_type, verification\_status\_joint, dti\_joint, annual\_inc\_joint

**CONVERTING FACTOR VARIABLES INTO NUMERIC:**

We have few factor variables which we need to convert to numeric, below are the details of factor variables which needs conversion into numeric.

**Variable name: “term”** | Description: The number of payments on the loan. Values are in months and can be either 36 or 60:We replaced the string “months” & then converted the value into 1 & 0s. If it is 60 then we populated with 1s & if it is 36, we populated with 0s.

**Variable name: “emp\_length”** | Description: 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:For this variable we replace “years” with space & then perform below replacements.

< 1 = 0

1-9 = 1-9

10 & above = 10

**Variable name: “verification\_status”** | Description:Status of the loan application

There are 3 possible value “Verified”, “Source Verified” & “Not Verified”.

We first combine labels “verified” & “Source Verified” into “Source Verified” & then replaced the labels with numeric values 0s & 1s.

**Variable name: “initial\_list\_status”** | Description**:** The initial listing status of the loan. Possible values are – W, F

We replace labels “W” & “F” with values 1 & 0 respectively.

**Variable name: “loan\_status”** | Description:status of the loan currently

There are multiple labels, we grouped loan\_status labels as below, based on the level of risk of default:

1. LOW-RISK

Current, Issued, Fully Paid, does not meet the credit policy. Status: Fully Paid = 1

1. HIGH RISK

In Grace Period, Late (16-30 days), Late (31-120 days), Charged Off, Default, does not meet the credit policy. Status: Charged Off = 2

**Variable name: “issue\_d”** | Description:The month which the loan was funded

For this variable we have value in below format.

Month-Year [e.g. Jan-15]

We have converted this date into dd-mm-yyyy [e.g. 01-01-2015]

Same way we have handled other variables such as “**earliest\_cr\_line”,” last\_pymnt\_d” & “next\_pymnt\_d”**

**C) OUTLIER DETECTION USING BOX PLOTS:**

Now we have performed one more round of explanatory data analysis (EDA) by plotting bivariate & univariate boxplots.

**Below is the boxplot analysis document attached:**



**FINDINGS FROM BOXPLOT ANALYSIS:**

1. Bivariate box-plot analysis: All the numeric variables have outliers
2. Univariate box-plot analysis: From univariate box plots of all the variables, following observations are made:

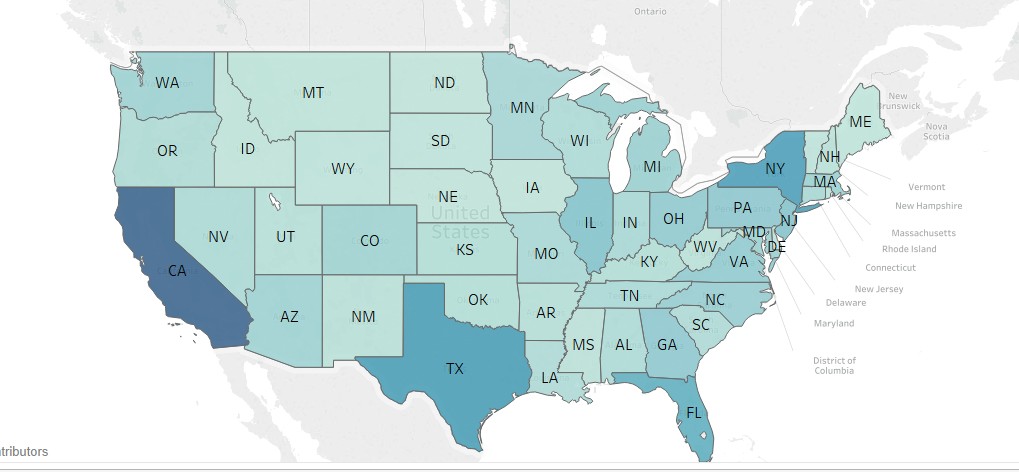
* **Variables - loan\_amnt, funded\_amnt, funded\_amnt\_inv, term, grade, emp\_length, home\_ownership, verification\_status, initial\_list\_status do not have outliers**
* **Remaining all other variables have outliers**

**ADDITIONAL EDA:**

**FINDINGS:**from bar plot of Loan amount against each state we can conclude that maximum loan is sanctioned in state California (CA)

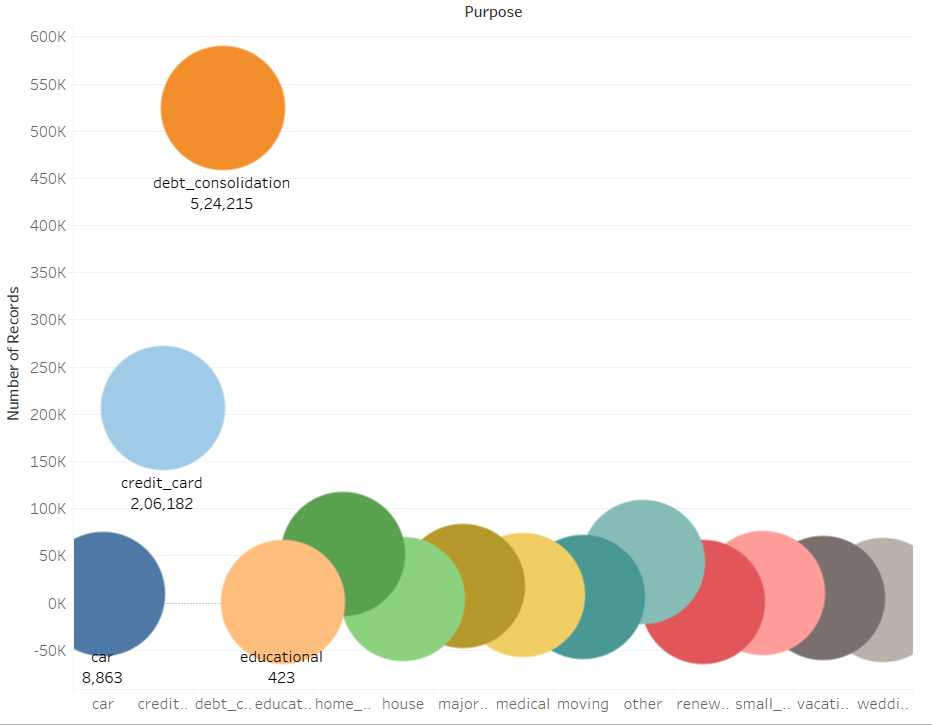


**Geo plot for state wise Loan disbursement:**



**FINDINGS:**From Geo plot also, it is evident that maximum loan is disbursed from state California.

**Purpose of Loan*:***

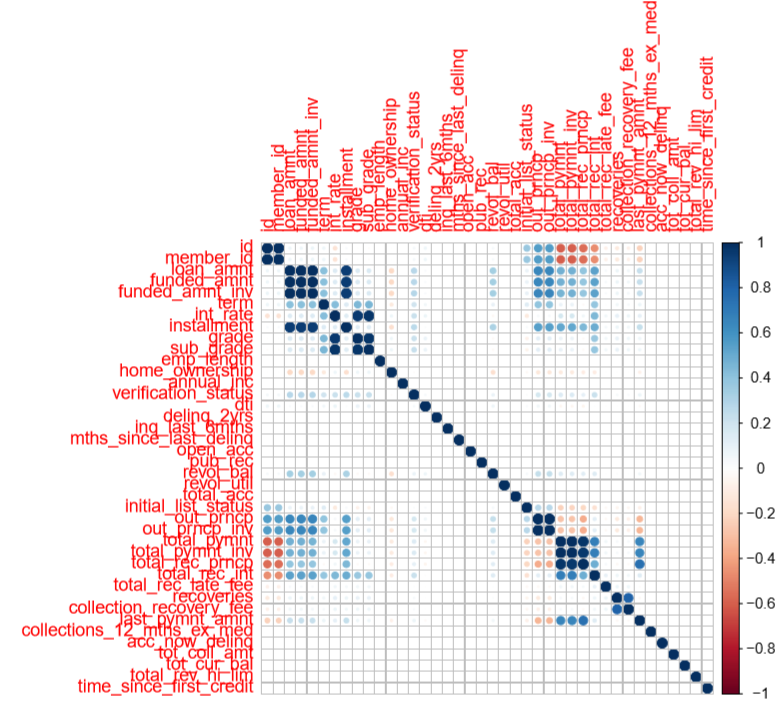


**FINDINGS:** From this plot we can say very easily that maximum loan has been disbursed for the purpose “debt consolidation”

**D) MULTI-VARIATE ANALYSIS USING CORR PLOT:**

Corrplot: Now we draw corrplot for given dataset, below is the plot:

FIGURE 3.3: CORRELATION PLOT OF ALL VARIABLES



**FINDINGS FROM MULTI-VARIATE ANALYSIS USING CORR PLOT:**

* loan\_amnt, funded\_amnt, funded\_amnt\_inv have very high correlation (0.9) with each other. Therefore funded\_amnt, funded\_amnt\_inv can be dropped before going for imputation and predictive modelling.
* out\_prncp, out\_prncp\_inv are also highly correlated with other. Hence, out\_prncp\_inv can be dropped before going for imputation and predictive modelling.
* sub\_grade and int\_rate have high correlation (0.7) with each other. For modelling purpose, subgrade variable is dropped.
* total\_pymnt, total\_pymnt\_inv are also highly correlated. Hence, total\_pymnt\_inv can be dropped before going for imputation and predictive modelling.

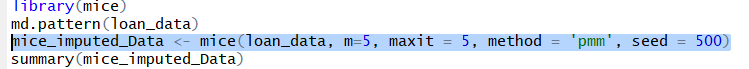
**DROPPING VARIABLES:** sub grade, fundedAmt, fundedAmt\_inv, out\_prncp\_inv, total\_pymnt\_inv

**HANDLING OUTLIERS:**

* Handling outliers for some variables might result in loss of data patterns and insights
* for example: defaulted customers will have higher number of delinquencies, which might be shown in the box plot as outliers, if these outliers are handled then it results in loss of defaulted customers’ data which is very important for the model
* Therefore, only few variables for which handling outliers will not affect the model output are selected for outlier treatment, and they are as follows:
* dti, revol\_util, out\_prncp
* **Outliers of the variables dti, revol\_util, out\_prncp are replaced with NAs and then IMPUTED using mice package.**

**IMPUTATION:**

* We used **MICE (Multivariate Imputation via Chained Equations)** package for imputation of missing data, and its good approach to create multiple imputations as compared to a single imputation (such as mean), and it will take care of uncertainty in missing values.
* MICE assume that the missing data are Missing at Random (MAR), which means that the probability that a value is missing depends only on observed value and can be predicted using them.
* For example: Suppose we have X1, X2….Xk variables. If X1 has missing values, then it will be regressed on other variables X2 to Xk. The missing values in X1 will be then replaced by predictive values obtained. Similarly, if X2 has missing values, then X1, X3 to Xk variables will be used in prediction model as independent variables. Later, missing values will be replaced with predicted values.
* By default, linear regression is used to predict numeric missing values.



**SUMMARY OF EDA AND DATA CLEANING**

|  |  |
| --- | --- |
| **VARIABLE NAMES** | **EDA/ DATA CLEANING/ TRANSFORMATIONS/ IMPUTATION/OUTLIER HANDLING OF VARIABLES** |
| all\_util, annual\_inc\_joint, dti\_joint, il\_util , inq\_fi, inq\_last\_12m, max\_bal\_bc ,mths\_since\_last\_major\_derog , mths\_since\_last\_record , mths\_since\_rcnt\_il open\_acc\_6m, open\_il\_12m, open\_il\_24m, open\_il\_6m, open\_rv\_12m, open\_rv\_24m,total\_bal\_il, total\_cu\_tl | All these 18 variables have missing values more than 50% of the data, hence all these variables are dropped |
| application\_type  verification\_status\_joint, dti\_joint, annual\_inc\_joint | This variable has ‘0’ variance and hence can be dropped (all records take “SINGLE” type).  And hence variables related to joint are dropped |
| zip\_code | zip\_code contains partially masked data and hence dropped |
| Desc, title, url , id , member\_id | These variables DO NOT add any value to the modelling and hence dropped |
| term, emp\_length, verification\_status, initial\_list\_status,  loan\_status | All these variables are converted into numeric |
| funded\_amnt, funded\_amnt\_inv | These two variables have high correlation with loan\_amnt and hence dropped |
| out\_prncp\_inv | This variable is dropped since it has high correlation with out\_prncp |
| sub\_grade | This variable is dropped since it has high correlation with int\_rate |
| total\_pymnt\_inv | This variable is dropped since it has high correlation with total\_pymnt |
| dti, revol\_util, out\_prncp | These variables have outliers and they are replaced with NAs and then imputed using MICE package |
| emp\_title,emp\_length, annual\_inc, earliest\_cr\_line, inq\_last\_6mths, mths\_since\_last\_delinq, open\_acc, pub\_rec, revol\_util, total\_acc, acc\_now\_delinq, total\_coll\_amt, total\_cur\_bal, tot\_rev\_hi\_lim | All these variables have missing values and hence imputed using MICE package |

**CHAPTER - 4**

**PREEDICTIVE MODELS**

**LOGISTIC REGRESSION MODEL FOR EXISTING CUSTOMERS**

**PURPOSE:** in the process of building the model, the dataset is partitioned into train data and test data and the model is built on train data. So, the purpose of this model is to predict the risk level of default for the customer records in test data based on the model built from train data. Here, we are trying to predict the level of default risk for existing customers based on historical records of past customers.

Before we proceed for modelling, we can drop two more variables “id”, “member\_id”. After removing these variables, we are left with **36 variables.**

**LIST OF VARIABLES USED FOR THE MODEL BUILDING:**

[1] "loan\_amnt" "term" "int\_rate"

[4] "installment" "grade" "emp\_length"

[7] "home\_ownership" "annual\_inc" "verification\_status"

[10] "loan\_status" "purpose" "dti"

[13] "delinq\_2yrs" "inq\_last\_6mths" "mths\_since\_last\_delinq"

[16] "open\_acc" "pub\_rec" "revol\_bal"

[19] "revol\_util" "total\_acc" "initial\_list\_status"

[22] "out\_prncp" "total\_pymnt" "total\_rec\_prncp"

[25] "total\_rec\_int" "total\_rec\_late\_fee" "recoveries"

[28] "collection\_recovery\_fee" "last\_pymnt\_amnt" "collections\_12\_mths\_ex\_med"

[31] "acc\_now\_delinq" "tot\_coll\_amt" "tot\_cur\_bal"

[34] "total\_rev\_hi\_lim" "time\_since\_first\_credit" “addr\_state”

**DEPENDENT and INDEPENDENT VARIABLES: loan\_status is dependent variable and remaining all others are independent variables**

loan\_status

loan\_status = 0 (low risk)

loan\_status = 1 (high risk)

**SPLIT THE DATA INTO TRAIN AND TEST DATA SETS:**

Now we will divide the data into train & test in **70:30** ratio. After dividing data, we are left with below number of records in Train & test sample.

Train -> 621166 ; Test -> 266213

**BUILD LOGISTIC REGRESSION MODEL:**

Now as part of next step we will build the binomial logistic regression model. In order to perform the same, we will use **“glm” function**

logit = glm(loan\_status ~ ., data=Training, family= binomial)

**OVERALL SIGNIFICANCE/VALIDITY OF THE MODEL (log likelihood test):**

Now let us do likelihood ratio test using function “lrtest”, upon doing this test we got **p value < 2.2e-16**. This means p value is highly significant. It implies that the null hypothesis of "all the coefficient estimates are Zero" is rejected and we conclude that at least one coefficient estimate is non-Zero.

**Inference: overall model significance is high**

**MEASURE OF FIT: CALCULATE McFadden R-Square:**

|  |  |  |
| --- | --- | --- |
| Measure | Value | Description |
| llh | -7.325561e+04 | Log-likelihood from the fitted model |
| llhNull | -1.668736e+05 | Log-likelihood from intercept-only restricted model |
| G2 | 1.872360e+05 | Minus 2 times the difference in the log-likelihoods |
| McFadden | **5.610114e-01** | **McFadden’s pseudo r-squared** |
| r2ML | 2.602379e-01 | Maximum Likelihood pseudo r-squared |
| r2CU | 6.260667e-01 | Cragg and Uhler’s pseudo r-squared |

**Interpretation of McFadden’s r square:**

McFadden’s R squared for the model is 0.561, and it is known that if McFadden’s R square value is greater than 0.5, then it indicates an excellent goodness of fit for the model.

**Inference: Mc-fadden R-sq of 56.1 % indicates that** **goodness of fit is EXCELLENT and very good predictive ability**.

It also indicates- 56.1 % of the uncertainty of the intercept only model has been explained by the model we have built.

**SUMMARY OF LOGIT FUNCTION:**

* Summary of logit function gives the **coefficient estimate values, standard errors and p-values** from which **significance of the variables can be determined**.
* Variables for which p-value is less than 0.05 are considered statistically significant.
* Variables for which p-value is more than 0.05 are considered statistically Insignificant.

**LIST OF STATISTICALLY SIGNIFICANT VARIABLES FROM SUMMARY**

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.703e+00 1.897e-01 -24.786 < 2e-16 \*\*\*

loan\_amnt 2.479e-04 6.817e-06 36.357 < 2e-16 \*\*\*

term 6.140e-01 4.136e-02 14.846 < 2e-16 \*\*\*

int\_rate 1.756e-01 6.467e-03 27.153 < 2e-16 \*\*\*

installment 7.785e-03 2.234e-04 34.854 < 2e-16 \*\*\*

gradeB -2.395e-01 4.070e-02 -5.884 4.00e-09 \*\*\*

gradeC -6.133e-01 5.327e-02 -11.513 < 2e-16 \*\*\*

gradeD -1.027e+00 7.054e-02 -14.557 < 2e-16 \*\*\*

gradeE -1.555e+00 8.798e-02 -17.679 < 2e-16 \*\*\*

gradeF -2.353e+00 1.120e-01 -21.017 < 2e-16 \*\*\*

gradeG -2.735e+00 1.366e-01 -20.028 < 2e-16 \*\*\*

emp\_length -5.455e-03 2.227e-03 -2.450 0.014302 \*

home\_ownershipRENT 5.251e-02 2.060e-02 2.549 0.010798 \*

annual\_inc -1.524e-06 2.566e-07 -5.939 2.86e-09 \*\*\*

purposecredit\_card 2.149e-01 8.389e-02 2.562 0.010417 \*

purposedebt\_consolidation 3.262e-01 8.256e-02 3.952 7.76e-05 \*\*\*

purposehome\_improvement 2.664e-01 8.812e-02 3.023 0.002503 \*\*

purposehouse 3.850e-01 1.313e-01 2.932 0.003362 \*\*

purposemedical 2.723e-01 1.040e-01 2.617 0.008874 \*\*

purposesmall\_business 3.626e-01 1.017e-01 3.567 0.000362 \*\*\*

addr\_stateND -1.345e+00 5.803e-01 -2.317 0.020482 \*

addr\_stateNE -8.166e-01 3.247e-01 -2.515 0.011892 \*

addr\_stateSC -3.498e-01 1.749e-01 -2.000 0.045488 \*

dti 7.154e-03 1.038e-03 6.892 5.50e-12 \*\*\*

inq\_last\_6mths 7.815e-02 7.260e-03 10.764 < 2e-16 \*\*\*

mths\_since\_last\_delinq -4.153e-03 5.809e-04 -7.150 8.67e-13 \*\*\*

pub\_rec -3.716e-02 1.385e-02 -2.684 0.007274 \*\*

revol\_util 9.642e-04 4.402e-04 2.190 0.028507 \*

total\_acc 7.701e-03 9.438e-04 8.159 3.37e-16 \*\*\*

initial\_list\_status -1.416e-01 1.613e-02 -8.779 < 2e-16 \*\*\*

out\_prncp -5.414e-04 3.473e-06 -155.894 < 2e-16 \*\*\*

last\_pymnt\_amnt -5.123e-04 1.327e-05 -38.607 < 2e-16 \*\*\*

tot\_cur\_bal -3.476e-07 8.758e-08 -3.970 7.20e-05 \*\*\*

total\_rev\_hi\_lim -1.605e-06 7.719e-07 -2.079 0.037650 \*

time\_since\_first\_credit -1.610e-05 3.220e-06 -5.000 5.74e-07 \*\*\*

**INTERPRETATION OF COEFFICIENTS FROM SUMMARY:**

* Interpretation of the coefficient of the “loan\_amnt”:

The coefficient of loan\_amnt is 2.479e-04, which indicates that for one unit increase in loan\_amnt (holding all other variables constant), the log odds of defaulting over not defaulting is 2.479e-04

* Similar interpretations can be made with the coefficients of all other variables.
* Standard error is very less for all the variables

**ODDS AND PROBABILITES:**

**Following is the list of variables for which odds ratio >1: - practically significant variables**

**VARIABLE OODS-RATIO**

loan\_amnt 1.002

term 1.84

int\_rate 1.19

installment 1.00

home\_ownershipOWN 1.01

home\_ownershipRENT 1.05

purposecredit\_card 1.23

purposedebt\_consolidation 1.38

purposehome\_improvement 1.30

purposehouse 1.46

purposemajor\_purchase 1.16

purposemedical 1.31

purposemoving 1.19

purposeother 1.10

purposerenewable\_energy 1.30

purposesmall\_business 1.43

purposevacation 1.17

purposewedding 1.03

addr\_stateAL 1.10

addr\_stateAR 1.07

addr\_stateAZ 1.03

addr\_stateCA 1.10

addr\_stateDE 1.03

addr\_stateFL 1.10

addr\_stateHI 1.20

addr\_stateIA 3.09

addr\_stateID 4.28

addr\_stateIN 1.00

addr\_stateLA 1.06

addr\_stateMA 1.09

addr\_stateMD 1.10

addr\_stateMN 1.02

addr\_stateNC 1.05

addr\_stateNJ 1.08

addr\_stateNM 1.05

addr\_stateNV 1.37

addr\_stateNY 1.14

addr\_stateOR 1.08

addr\_statePA 1.09

addr\_stateRI 1.01

addr\_stateSD 1.17

addr\_stateTN 1.05

addr\_stateUT 1.01

addr\_stateVA 1.11

dti 1.00

**delinq\_2yrs 1.01**

inq\_last\_6mths 1.08

revol\_util 1.00

total\_acc 1.00

**total\_pymnt 1.865897e+07**

**recoveries 1.983156e+05**

**Though the variables delinq\_2yrs, total\_pymnt, recoveries are not significant statistically, odds ratio for these variables is greater than 1, indicating that these variables are practically significant for the model.**

**Interpreting odds ratio and probability:**

* **Odds ratio for loan\_amnt is 1.0002 and probability is 0.50**

Keeping all other variables constant, if loan amount increases by one unit, the odds of defaulting become 1.0002 times the odds of not defaulting and the probability of defaulting is 50 %.

* **Odds ratio for term is 1.84 and probability is 0.64**

Keeping all other variables constant, if term increases by one unit, the odds of defaulting become 1.84 times the odds of not defaulting and the probability of defaulting is 64 %.

* **Odds ratio for int\_rate is 1.19 and probability is 0.54**

Keeping all other variables constant, if interest rate increases by one unit, the odds of defaulting become 1.19 times the odds of not defaulting and the probability of defaulting is 54 %.

* **Odds ratios and probabilities of all other variables can be interpreted in a similar way as above.**

**PREDICTION ON TRAIN DATA SET:**

The model built is used on the train data to predict the probabilities and class level of loan\_status and then confusion matrix, accuracy, AUC-ROC, Mis-Classification error, sensitivity and specificity are derived for the train data.

**CONFUSION MATRIX FOR TRAIN DATA:**

horizontal class labels represent predicted classes and vertical class labels represent actual classes

**0 1**

**0 570641 15782**

**1 3377 31366**

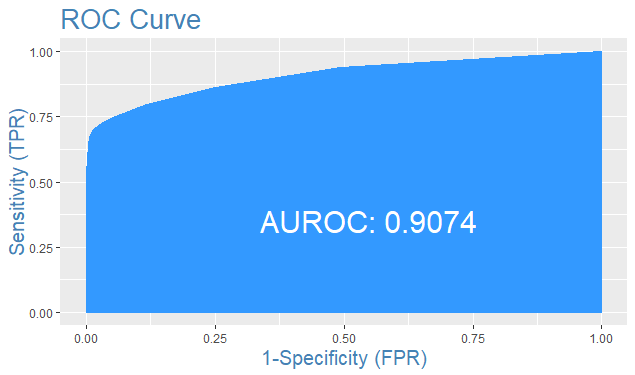
**ACCURACY FOR TRAIN DATA: 96.9 %**

**MIS-CLASSIFICATION ERROR FOR TRAIN DATA: 3.08 %**

**SENSITIVITY FOR TRAIN DATA: 66.5 %**

**SPECIFICITY FOR TRAIN DATA: 99.4 %**

**AUC-ROC FOR TRAIN DATA: 90.74 %**



**FINDINGS FROM TRAIN DATA PREDICTION:**

***“ACCURACY (96.9 %), AUC-ROC (90.74 %) INDICATE A VERY GOOD PREDICITABILITY POWER OF THE MODEL. SPECIFICITY OF 99.4 % TELLS THAT IT IS ABLE TO PREDICT LOW RISK CATEGORY VERY WELL, HOWEVER SENSITIVITY OF 66.5 % INDICATES THAT PREDICTION ABILITY ON HIGH RISK CATEGORY CAN BE IMPROVED FURTHER.”***

**PREDICTION ON TEST DATA SET:**

The model built is used on the test data to predict the probabilities and class level of loan\_status and then confusion matrix, accuracy, AUC-ROC, Mis-Classification error, sensitivity and specificity are derived for the test data.

**CONFUSION MATRIX FOR TEST DATA:**

horizontal class labels represent predicted classes and vertical class labels represent actual classes

**0 1**

**0 244398 6806**

**1 1534 13475**

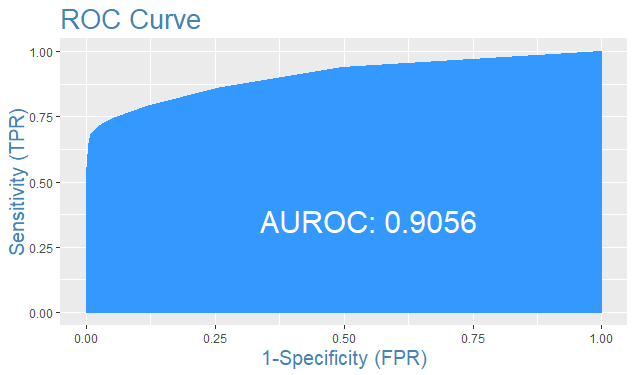
**ACCURACY FOR TEST DATA: 96.8 %**

**MIS-CLASSIFICATION ERROR FOR TEST DATA: 3.13 %**

**SENSITIVITY FOR TEST DATA: 66.44 %**

**SPECIFICITY FOR TEST DATA: 99.37 %**

**AUC-ROC FOR TEST DATA: 90.56%**



**FINDINGS FROM TEST DATA PREDICTION:**

***“ACCURACY (96.8 %), AUC-ROC (90.56 %) INDICATE A VERY GOOD PREDICITABILITY POWER OF THE MODEL. SPECIFICITY OF 99.37 % TELLS THAT IT IS ABLE TO PREDICT LOW RISK CATEGORY VERY WELL, HOWEVER SENSITIVITY OF 66.44 % INDICATES THAT PREDICTION ABILITY ON HIGH RISK CATEGORY CAN BE IMPROVED FURTHER.”***

**SUMMARY ON PERFORMANCE OF THE MODEL:**

|  |  |  |
| --- | --- | --- |
| **Log likelihood test** | **p value< 2.2e-16** | **MODEL IS OVERALL SIGNIFICANT ENOUGH** |
| **McFadden R-square** | **0.561** | **EXCELLENT GOODNESS OF FIT** |

|  |  |  |
| --- | --- | --- |
| **Measure** | **Train data** | **Test data** |
| **Accuracy** | **96.9 %** | **96.8 %** |
| **AUC-ROC** | **90.74 %** | **90.56 %** |
| **misClassError** | **3.08 %** | **3.13 %** |
| **Sensitivity** | **66.5 %** | **66.44 %** |
| **Specificity** | **99.4 %** | **99.37 %** |

**LOGISTIC REGRESSION MODEL FOR NEW CUSTOMERS**

**PURPOSE:** in the process of building the model, the dataset is partitioned into train data and test data and the model is built on train data. The purpose of this model is to predict the level of default risk for new customers based on historical records of existing customer data.

Before we proceed for modelling, we can drop two more variables “id”, “member\_id”. After removing these variables, we are left with **11 variables which are relevant for new customer data.**

**LIST OF VARIABLES USED FOR THE MODEL BUILDING:**

[1] "loan\_amnt" "term" "grade" "emp\_length" "home\_ownership"

[6] "annual\_inc" "loan\_status" "purpose" "addr\_state" "inq\_last\_6mths"

[11] "pub\_rec"

**DEPENDENT and INDEPENDENT VARIABLES: loan\_status is dependent variable and remaining all others are independent variables**

loan\_status = 0 (low risk)

loan\_status = 1 (high risk)

**SPLIT THE DATA INTO TRAIN AND TEST DATA SETS:**

Now we will divide the data into train & test in **70:30** ratio. After dividing data, we are left with below number of records in Train & test sample.

Train -> 621166

Test -> 266213

**BUILD LOGISTIC REGRESSION MODEL:**

Now as part of next step we will build the binomial logistic regression model. In order to perform the same, we will use **“glm” function**

logit = glm(loan\_status ~ ., data=Training, family= binomial)

**OVERALL SIGNIFICANCE/VALIDITY OF THE MODEL (log likelihood test):**

Now let us do likelihood ratio test using function “lrtest”, upon doing this test we got **p value < 2.2e-16**. This means p value is highly significant. It implies that the null hypothesis of "all the coefficient estimates are Zero" is rejected and we conclude that at least one coefficient estimate is non-Zero.

**Inference: overall model significance is high**

**MEASURE OF FIT: CALCULATE McFadden R-Square:**

|  |  |  |
| --- | --- | --- |
| Measure | Value | Description |
| llh | -1.578858e+05 | Log-likelihood from the fitted model |
| llhNull | -1.669036e+05 | Log-likelihood from intercept-only restricted model |
| G2 | 1.803558e+04 | Minus 2 times the difference in the log-likelihoods |
| McFadden | 5.402994e-02 | McFadden’s pseudo r-squared |
| r2ML | 2.861758e-02 | Maximum Likelihood pseudo r-squared |
| r2CU | 6.883732e-02 | Cragg and Uhler’s pseudo r-squared |

**Interpretation of McFadden’s r square:**

McFadden’s R squared for the model is 0.054, and this McFadden’s R square value indicates that fit of the model is not robust enough (to improve the fit of the model, Support Vector Machine and Random Forest algorithms are used on the same data, which are presented in next coming sections of this report)

**SUMMARY OF LOGIT FUNCTION:**

* Summary of logit function gives the coefficient estimate values, standard errors and p-values from which significance of the variables can be determined.
* **Variables for which p-value is less than 0.05 are considered statistically significant.**

**LIST OF STATISTICALLY SIGNIFICANT VARIABLES FROM SUMMARY**

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.606e+00 1.120e-01 -32.197 < 2e-16 \*\*\*

loan\_amnt 8.036e-06 7.565e-07 10.622 < 2e-16 \*\*\*

term -1.466e-01 1.262e-02 -11.616 < 2e-16 \*\*\*

gradeB 7.465e-01 2.278e-02 32.765 < 2e-16 \*\*\*

gradeC 1.104e+00 2.256e-02 48.950 < 2e-16 \*\*\*

gradeD 1.477e+00 2.353e-02 62.746 < 2e-16 \*\*\*

gradeE 1.709e+00 2.586e-02 66.080 < 2e-16 \*\*\*

gradeF 2.052e+00 3.087e-02 66.456 < 2e-16 \*\*\*

gradeG 2.227e+00 4.634e-02 48.052 < 2e-16 \*\*\*

emp\_length -9.197e-03 1.403e-03 -6.554 5.61e-11 \*\*\*

home\_ownershipOtherNone 7.810e-01 2.124e-01 3.677 0.000236 \*\*\*

home\_ownershipRENT 1.389e-01 1.141e-02 12.177 < 2e-16 \*\*\*

annual\_inc -4.071e-06 1.558e-07 -26.129 < 2e-16 \*\*\*

purposeeducational 9.040e-01 1.583e-01 5.709 1.13e-08 \*\*\*

purposehouse 1.719e-01 8.103e-02 2.121 0.033928 \*

purposemedical 1.441e-01 6.854e-02 2.102 0.035545 \*

purposeother 1.194e-01 5.548e-02 2.153 0.031344 \*

purposerenewable\_energy 3.359e-01 1.627e-01 2.065 0.038943 \*

purposesmall\_business 6.761e-01 6.107e-02 11.071 < 2e-16 \*\*\*

purposewedding 3.301e-01 9.413e-02 3.507 0.000453 \*\*\*

addr\_stateAL 2.259e-01 1.046e-01 2.160 0.030757 \*

addr\_stateDC -5.145e-01 1.559e-01 -3.299 0.000970 \*\*\*

addr\_stateHI 2.401e-01 1.139e-01 2.108 0.035005 \*

addr\_stateME -3.316e+00 9.957e-01 -3.330 0.000868 \*\*\*

addr\_stateMS -3.539e-01 1.299e-01 -2.724 0.006452 \*\*

addr\_stateND -1.730e+00 5.134e-01 -3.369 0.000755 \*\*\*

addr\_stateNE -1.335e+00 2.643e-01 -5.051 4.39e-07 \*\*\*

addr\_stateNV 2.978e-01 1.032e-01 2.884 0.003924 \*\*

addr\_stateNY 2.004e-01 9.772e-02 2.051 0.040299 \*

inq\_last\_6mths 1.726e-01 4.330e-03 39.865 < 2e-16 \*\*\*

pub\_rec -1.989e-01 1.045e-02 -19.028 < 2e-16 \*\*\*

**INTERPRETATIONS FROM SUMMARY:**

**a) Coefficient interpretation:**

* Interpretation of the coefficient of the “loan\_amnt”:

The coefficient of loan\_amnt is 8.036e-06, which indicates that for one unit increase in loan\_amnt (holding all other variables constant), the log odds of defaulting over not defaulting is 8.036e-06

* Similar interpretations can be made with the coefficients of all other variables.
* Standard error is less for all the variables

**b) VARIABLE IMPORTANCE BASED ON Z-VALUES:**

TABLE 4.1: VARIABLE IMPORTANCE BASED ON Z-VALUES

Adjacent table contains the most important variables (top 8) ordered by z-value

|  |  |
| --- | --- |
| **VARIABLE** | **Z-VALUE (absolute)** |
| **grade** | **as high as 66.4** |
| **inq\_last\_6mths** | **39.865** |
| **annual\_inc** | **26.129** |
| **pub\_rec** | **19.028** |
| **home\_ownershipRENT** | **12.177** |
| **term** | **11.616** |
| **purposesmall\_business** | **11.071** |
| **loan\_amnt** | **10.622** |

**CONCLUSION/IMPORTANT VARIABLES:** **from the above table, it can be concluded that “grade” (which is CIBIL score in banking terms) has highest z-value indicating that “grade” is the MOST important variable to predict the level of risk involved in giving a loan to a new customer.**

**ODDS AND PROBABILITES:**

**Following is the list of variables for which odds ratio >1: - practically significant variables**

**VARIABLE OODS-RATIO**

loan\_amnt 1.00

gradeB 2.10

gradeC 3.01

gradeD 4.37

gradeE 5.52

gradeF 7.78

gradeG 9.27

home\_ownershipOtherNone 2.18

home\_ownershipRENT 1.14

purposedebt\_consolidation 1.05

purposeeducational 2.46

purposehome\_improvement 1.04

purposehouse 1.18

purposemajor\_purchase 1.03

purposemedical 1.15

purposemoving 1.12

purposeother 1.12

purposerenewable\_energy 1.39

purposesmall\_business 1.96

purposevacation 1.02

purposewedding 1.39

addr\_stateAL 1.25

addr\_stateAR 1.11

addr\_stateAZ 1.14

addr\_stateCA 1.16

addr\_stateCT 1.00

addr\_stateDE 1.04

addr\_stateFL 1.19

addr\_stateGA 1.02

addr\_stateHI 1.27

addr\_stateIA 2.27

addr\_stateID 1.55

addr\_stateIN 1.00

addr\_stateMA 1.12

addr\_stateMD 1.18

addr\_stateMI 1.09

addr\_stateMN 1.07

addr\_stateMO 1.09

addr\_stateNC 1.14

addr\_stateNJ 1.19

addr\_stateNM 1.23

addr\_stateNV 1.34

addr\_stateNY 1.22

addr\_stateOH 1.02

addr\_stateOK 1.17

addr\_stateOR 1.05

addr\_statePA 1.02

addr\_stateRI 1.13

addr\_stateUT 1.20

addr\_stateVA 1.20

addr\_stateWA 1.02

inq\_last\_6mths 1.18

**Interpreting odds ratio and probability:**

* **Odds ratio for loan\_amnt is 1.00000804 and probability is 0.50**

Keeping all other variables constant, if loan amount increases by one unit, the odds of defaulting become 1.000008 times the odds of not defaulting and the probability of defaulting is 50%.

* **Odds ratio for gradeB is 2.10 and probability is 0.67**

Keeping all other variables constant, if term increases by one unit, the odds of defaulting become 2.10 times the odds of not defaulting and the probability of defaulting is 67%.

* **Odds ratios and probabilities of all other variables can be interpreted in a similar way as above.**

**PREDICTION ON TRAIN DATA SET:**

The model built is used on the train data to predict the probabilities and class level of loan\_status and then confusion matrix, accuracy, AUC-ROC, Mis-Classification error, sensitivity and specificity are derived for the train data.

**CONFUSION MATRIX FOR TRAIN DATA:**

horizontal class labels represent predicted classes and vertical class labels represent actual classes

**0 1**

**0 573946 47115**

**1 60 45**

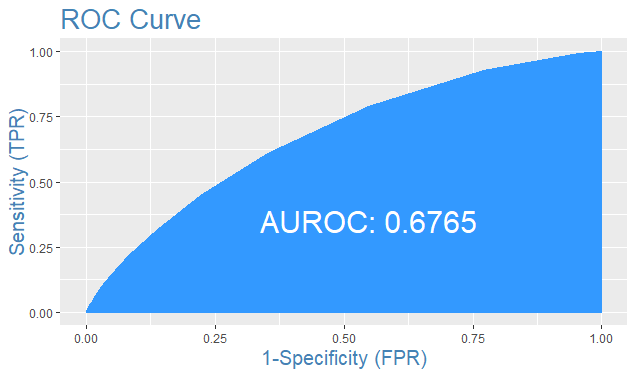
**ACCURACY FOR TRAIN DATA: 92.4 %**

**MIS-CLASSIFICATION ERROR FOR TRAIN DATA: 7.59 %**

**SENSITIVITY FOR TRAIN DATA: 0.1 %**

**SPECIFICITY FOR TRAIN DATA: 99.98 %**

**AUC-ROC FOR TRAIN DATA:**



**FINDINGS FROM TRAIN DATA PREDICTION:**

***“ACCURACY (92.4 %) INDICATE A GOOD PREDICITABILITY POWER OF THE MODEL. SPECIFICITY OF 99.9% TELLS THAT IT IS ABLE TO PREDICT LOW RISK CATEGORY VERY WELL. HOWEVER, SENSITIVITY OF 0.1 % INDICATES THAT PREDICTION ABILITY ON HIGH RISK CATEGORY CAN BE IMPROVED FURTHER.AUC-ROC OF 67.65 % CAN BE ALSO BE IMPROVED”***

**PREDICTION ON TEST DATA SET:**

The model built is used on the test data to predict the probabilities and class level of loan\_status and then confusion matrix, accuracy, AUC-ROC, Mis-Classification error, sensitivity and specificity are derived for the test data.

**CONFUSION MATRIX FOR TEST DATA:**

horizontal class labels represent predicted classes and vertical class labels represent actual classes

**0 1**

**0 245920 20247**

**1 24 22**

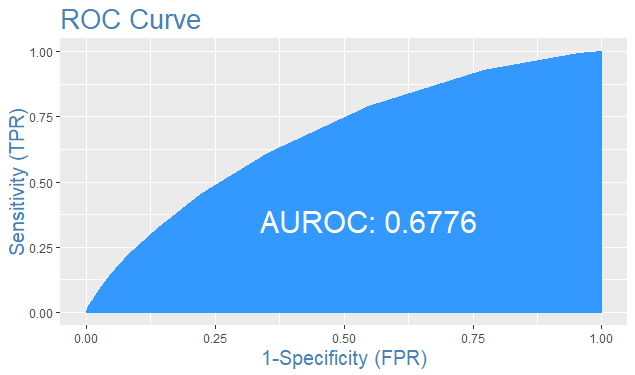
**ACCURACY FOR TEST DATA: 92.3 %**

**MIS-CLASSIFICATION ERROR FOR TEST DATA: 7.6 %**

**SENSITIVITY FOR TEST DATA: 0.1 %**

**SPECIFICITY FOR TEST DATA: 99.9 %**

**AUC-ROC FOR TEST DATA:**



**FINDINGS FROM TEST DATA PREDICTION:**

***“ACCURACY (92.3 %) INDICATE A GOOD PREDICITABILITY POWER OF THE MODEL. SPECIFICITY OF 99.9% TELLS THAT IT IS ABLE TO PREDICT LOW RISK CATEGORY VERY WELL. HOWEVER, SENSITIVITY OF 0.1 % INDICATES THAT PREDICTION ABILITY ON HIGH RISK CATEGORY CAN BE IMPROVED FURTHER.AUC-ROC OF 67.76 % CAN BE ALSO BE IMPROVED”***

**SUMMARY ON PERFORMANCE OF LOGISTIC REGRESSION MODEL:**

TABLE 4.2: PERFORMANCE OF LR MODEL FOR NEW CUSTOMER

|  |  |  |
| --- | --- | --- |
| **Log likelihood test** | **p value< 2.2e-16** | **MODEL IS OVERALL SIGNIFICANT ENOUGH** |
| **McFadden R-square** | **5.402994e-02** | **GOODNESS OF FIT IS NOT ROBUST** |

|  |  |  |
| --- | --- | --- |
| **Measure** | **Train data** | **Test data** |
| **Accuracy** | **92.4 %** | **92.3 %** |
| **AUC-ROC** | **67.65 %** | **67.76 %** |
| **misClassError** | **7.59 %** | **7.6 %** |
| **Sensitivity** | **0.1%** | **0.1%** |
| **Specificity** | **99.9%** | **99.9%** |

**NOTE: since the results of logistic regression with SMOTE are also not satisfactory to the level expected, we are employing SVM (support vector machine) and Random Forest algorithms to improve the predictability power of the model, in regard to minority class (that is 1’s).**

**SVM (Support Vector Machine) PREDICTION MODEL FOR NEW CUSTOMERS**

**PURPOSE:** in the process of building the model, the dataset is partitioned into train data and test data and the model is built on train data. The purpose of this model is to predict the level of default risk for new customers based on historical records of existing customer data.

Before we proceed for modelling, we can drop two more variables “id”, “member\_id”. After removing these variables, we are left with **11 variables which are relevant for new customer data.**

**LIST OF VARIABLES USED FOR THE MODEL BUILDING:**

[1] "loan\_amnt" "term" "grade" "emp\_length" "home\_ownership"

[6] "annual\_inc" "loan\_status" "purpose" "addr\_state" "inq\_last\_6mths"

[11] "pub\_rec"

**DEPENDENT and INDEPENDENT VARIABLES: loan\_status is dependent variable and remaining all others are independent variables**

loan\_status = 0 (low risk)

loan\_status = 1 (high risk)

**SPLIT THE DATA INTO TRAIN AND TEST DATA SETS:**

A stratified sample of 221845 records is considered for this model. Now we will divide the data into train & test in the ratio of **70:30**. After dividing data, we are left with below number of records in Train & test sample.

Train -> 155292

Test -> 66553

**BUILD SUPPORT VECTOR MACHINE MODEL:**

Now as part of next step we built the support vector machine model. In order to perform the same, from the library **“e1071”** we use the **“svm”** function with gamma=4 and cost=5 to get better prediction

model\_svm <- svm(loan\_status~. , data=Training, method="C-classification", gamma=4, cost=5)

**CHECK THE SUMMARY OF THE MODEL:**

Parameters:

SVM-Type: C-classification

SVM-Kernel: radial

cost: 5

gamma: 4

Number of Support Vectors: 120174

(108412 11762)

Number of Classes: 2

Levels: 0 1

**PREDICTION ON TRAIN DATA SET:**

The model built is used on the train data to predict the loan\_status and then confusion matrix, accuracy, kappa, sensitivity and specificity are derived for the train data.

**CONFUSION MATRIX FOR TRAIN DATA:**

**Reference**

**Prediction 0 1**

**0 143428 64**

**1 2147 9653**

**McNemar's Test P-Value: < 2.2e-16**

**ACCURACY FOR TRAIN DATA: 98.58 %**

**SENSITIVITY FOR TRAIN DATA: 98.53%**

**SPECIFICITY FOR TRAIN DATA: 99.34 %**

**FINDINGS FROM TEST DATA PREDICTION:**

***“ACCURACY (98.58 %) INDICATE A VERY GOOD PREDICITABILITY POWER OF THE MODEL.VALUES OF SPECIFICITY AND SENSITIVITY INDICATE THAT BOTH 0’S AND 1’S ARE PREDICTED WITH HIGH ACCURACY”***

**PREDICTION ON TEST DATA SET:**

The model built is used on the test data to predict the loan\_status and then confusion matrix, accuracy, kappa, sensitivity and specificity are derived for the test data.

**CONFUSION MATRIX FOR TEST DATA:**

**Reference**

**Prediction 0 1**

**0 143428 64**

**1 2147 9653**

**ACCURACY FOR TEST DATA: 90.55 %**

**SENSITIVITY FOR TEST DATA: 92.47%**

**SPECIFICITY FOR TEST DATA: 100%**

**McNemar's Test P-Value: < 2.2e-16**

**FINDINGS FROM TEST DATA PREDICTION:**

***“ACCURACY (90.55 %) INDICATE A VERY GOOD PREDICITABILITY POWER OF THE MODEL.VALUES OF SPECIFICITY AND SENSITIVITY INDICATE THAT BOTH 0’S AND 1’S ARE PREDICTED WITH HIGH ACCURACY”***

**SUMMARY ON PERFORMANCE OF SUPPORT VECTOR MACHINE MODEL:**

TABLE 4.3: PERFORMANCE OF SVM FOR NEW CUSTOMER

|  |  |  |
| --- | --- | --- |
| **Measure** | **Train data** | **Test data** |
| **Accuracy** | **98.58 %** | **90.55 %** |
| **Sensitivity** | **98.53 %** | **92.47 %** |
| **Specificity** | **99.34 %** | **100 %** |

**RANDOM FOREST PREDICTION MODEL FOR NEW CUSTOMERS**

**PURPOSE:** in the process of building the model, the dataset is partitioned into train data and test data and the model is built on train data. The purpose of this model is to predict the level of default risk for new customers based on historical records of existing customer data.

Before we proceed for modelling, we can drop two more variables “id”, “member\_id”. After removing these variables, we are left with **11 variables which are relevant for new customer data.**

**LIST OF VARIABLES USED FOR THE MODEL BUILDING:**

[1] "loan\_amnt" "term" "grade" "emp\_length" "home\_ownership"

[6] "annual\_inc" "loan\_status" "purpose" "addr\_state" "inq\_last\_6mths"

[11] "pub\_rec"

**DEPENDENT and INDEPENDENT VARIABLES: loan\_status is dependent variable and remaining all others are independent variables**

loan\_status = 1 (low risk)

loan\_status = 2 (high risk)

**SPLIT THE DATA INTO TRAIN AND TEST DATA SETS:**

Now we will divide the data into train & test in **70:30** ratio. After dividing data, we are left with below number of records in Train & test sample.

**BUILD RANDOM FOREST MODEL:**

Now as part of next step we will build the random forest model by using the **“randomForest”** function and we have taken the number of tress as 10 for building the model on training dataset.

rf <-randomForest(loan\_status~. ,data=Training, ntree=10)

**FIND OPTIMAL MTRY VALUE BY TUNING:**

We have to identified the number of variables selected at each split (denoted by mtry in randomforest function) as mentioned below.

mtry <- tuneRF(Training[-7],Training$loan\_status, ntreeTry=10, stepFactor=1.5, improve=0.01, trace=TRUE, plot=TRUE)

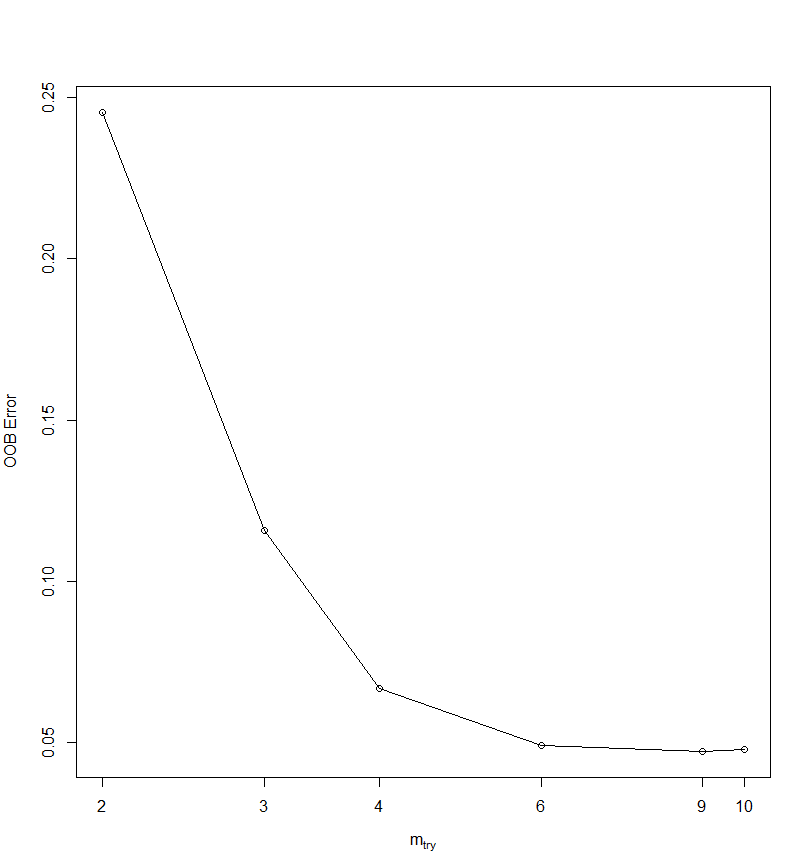
ntreeTry: number of trees used at the tuning step.

stepFactor: at each iteration, mtry is inflated (or deflated) by this value.

improve: the (relative) improvement in OOB error must be by this much for the search to continue.

trace: whether to print the progress of the search.

plot**:** whether to plot the OOB error as function of mtry.



**“Based on the above results and minimum out of bag error rate we got the best mtry as 9 hence we re-ran the model with best mtry value i.e. 9” and following are the results.**

**PREDICTION ON TRAIN DATA SET:**

The model is built used on the train data to predict the class level of loan\_status and then confusion matrix, accuracy and AUC-ROC are derived for the train data.

**CONFUSION MATRIX FOR TRAIN DATA:**

horizontal class labels represent predicted classes and vertical class labels represent actual classes.

**1 2**

**1 515236 52730**

**2 1061 566321**

**0.952, that is 95.2%**

**ACCURACY FOR TRAIN DATA:**

**Inference: overall accuracy of 95.2% on train data indicates a very good predictive ability of the model.**

**VARIABLE IMPORTANCE:**

From the below plots it is identified that **grade, addr\_state, annual\_inc and loan\_amnt are the more important variables** as mean decrease in accuracy and mean decrease in gini is highest for them.

FIGURE 4.4: VARIABLE IMPORTANCE FROM RF

****

**ROC CURVE FOR TRAIN DATA SET: AUC-ROC : 98.65%**



**FINDINGS FROM TRAIN DATA PREDICTION:**

***“ACCURACY (95.2 %) AND AUC-ROC (98.65%) INDICATE A VERY GOOD PREDICITABILITY POWER OF THE MODEL”***

**PREDICTION ON TEST DATA SET:**

The model built is used on the test data to predict the class level of loan\_status and then confusion matrix, accuracy and AUC-ROC are derived for the test data.

**CONFUSION MATRIX FOR TEST DATA:**

horizontal class labels represent predicted classes and vertical class labels represent actual classes

**1 2**

**1 193459 49995**

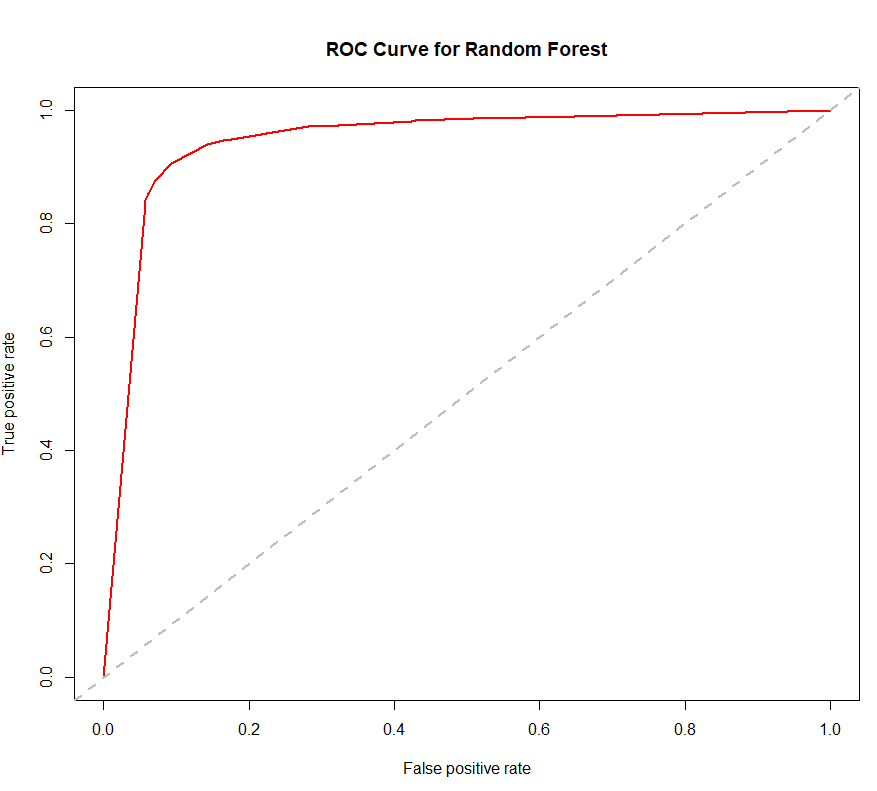
**2 10951 232342**

**ACCURACY FOR TEST DATA:**

**0.87478, that is 87.4%**

**Inference: overall accuracy of 87.4% on test data indicates a very good predictive ability of the model.**

**ROC CURVE FOR TEST DATA : AUC-ROC : 94.4%**



**FINDINGS FROM TEST DATA PREDICTION:**

***“ACCURACY (87.4 %) AND AUC-ROC (94.4%) INDICATE A VERY GOOD PREDICITABILITY POWER OF THE MODEL”***

**SUMMARY ON PERFORMANCE OF RANDOM FORESET MODEL:**

TABLE 4.5: PERFORMANCE OF RF FOR NEW CUSTOMER

|  |  |  |
| --- | --- | --- |
| **Measure** | **Train data** | **Test data** |
| **Accuracy** | **95.2%** | **87.4%** |
| **AUC** | **98.65%** | **94.4%** |
| **KS** | **96.2%** | **81.3%** |

**CHAPTER - 5**

**CLUSTERING FOR CUSTOMER PROFILING**

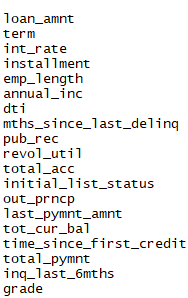
**PURPOSE OF CLUSTERING:**

For meaningful exploratory data analysis (EDA) where observations of the credit history are divided into meaningful groups that share common characteristics (features) leading to REATIVE PROBABILUITY OF DEFAULT EXTRAPOLATION

**VARIABLES UNDER STUDY:**

Initially we consider the variables which were significant in the logistic model out-come to cluster our data.

We used the following significant variables for clustering



We have now 8,87,379 observations and 19 variables for clustering. There is only 1 categorical variable i.e. GRADE – we put some ordinal value to the grades and convert it to the numeric.

**SCALING:**

The numerical data must be scaled to fetch proper clusters

**K-MEANS CLUSTERING:**

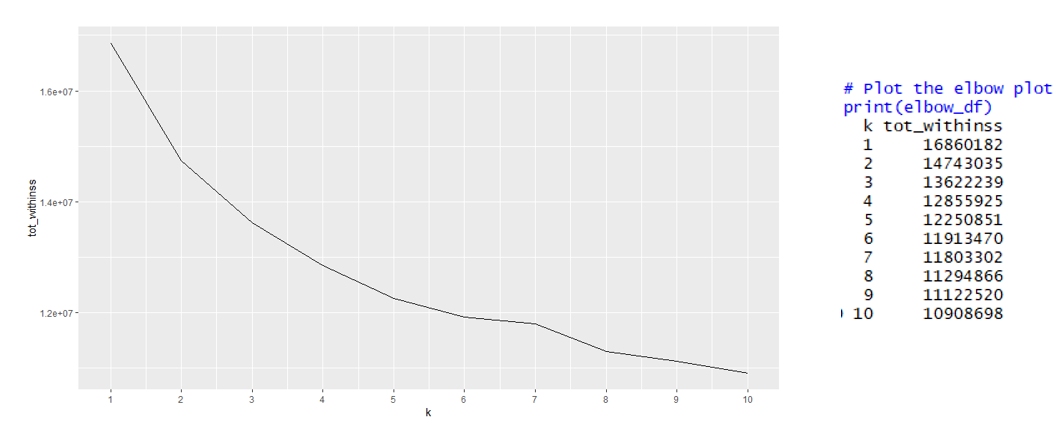
**FINDING THE OPTIMIZED NUMBER OF CLUSTERS**

Once the data is prepared, we apply K-MEANS to our data as our data is huge containing 8 lakhs observations and K-MEANS has relatively lower Computation Complexity than hierarchical clustering, before that we find out the optimal number of clusters through elbow method to set the criterion for K-MEANS.

As K increases the total within sum of squares decreases, because we want to call for multiple values of K, we need to create multiple model and extract the corresponding values

To do this we use map\_double (map\_dbl) function from purrr library

Below is the iterated value to total sum-of-square within for all values of K, from this we can create the elbow plot. ELBOW: - point at which the curve flattens out gives us the optimal no of clusters.



It seems for a huge data with around 8 lakhs records we take K= 10 as the optimal number of clusters where tot\_withinss is the lowest, prior to k=10 we have a steady decreasing slope of the curve.

**CLUSTERING WITH K-MEANS TAKING 10 AS OPTIMAL NUMBER OF THE CLUSTERS**

We build the KMEANS model and then extract the cluster assignment vector from the k-means model. Then create a new data frame appending the cluster assignment

|  |  |  |
| --- | --- | --- |
| **CLUSTER** | **Number of observations** | **% of observations** |
| 1 | 1,12,967.00 | 13% |
| 2 | 82,906.00 | 9% |
| 3 | 77,190.00 | 9% |
| 4 | 36,411.00 | 4% |
| 5 | 1,00,318.00 | 11% |
| 6 | 64,398.00 | 7% |
| 7 | 72,818.00 | 8% |
| 8 | 1,52,285.00 | 17% |
| 9 | 1,42,782.00 | 16% |
| 10 | 45,304.00 | 5% |
| **TOTAL OBS** | **8,87,379.00** | 100% |

**PROFILING OF THE CLUSTERS:**

We calculate the mean for the clustering variable for each cluster

TABLE 5.1: CUSTOMER PROFILING



**SUMMARIZING THE CLUSTERS ON THEIR RISK FACTORS - RESULTS AND INTERPRETATION:**

TABLE 5.2: ACTIONABLE INSIGHTS FROM CLUSTERING

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LEVEL OF RISK** | **CLUSTER** | **dti** | **grade** | **ACTIONABLE INSIGHTS** |
| **HIGH-RISK** | Cluster 3 | 21.2 | 3.13 | DETERMINE THE PROBABILITY OF DEFAULT OF THESE GROUP. PERSONALISED FAST RECOVERY ACTIONS ON NPAs: high risk borrowers cluster can be further sub-clustered based on the delinquency period as follows: Sub-cluster  NPA  Repossession  Valuation  Sale |
| **MEDIUM RISK** | cluster 2 | 16 | 5.77 | Give away offers that would increase the clv of the customers like balance transfer, loan refinance, top up loans in calculative amount. |
|  | cluster 5 | 16.4 | 6.28 |
|  | cluster 7 | 19.9 | 5.11 |
| **LOW RISK** | Cluster 9 | 13.4 | 6.47 | CUSTOMER RETENTION/ACQUISITION STRATEGY:  low risk borrowers cluster can be rewarded with personalised retail, cash back, cross selling and up selling offers. Such kind of actions are a win-win situation for both the customers and the bank. These actions also help in retaining the existing customers in long run and also helps in attracting new customers. |

**FURTHER CLUSTERING OF HIGH-RISK CUSTOMERS:**

**FINDING THE OPTIMIZED NO OF SUB CLUSTERS FOR HIGH-RISK CUSTOMERS**

From the Elbow curve below that the high-risk customer can be further divided into **3** more clusters depending on the severity of their risk.



**#CLUSTERING WITH K-MEANS TAKING 3 AS OPTIMAL NO OF THE SUB CLUSTERS FOR HIGH-RISK CUSTOMERS**

|  |  |  |
| --- | --- | --- |
| **HIGH RISK SUB CLUSTERS** | **NUMBER OF OBSERVATIONS** | **% OF OBSERVATIONS** |
| 1 | 5816 | 8% |
| 2 | 48579 | 63% |
| 3 | 22795 | 30% |
| Total | 77190 | 100% |



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LEVEL OF RISK** | **CLUSTER NUMBER** | **dti** | **grade** | **ACTIONABLE INSIGHTS** |
| **VERY HIGH RISK** | CLUSTER2 | 19.4 | 3.18 | Probability of Default is very High, there are chances of NPA |
| **HIGH-RISK** | CLUSTER 1 | 21 | 3.22 | Chances of defaulting in near future is high. Might require REPETATIVE REMINDERS from collection team |
| CLUSTER 3 | 21 | 3.22 |

**BUSINESS IMPLICATIONS FROM CLUSTERING:**

* **OFFERS FOR LOW RISK CUSTOMERS /SAFE CUSTOMERS:**

CUSTOMER RETENTION/ACQUISITION STRATEGY: low risk borrowers cluster can be rewarded with personalised retail, cash back, cross selling and up selling offers. Such kind of actions are a win-win situation for both the customers and the bank. These actions also help in retaining the existing customers in long run and also helps in attracting new customers.

* **ACTIONS ON HIGH RISK CUSTOMERS**: **Customers with very high debt to income ratio and low grades**

PERSONALISED FAST RECOVERY ACTIONS OR REMINERS OF PAYMENT AND FOLLOWUPS ON PROBABLE HIGH RISK CUSTOMERS:

High risk borrowers cluster can be further sub-clustered into:

1. **VERY HIGH RISKY CUSTOMERS** -- determine the probability of default of the customers and arrange for follow ups and reminders as they have a good chance of turning into NPA’S
2. **HIGH RISKY CUSTOMERS** --- can be given offers of certain % interest exemption if they repay the loans properly

**CHAPTER - 6**

**TIME SERIES FORECAST OF CREDIT DEMAND**

**PURPOSE:** The main purpose of forecasting credit demand is to predict loan amount for “Lending Club” for future.

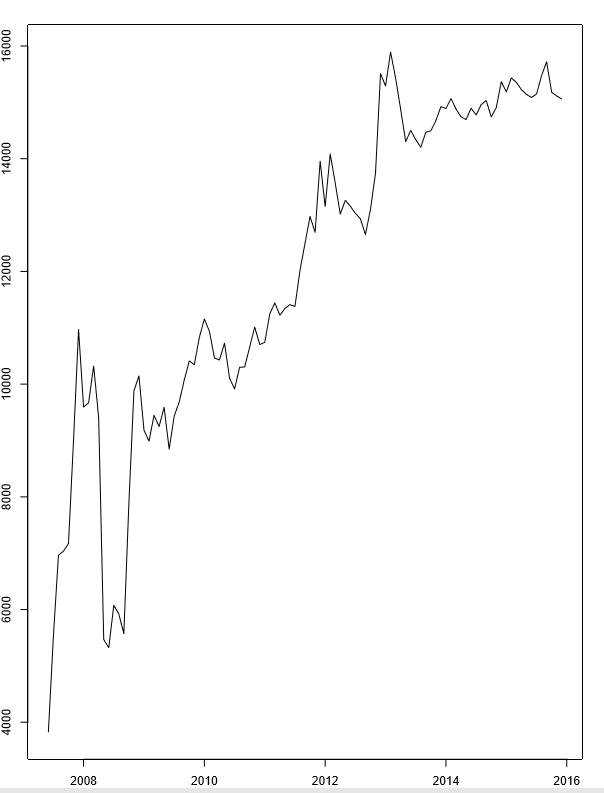
We have issue date in database & Loan amount, we will use this historical information to forecast future loan amount. Before we proceed for modelling, we will segregate “issue date” into Year & Month. In this process we will use subset of data having below variables.

* Year
* Month
* Loan Amount issues in that month
* Average Loan amount issued in that month
* Loan Application Count

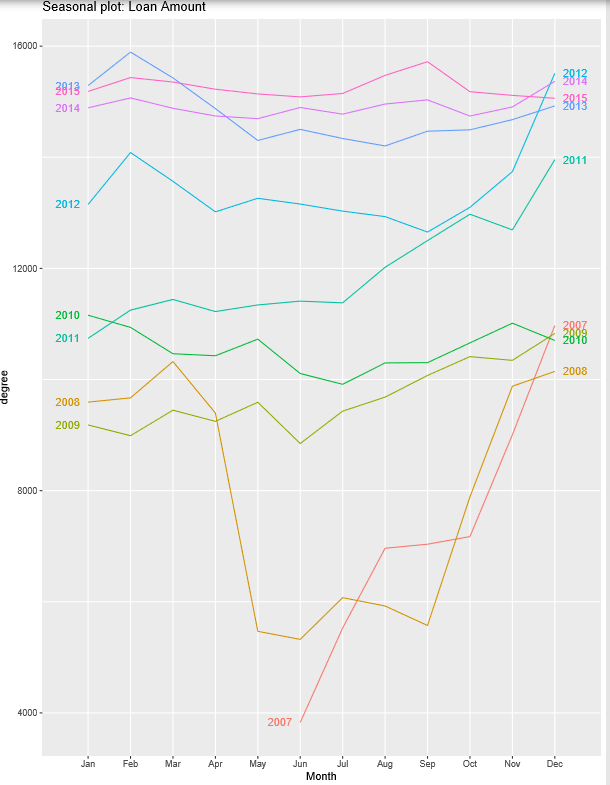
We could see that data is not sorted therefore we sorted data based on year & month. We then applied Time series technique on Total Loan amount, but MPEA value obtained is more than “1” which means Time series modelling for Total loan amount is not feasible. This is because total loan amount increased exponentially over the years. We therefore performed Time series on Average loan amount & Loan Application count.

Below are the requisite steps for Time series modelling.

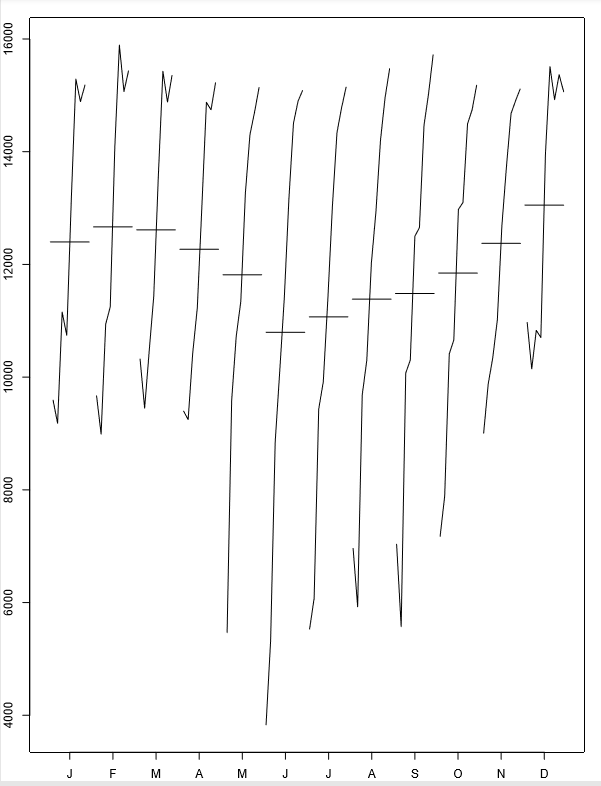
**Time Series Plot for Average Loan Amount.**



**ggseasonal plot of loan amount:**

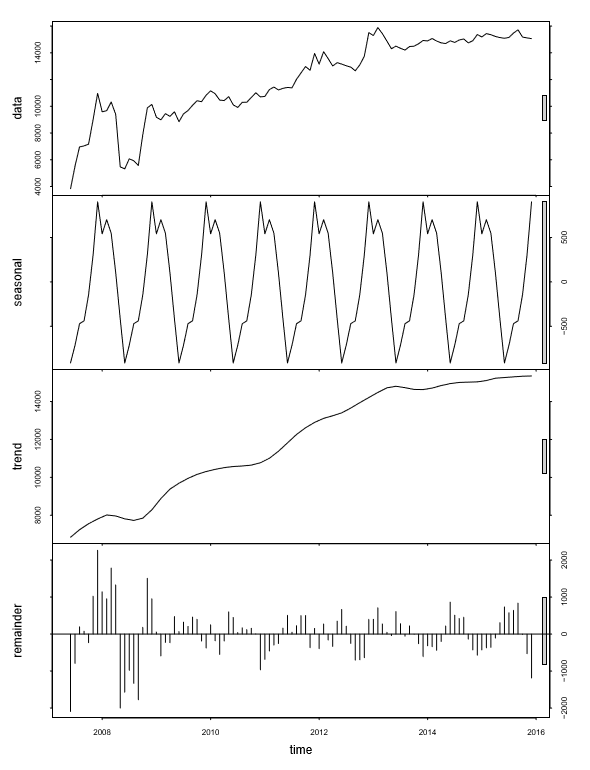


**Month Plot:**



From these 3 plots we can infer that there is an increasing trend present in data though there is no notable seasonality observed in initial years (2007-2010) but in later years (2011-2015) we could observe some degree of seasonality.

**Decompose data into Trend, seasonality & residuals:**



**Check for stationarity: [Augmented Dickey Fuller test]**

Below is the out of ADF test. From the values it can be conferred that series is stationary.

Augmented Dickey-Fuller Test

data: forecst.ItemA$residuals

Dickey-Fuller = -4.8044, Lag order = 4, p-value = 0.01

alternative hypothesis: stationary

**SPLIT THE DATA INTO TRAIN AND TEST DATA SETS:**

Now we will divide the data into train & test. We have data from June 2007 till December 2015, we put data from June 2007 till 2014 June in train & data from July 2014 to December 2015 into test.

We now apply below 3 models & find corresponding MPEA value

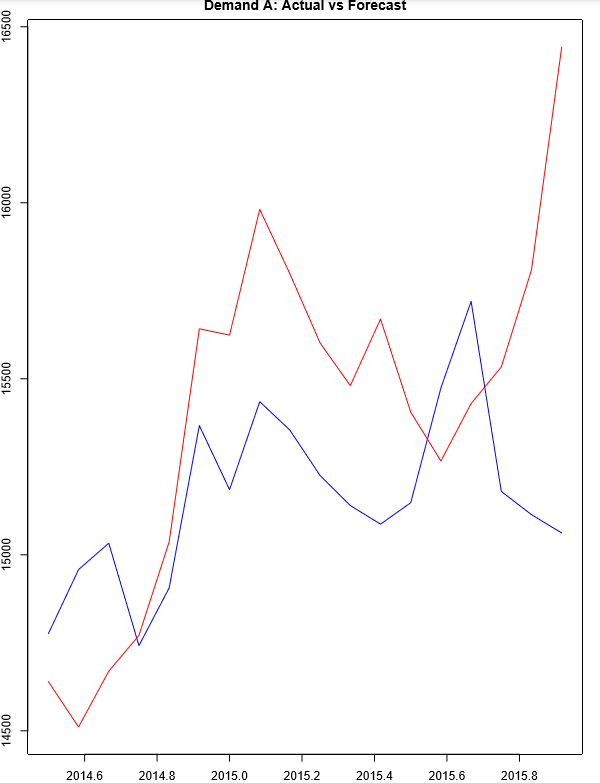
* Random walk with drift
* Holt winters
* ARIMA

Upon applying these 3 models we obtained below MAPE values.

|  |  |
| --- | --- |
| **Model** | **MAPE** |
| **Random walk with drift** | **0.1269** |
| **Holt winters** | **0.0259** |
| **ARIMA** | **0.0550** |

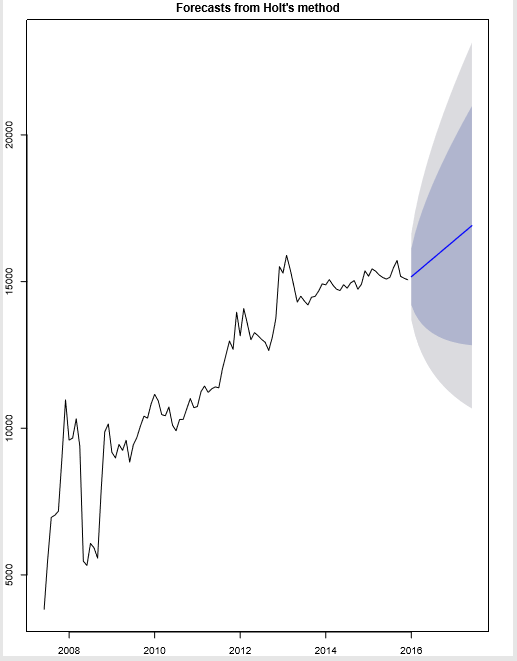
We could see that MAPE value is least for Holt winter’s model, therefore we will forecast using Holt winter’s model.

**Actual vs Forecast using holt winters model**



We also applied Holts double exponential method to forecast values, below is the plot using holt’s method.

FIGURE 6.1: AVERAGE LOAN AMOUNT FORECAST



**Forecasting value for next 18 months that is for from Jan 2016 till June 2017.**

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

Jan 2016 15675.66 14418.21 16933.11 13752.56 17598.77

Feb 2016 15974.64 14710.09 17239.19 14040.67 17908.61

Mar 2016 15900.60 14629.03 17172.16 13955.91 17845.29

Apr 2016 15775.80 14497.13 17054.46 13820.25 17731.35

May 2016 15695.33 14409.24 16981.42 13728.42 17662.23

Jun 2016 15699.91 14405.80 16994.02 13720.74 17679.08

Jul 2016 15745.34 14442.68 17047.99 13753.09 17737.58

Aug 2016 16045.37 14732.09 17358.65 14036.88 18053.86

Sep 2016 16250.34 14926.46 17574.22 14225.65 18275.03

Oct 2016 15726.81 14398.29 17055.33 13695.01 17758.61

Nov 2016 15710.15 14372.35 17047.95 13664.16 17756.14

Dec 2016 15787.47 15299.86 16275.09 15041.73 16533.22

Jan 2017 16328.31 14455.16 18201.46 13463.57 19193.04

Feb 2017 16637.43 14753.85 18521.01 13756.74 19518.12

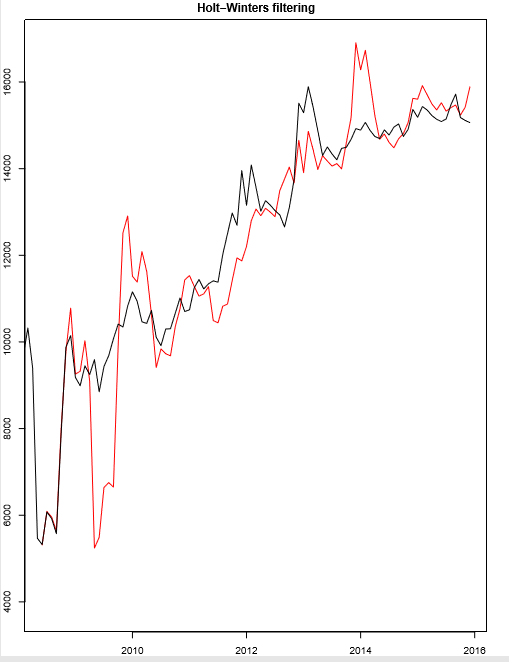
Mar 2017 16558.04 14667.49 18448.59 13666.70 19449.39

Apr 2017 16425.84 14528.79 18322.89 13524.55 19327.13

May 2017 16339.84 14435.66 18244.03 13427.64 19252.04

Jun 2017 16342.42 14429.81 18255.02 13417.34 19267.49

Now lets us apply Holt-Winters exponential smoothing with trend & see how the forecasted values looks like.



**Below are the forecasted values using holt winters model using exponential smoothing.**

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

Jan 2016 15675.66 14418.21 16933.11 13752.56 17598.77

Feb 2016 15974.64 14710.09 17239.19 14040.67 17908.61

Mar 2016 15900.60 14629.03 17172.16 13955.91 17845.29

Apr 2016 15775.80 14497.13 17054.46 13820.25 17731.35

May 2016 15695.33 14409.24 16981.42 13728.42 17662.23

Jun 2016 15699.91 14405.80 16994.02 13720.74 17679.08

Jul 2016 15745.34 14442.68 17047.99 13753.09 17737.58

Aug 2016 16045.37 14732.09 17358.65 14036.88 18053.86

Sep 2016 16250.34 14926.46 17574.22 14225.65 18275.03

Oct 2016 15726.81 14398.29 17055.33 13695.01 17758.61

Nov 2016 15710.15 14372.35 17047.95 13664.16 17756.14

Dec 2016 15787.47 15299.86 16275.09 15041.73 16533.22

Jan 2017 16328.31 14455.16 18201.46 13463.57 19193.04

Feb 2017 16637.43 14753.85 18521.01 13756.74 19518.12

Mar 2017 16558.04 14667.49 18448.59 13666.70 19449.39

Apr 2017 16425.84 14528.79 18322.89 13524.55 19327.13

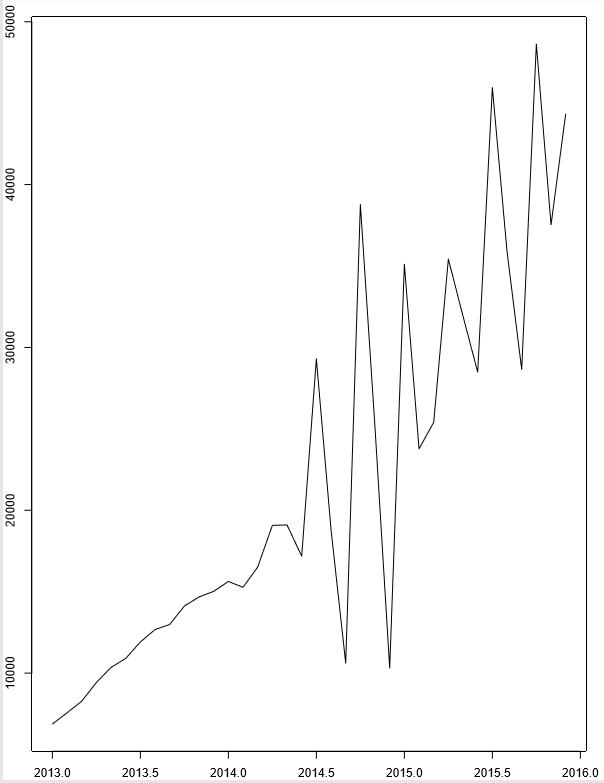
May 2017 16339.84 14435.66 18244.03 13427.64 19252.04

Jun 2017 16342.42 14429.81 18255.02 13417.34 19267.49

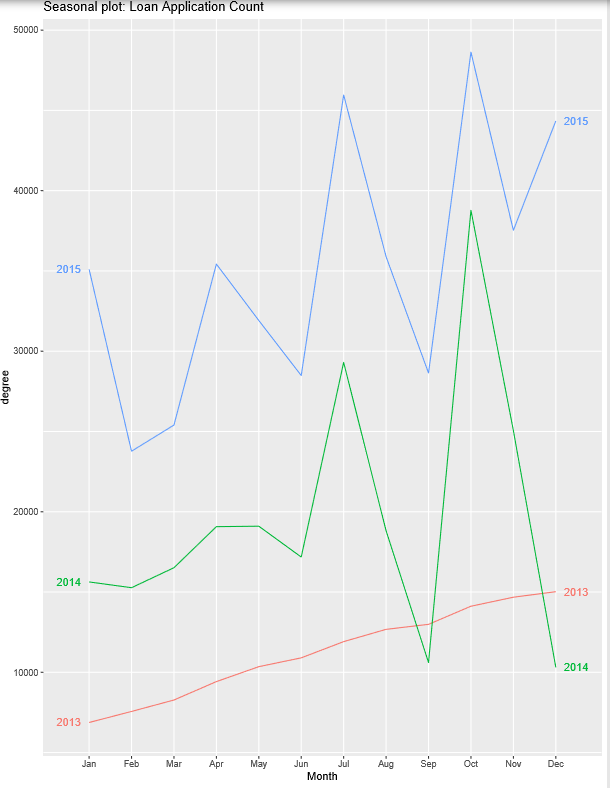
**Forecasting Loan Application count:**

For Loan application we can see that there is an exponential increase of total application from year 2007 till 2012, from year 2013 the loan application got stabilised to some extent.

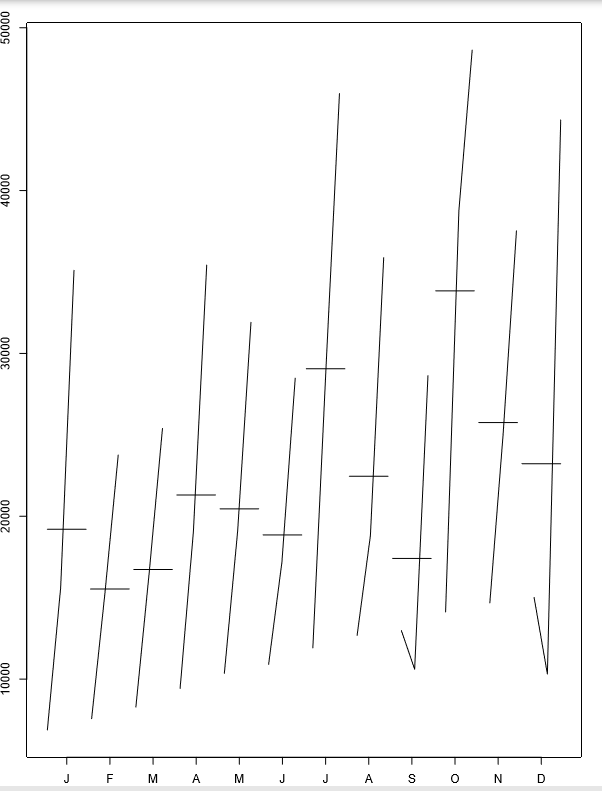
We therefore will consider records from year 2013 for forecasting loan application.



**ggseasonal plot for Loan Application Count:**

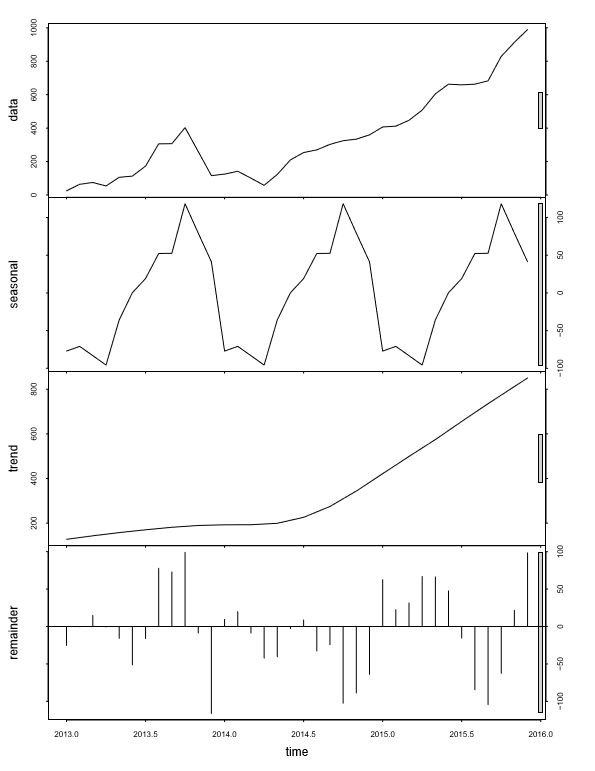


**Month plot for Loan Application:**



From above 3 plots we can infer that there is increasing trend present whereas for year 2014-15 some degree of seasonality is present. Overall trend looks more prominent & we can infer that series has increasing trend.

**Decompose data into Trend, seasonality & residuals:**



**SPLIT THE DATA INTO TRAIN AND TEST DATA SETS:**

Now we will divide the data into train & test. We have observed that loan application count has increased exponentially from year 2007 till 2012, we therefore put data from Jan 2013 till 2015 June in train & data from July 2015 to December 2015 into test.

**We now apply below 3 models & find corresponding MAPE value**

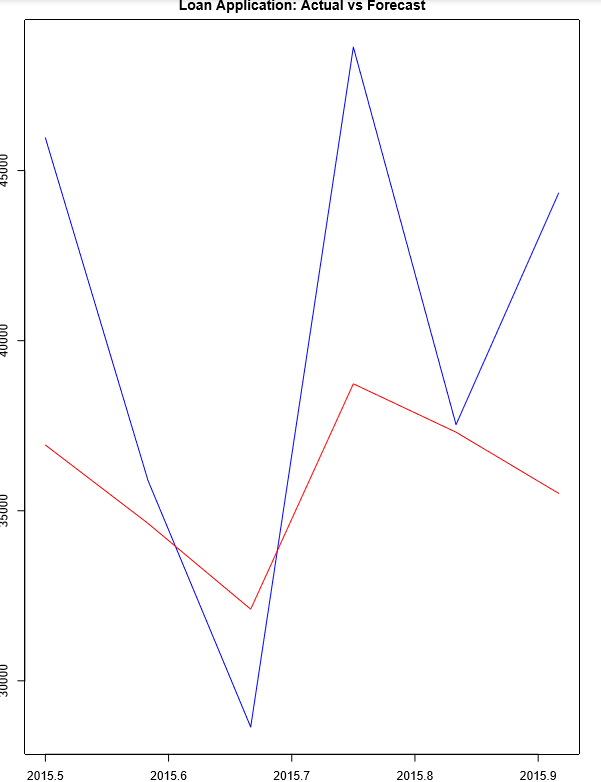
* Random walk with drift
* Holt winters
* ARIMA

Upon applying these 3 models we obtained below MAPE values.

|  |  |
| --- | --- |
| **Model** | **MAPE** |
| **Random walk with drift** | **0.5106** |
| **Holt winters** | **0.5800** |
| **ARIMA** | **0.1284** |

**We could see that MAPE value is least for ARIMA model, therefore we will forecast using ARIMA model.**

**Actual vs Forecast using ARIMA model**



**Forecast using ARIMA model:**

Below is the forecasted value for next 18 months.

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

Jan 2016 45792.02 40117.55 51466.49 37113.67 54470.38

Feb 2016 42106.63 36230.56 47982.71 33119.95 51093.31

Mar 2016 46810.92 40930.54 52691.29 37817.66 55804.17

Apr 2016 48354.06 41118.04 55590.07 37287.53 59420.58

May 2016 46189.15 38937.87 53440.44 35099.28 57279.03

Jun 2016 49542.87 42278.07 56807.68 38432.31 60653.43

Jul 2016 51068.21 43116.92 59019.50 38907.76 63228.65

Aug 2016 49943.31 41989.69 57896.92 37779.30 62107.32

Sep 2016 52435.62 44456.86 60414.38 40233.17 64638.08

Oct 2016 53896.31 45514.10 62278.52 41076.83 66715.80

Nov 2016 53479.56 45075.47 61883.65 40626.61 66332.51

Dec 2016 55425.20 46984.24 63866.15 42515.87 68334.52

Jan 2017 56808.33 48104.03 65512.63 43496.26 70120.41

Feb 2017 56871.73 48124.84 65618.62 43494.52 70248.95

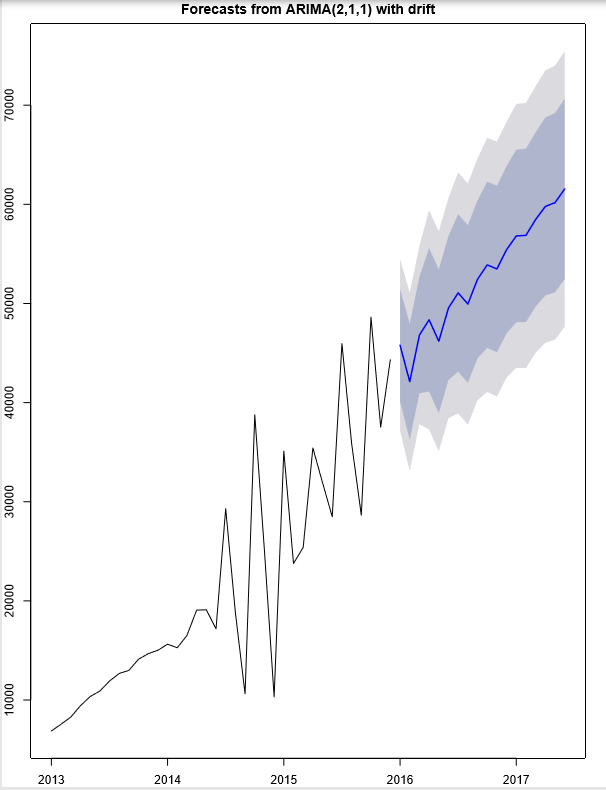
Mar 2017 58472.29 49678.22 67266.37 45022.91 71921.67

Apr 2017 59781.32 50799.24 68763.39 46044.42 73518.21

May 2017 60168.96 51129.23 69208.69 46343.88 73994.04

Jun 2017 61552.99 52457.81 70648.17 47643.12 75462.87

FIGURE 6.2: FORECAST OF NUMBER OF LOAN APPLICATIONS



**Summary of findings:**

We have seen that average loan amount & loan application count has increasing trend present whereas we cannot confirm that both the series has presence of strong seasonality. We applied 3 models & the model which has given least MAPE we have forecasted with that model.

Below is the table containing forecasted values for Average Loan amount & Loan application count

TABLE 6.3: FORECAST SUMMARY FOR AVERAGE LOAN AMOUNT AND NUMBER OF LOAN APPLICATIONS

|  |  |  |
| --- | --- | --- |
| Period | Point forecast for Loan Amount | Point Forecast for Loan application |
| Jan 2016 | 15675.66 | 45792 |
| Feb 2016 | 15974.64 | 42106 |
| Mar 2016 | 15900.60 | 46810. |
| Apr 2016 | 15775.80 | 48354 |
| May 2016 | 15695.33 | 46189 |
| June 2016 | 15699.91 | 49542 |
| July 2016 | 15745.34 | 51068 |
| Aug 2016 | 16045.37 | 49943 |
| Sep 2016 | 16250.34 | 52435 |
| Oct 2016 | 15726.81 | 53896 |
| Nov 2016 | 15710.15 | 53479 |
| Dec 2016 | 15787.47 | 55425 |
| Jan 2017 | 16328.31 | 56808 |
| Feb 2017 | 16637.43 | 56871 |
| Mar 2017 | 16558.04 | 58472 |
| Apr 2017 | 16425.84 | 59781 |
| May 2017 | 16339.84 | 60168 |
| June 2017 | 16342.42 | 61552 |

**CHAPTER – 7**

**SENTIMENT ANALYSIS ON TWITTER DATA OF LENDING CLUB**

**SENTIMENT ANALYSIS OF LENDING CLUB WITH TWITTER DATA:**

Now, we will try to analyze the sentiments of tweets made by a Twitter handle.

1. **EXTRACTING TWEETS USING TWITTER APPLICATION**

* Invoke Twitter API using the app we have created and using the keys and access tokens we got through the app.
* Check the extracted Tweets format.

1. **CLEANING THE TWEETS FOR FURTHER ANALYSIS**

* Total of 16 variables using ‘userTimeline’ function, out of which Text variable is significant for use.
* Remove hashtags and URLs from the text field.

1. **WORDCLOUD FOR TWEETS**



We can see repetitive promotional Tweets like "Who do you know that would benefit from a low rate". So, removing those from, below word could has been formed



FIGURE 7.1: WORD CLOUD OF LENDING CLUB TWEETS

1. **GET SENTIMENT SCORE OF TWEETS USING “SYUZHET” PACKAGE**

* get\_sentiment function of ‘Syuzhet’ package will identify the sentiment score, and these scores are used to segregate the tweets as positive, neutral and negative. The results of this package are as follows:

|  |  |
| --- | --- |
| **Nature of the tweet** | **Number of tweets** |
| **Positive** | **124** |
| **Neutral** | **154** |
| **Negative** | **62** |

* Most Positive and Most Negative Tweets can also be identified

**Most Positive Tweet:**

[1] "Happy Monday everyone\n\nAs we get ready for another great week we want we want to send out a friendly reminder\n\nFe<U+2026>"

[2] "Happy Monday everyone \n\nAs we get ready for another great week we want we want to send out a friendly reminder<U+2026>"

**Most Negative Tweet:**

[1] "Beware of this company installed on one of its members converted to him thousand Saudi Riyals taxes until the mo<U+2026>"

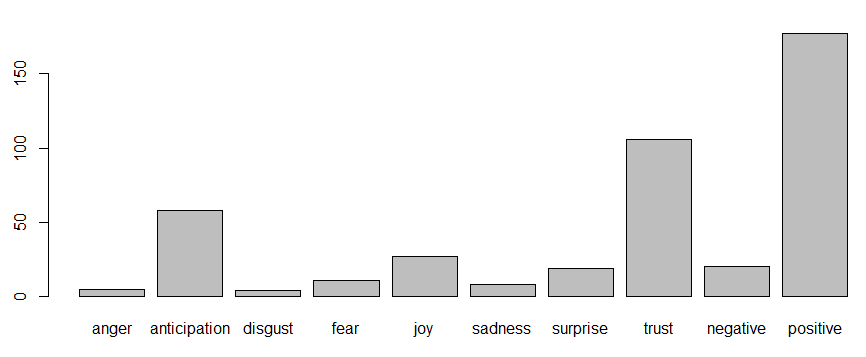
[2] "Lendingclub Corporation LC Shorted Shares Declined By"

[3] "The failure rates for"

1. **CALL NRC SENTIMENT DICTIONARY TO CATEGORIZE THE TWEETS BASED ON DIFFERENT EMOTIONS:**

get\_nrc\_sentiment function is used to call NRC sentiment dictionary to calculate the presence of 10 different emotions – anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative and positive, and scores them.

And it will categorise the tweets into one of the emotion types, as follows:



**LABELS**

X-AXIS: EMOTION

Y-AXIS: NUMBER OF TWEETS

FIGURE 7.2: CLASSFICATION OF LENDING CLUB TWEETS BASED ON EMOTIONS

1. **CART MODEL**

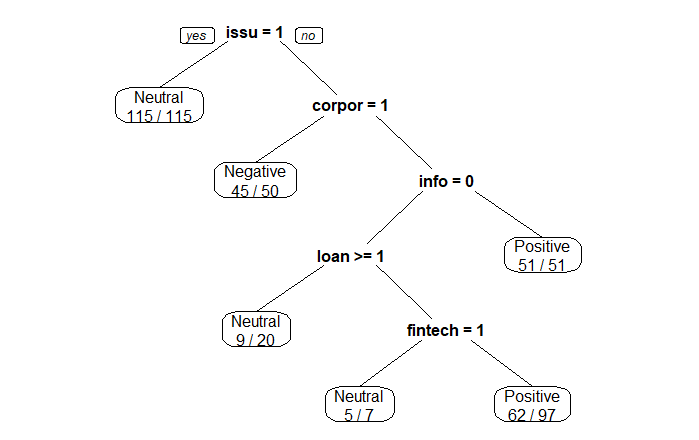


FIGURE 7.3: CART MODEL OUPUT FOR TWEETS OF LENDING CLUB

**INTERPRETATION OF CART MODEL OUTPUT:**

Word “issu” seems to be the most significant word in determining the nature of the tweet, as the CART output shows that if the word “issu” appears atleast once, then the tweet is a neutral one and if “issu” does not appear atleast once, then the nature of the tweet depends on the presence of the word “corpor”

**FINDINGS FROM SENTIMENT ANALYSIS:**

From the above analysis, we could conclude that there are **fairly positive sentiments towards Lending Club among people** in Twitter.

More percentage of extracted tweets belong to **“positive” and “trust**” emotion, indicating that Lending Club has good reputation among the people

**TEXT MINING ON “DESCRIPTION OF PURPOSE OF LOAN”:**

“desc” is a variable in the Lending club loan data, which describes the purpose of loan written in the words of customer. Check which words has been used most for borrowing loans. We would create Word cloud for Defaulters and Non-Defaulters.

**Word Cloud for Defaulters:**



**Word Cloud for Non-Defaulters:**



**FINDINGS:** Words like credit card, debt consolidation, pay, loan are most used for both the default and non-default cases. There is no significant difference between defaulters and non-defaulters in terms of words used by them while mentioning the purpose of the loan in the loan application

**CHAPTER - 8**

**RECOMMENDATIONS AND CONCLUSIONS**

**SCOPE OF RECOMMENDATION:**

* For our analysis, Lending club customer data, for the duration 2007-2015, is taken
* Loan forecast has been given for next 18 months i.e. Jan 2016 till June 2017.
* Sentimental analysis is done on Twitter data.
* Targeted Audience: Lending Club management

**LIST OF RECOMMENDATIONS:**

* We have done modelling on loan default using different techniques such as Logistic, SVM, Random Forest. Since these models have the ability to predict the risk involved in granting a loan to a new customer, Lending Club can use these models to avoid giving loans to high risk customers, thereby avoiding the bad debt/monetary loss.
* Among all variables, “grade”, “annual income”, “address State/province” & “employment length” are most important. Lending Club can give more importance to these variables while granting loan to a new customer.
* Customer profiling is done with the help of clustering (K means clustering), this helps to segregate borrowers based on low, medium & high risk. Lending Club can use this tool to devise suitable strategy for respective groups which will help to improve business revenue. At the same time Lending Club can come up with suitable business actions for different categories like:

For low-risk customers - Lending Club can provide promotional, retail, cross-sell and up sell offers to low-risk customers to retain them for longer periods.

For high-risk customers – suitable and personalized recovery actions can be employed

* Time series model helps to predict future loan/credit demand. Lending Club can utilize this information to take necessary steps to meet growing future demands. We can see that their good growth of Credit demand for upcoming 18 months therefore Lending club should take requisite steps to meet future credit needs of the borrower. This will help the Lending Club in liquidity management & financial planning.
* Sentimental analysis depicts fairly positive sentiments towards Lending club among people in Twitter. Lending Club can constantly monitor social media platform such as Twitter & can perceive people’s emotions. This will help Lending Club to know adverse future people’s emotions thus it can take corrective measures beforehand. Good percentage of tweets indicate neutral opinion of the people and therefore Lending Club can take measures to convert these into positive opinion.
* Top 3 states in loan disbursement are California (CA), Texas (TX) & New York (NY), Lending Club can target people from other states through various advertisement & marketing campaigns.

**BIBLIOGRAPHY**

* + https://www.roselladb.com/credit-risk-analysis.htm
  + https://towardsdatascience.com/a-gentle-introduction-to-credit-risk-modeling-with-data-science-part-2-d7b87806c9df
  + https://www.datasciencecentral.com/profiles/blogs/credit-risk-prediction-using-artificial-neural-network-algorithm
  + <ftp://ftp.repec.org/opt/ReDIF/RePEc/ami/articles/14_1_3.pdf>
* <https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/>
* <https://nrc-publications.canada.ca/eng/view/fulltext/?id=e8c7556d-9f94-466f-a1e5-72cdf9b9513f>
* <https://towardsdatascience.com/simple-fast-exploratory-data-analysis-in-r-with-dataexplorer-package-e055348d9619>
* http://r-statistics.co/Outlier-Treatment-With-R.html

**ANNEXURE**

**R-CODE OF THE MODEL BUILT:**

**Below are the attachments for code written in R:**

**    **

**DATA DICTIONARY:**

|  |  |
| --- | --- |
| LoanStatNew | Description |
| addr\_state | The state provided by the borrower in the loan application |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| annual\_inc\_joint | The combined self-reported annual income provided by the co-borrowers during registration |
| application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| collection\_recovery\_fee | post charge off collection fee |
| collections\_12\_mths\_ex\_med | Number of collections in 12 months excluding medical collections |
| delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| desc | Loan description provided by the borrower |
| 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. |
| dti\_joint | A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined 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. |
| emp\_title | The job title supplied by the Borrower when applying for the loan.\* |
| 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. |
| funded\_amnt | The total amount committed to that loan at that point in time. |
| funded\_amnt\_inv | The total amount committed by investors for that loan at that point in time. |
| grade | LC assigned loan grade |
| home\_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| id | A unique LC assigned ID for the loan listing. |
| initial\_list\_status | The initial listing status of the loan. Possible values are – W, F |
| inq\_last\_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| installment | The monthly payment owed by the borrower if the loan originates. |
| int\_rate | Interest Rate on the loan |
| is\_inc\_v | Indicates if income was verified by LC, not verified, or if the income source was verified |
| issue\_d | The month which the loan was funded |
| last\_credit\_pull\_d | The most recent month LC pulled credit for this loan |
| last\_fico\_range\_high | The upper boundary range the borrower’s last FICO pulled belongs to. |
| last\_fico\_range\_low | The lower boundary range the borrower’s last FICO pulled belongs to. |
| last\_pymnt\_amnt | Last total payment amount received |
| last\_pymnt\_d | Last month payment was received |
| 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. |
| loan\_status | Current status of the loan |
| member\_id | A unique LC assigned Id for the borrower member. |
| mths\_since\_last\_delinq | The number of months since the borrower's last delinquency. |
| mths\_since\_last\_major\_derog | Months since most recent 90-day or worse rating |
| mths\_since\_last\_record | The number of months since the last public record. |
| next\_pymnt\_d | Next scheduled payment date |
| open\_acc | The number of open credit lines in the borrower's credit file. |
| out\_prncp | Remaining outstanding principal for total amount funded |
| out\_prncp\_inv | Remaining outstanding principal for portion of total amount funded by investors |
| policy\_code | publicly available policy\_code=1 new products not publicly available policy\_code=2 |
| pub\_rec | Number of derogatory public records |
| purpose | A category provided by the borrower for the loan request. |
| pymnt\_plan | Indicates if a payment plan has been put in place for the loan |
| recoveries | post charge off gross recovery |
| 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. |
| sub\_grade | LC assigned loan subgrade |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| title | The loan title provided by the borrower |
| total\_acc | The total number of credit lines currently in the borrower's credit file |
| total\_pymnt | Payments received to date for total amount funded |
| total\_pymnt\_inv | Payments received to date for portion of total amount funded by investors |
| total\_rec\_int | Interest received to date |
| total\_rec\_late\_fee | Late fees received to date |
| total\_rec\_prncp | Principal received to date |
| url | URL for the LC page with listing data. |
| verified\_status\_joint | Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified |
| zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |
| open\_acc\_6m | Number of open trades in last 6 months |
| open\_il\_6m | Number of currently active installment trades |
| open\_il\_12m | Number of installment accounts opened in past 12 months |
| open\_il\_24m | Number of installment accounts opened in past 24 months |
| mths\_since\_rcnt\_il | Months since most recent installment accounts opened |
| total\_bal\_il | Total current balance of all installment accounts |
| il\_util | Ratio of total current balance to high credit/credit limit on all install acct |
| open\_rv\_12m | Number of revolving trades opened in past 12 months |
| open\_rv\_24m | Number of revolving trades opened in past 24 months |
| max\_bal\_bc | Maximum current balance owed on all revolving accounts |
| all\_util | Balance to credit limit on all trades |
| total\_rev\_hi\_lim | Total revolving high credit/credit limit |
| inq\_fi | Number of personal finance inquiries |
| total\_cu\_tl | Number of finance trades |
| inq\_last\_12m | Number of credit inquiries in past 12 months |
| acc\_now\_delinq | The number of accounts on which the borrower is now delinquent. |
| tot\_coll\_amt | Total collection amounts ever owed |
| tot\_cur\_bal | Total current balance of all accounts |

**IMPORTANT CAPTURED OUTPUTS FROM R:**

**Import data into R:**

|  |
| --- |
| > Loan1 = read.csv(file.choose(),header = TRUE)  > dim(Loan1)  [1] 887379 74  > loan\_data = Loan1 |

**Check the structure of the data:**

|  |
| --- |
| > str(loan\_data)  'data.frame': 887379 obs. of 74 variables:  $ id : int 1077501 1077430 1077175 1076863 1075358 1075269 1069639 1072053 1071795 1071570 ...  $ member\_id : int 1296599 1314167 1313524 1277178 1311748 1311441 1304742 1288686 1306957 1306721 ...  $ loan\_amnt : num 5000 2500 2400 10000 3000 ...  $ funded\_amnt : num 5000 2500 2400 10000 3000 ...  $ funded\_amnt\_inv : num 4975 2500 2400 10000 3000 ...  $ term : Factor w/ 2 levels " 36 months"," 60 months": 1 2 1 1 2 1 2 1 2 2 ...  $ int\_rate : num 10.6 15.3 16 13.5 12.7 ...  $ installment : num 162.9 59.8 84.3 339.3 67.8 ...  $ grade : Factor w/ 7 levels "A","B","C","D",..: 2 3 3 3 2 1 3 5 6 2 ...  $ sub\_grade : Factor w/ 35 levels "A1","A2","A3",..: 7 14 15 11 10 4 15 21 27 10 ...  $ emp\_title : Factor w/ 299273 levels "","'Property Manager",..: 1 224800 1 9376 282199 285977 246848 171062 1 256905 ...  $ emp\_length : Factor w/ 12 levels "< 1 year","1 year",..: 3 1 3 3 2 5 10 11 6 1 ...  $ home\_ownership : Factor w/ 6 levels "ANY","MORTGAGE",..: 6 6 6 6 6 6 6 6 5 6 ...  $ annual\_inc : num 24000 30000 12252 49200 80000 ...  $ verification\_status : Factor w/ 3 levels "Not Verified",..: 3 2 1 2 2 2 1 2 2 3 ...  $ issue\_d : Factor w/ 103 levels "Apr-2008","Apr-2009",..: 22 22 22 22 22 22 22 22 22 22 ...  $ loan\_status : Factor w/ 10 levels "Charged Off",..: 6 1 6 6 2 6 2 6 1 1 ...  $ pymnt\_plan : Factor w/ 2 levels "n","y": 1 1 1 1 1 1 1 1 1 1 ...  $ url : Factor w/ 887379 levels "https://www.lendingclub.com/browse/loanDetail.action?loan\_id=1000007",..: 21292 21256 21242 21220 20692 20684 19191 19811 19796 19657 ...  $ desc : Factor w/ 124471 levels "","- Pay off Dell Financial: $ 1300.00 - Pay off IRS for 2005: $ 1400.00 - Pay off Mac Comp : $ 1700.00 - Pay o"| \_\_truncated\_\_,..: 113401 113406 1 113257 113233 1 112346 111630 113231 111633 ...  $ purpose : Factor w/ 14 levels "car","credit\_card",..: 2 1 12 10 10 14 3 1 12 10 ...  $ title : Factor w/ 63146 levels "","'08 & '09 Roth IRA Investments",..: 10497 4976 52500 50874 50267 42595 36948 7266 24372 6113 ...  $ zip\_code : Factor w/ 935 levels "007xx","008xx",..: 810 296 572 856 909 803 267 839 897 729 ...  $ addr\_state : Factor w/ 51 levels "AK","AL","AR",..: 4 11 15 5 38 4 28 5 5 44 ...  $ dti : num 27.65 1 8.72 20 17.94 ...  $ delinq\_2yrs : num 0 0 0 0 0 0 0 0 0 0 ...    $ earliest\_cr\_line : Factor w/ 698 levels "","Apr-1955",..: 265 43 572 210 276 575 342 287 48 690 ...  $ inq\_last\_6mths : num 1 5 2 1 0 3 1 2 2 0 ...  $ mths\_since\_last\_delinq : num NA NA NA 35 38 NA NA NA NA NA ...  $ mths\_since\_last\_record : num NA NA NA NA NA NA NA NA NA NA ...  $ open\_acc : num 3 3 2 10 15 9 7 4 11 2 ...  $ pub\_rec : num 0 0 0 0 0 0 0 0 0 0 ...  $ revol\_bal : num 13648 1687 2956 5598 27783 ...  $ revol\_util : num 83.7 9.4 98.5 21 53.9 28.3 85.6 87.5 32.6 36.5 ...  $ total\_acc : num 9 4 10 37 38 12 11 4 13 3 ...  $ initial\_list\_status : Factor w/ 2 levels "f","w": 1 1 1 1 1 1 1 1 1 1 ...  $ out\_prncp : num 0 0 0 0 767 ...  $ out\_prncp\_inv : num 0 0 0 0 767 ...  $ total\_pymnt : num 5861 1009 3004 12226 3242 ...  $ total\_pymnt\_inv : num 5832 1009 3004 12226 3242 ...  $ total\_rec\_prncp : num 5000 456 2400 10000 2233 ...  $ total\_rec\_int : num 861 435 604 2209 1009 ...  $ total\_rec\_late\_fee : num 0 0 0 17 0 ...  $ recoveries : num 0 117 0 0 0 ...  $ collection\_recovery\_fee : num 0 1.11 0 0 0 0 0 0 2.09 2.52 ...  $ last\_pymnt\_d : Factor w/ 99 levels "","Apr-2008",..: 42 7 58 42 43 42 43 42 6 80 ...  $ last\_pymnt\_amnt : num 171.6 119.7 649.9 357.5 67.8 ...  $ next\_pymnt\_d : Factor w/ 101 levels "","Apr-2008",..: 1 1 1 1 35 1 35 1 1 1 ...  $ last\_credit\_pull\_d : Factor w/ 104 levels "","Apr-2009",..: 43 102 43 42 43 104 43 25 14 67 ...  $ collections\_12\_mths\_ex\_med : num 0 0 0 0 0 0 0 0 0 0 ...  $ mths\_since\_last\_major\_derog: num NA NA NA NA NA NA NA NA NA NA ...  $ policy\_code : num 1 1 1 1 1 1 1 1 1 1 ...  $ application\_type : Factor w/ 2 levels "INDIVIDUAL","JOINT": 1 1 1 1 1 1 1 1 1 1 ...  $ annual\_inc\_joint : num NA NA NA NA NA NA NA NA NA NA ...  $ dti\_joint : num NA NA NA NA NA NA NA NA NA NA ...  $ verification\_status\_joint : Factor w/ 4 levels "","Not Verified",..: 1 1 1 1 1 1 1 1 1 1 ...  $ acc\_now\_delinq : num 0 0 0 0 0 0 0 0 0 0 ...  $ tot\_coll\_amt : num NA NA NA NA NA NA NA NA NA NA ...  $ tot\_cur\_bal : num NA NA NA NA NA NA NA NA NA NA ...  $ open\_acc\_6m : num NA NA NA NA NA NA NA NA NA NA ...  $ open\_il\_6m : num NA NA NA NA NA NA NA NA NA NA ...  $ open\_il\_12m : num NA NA NA NA NA NA NA NA NA NA ...  $ open\_il\_24m : num NA NA NA NA NA NA NA NA NA NA ...  $ mths\_since\_rcnt\_il : num NA NA NA NA NA NA NA NA NA NA ...  $ total\_bal\_il : num NA NA NA NA NA NA NA NA NA NA ...  $ il\_util : num NA NA NA NA NA NA NA NA NA NA ...  $ open\_rv\_12m : num NA NA NA NA NA NA NA NA NA NA ...  $ open\_rv\_24m : num NA NA NA NA NA NA NA NA NA NA ...  $ max\_bal\_bc : num NA NA NA NA NA NA NA NA NA NA ...  $ all\_util : num NA NA NA NA NA NA NA NA NA NA ...  $ total\_rev\_hi\_lim : num NA NA NA NA NA NA NA NA NA NA ...  $ inq\_fi : num NA NA NA NA NA NA NA NA NA NA ...  $ total\_cu\_tl : num NA NA NA NA NA NA NA NA NA NA ...  $ inq\_last\_12m : num NA NA NA NA NA NA NA NA NA NA ... |

**Finding missing value percentage:**

|  |
| --- |
| > missingdata$nmsg  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 4  [15] 0 0 0 0 0 0 0 0 0 0 0 29 0 29  [29] 454312 750326 29 29 0 502 29 0 0 0 0 0 0 0  [43] 0 0 0 0 0 0 0 145 665676 0 0 886868 886870 0  [57] 29 70276 70276 866007 866007 866007 866007 866569 866007 868762 866007 866007 866007 866007  [71] 70276 866007 866007 866007  Levels: 0 145 29 4 454312 502 665676 70276 750326 866007 866569 868762 886868 886870  > missingdata$nmsg=as.numeric(levels(missingdata$nmsg))[missingdata$nmsg]  > missingdata$nmsg  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 4  [15] 0 0 0 0 0 0 0 0 0 0 0 29 0 29  [29] 454312 750326 29 29 0 502 29 0 0 0 0 0 0 0  [43] 0 0 0 0 0 0 0 145 665676 0 0 886868 886870 0  [57] 29 70276 70276 866007 866007 866007 866007 866569 866007 868762 866007 866007 866007 866007  [71] 70276 866007 866007 866007  > missingdata=cbind(missingdata,percmissing=as.integer(missingdata$nmsg/ncol\*100))  > missingdata$percmissing  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 51 84 0 0 0 0  [35] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 75 0 0 99 99 0 0 7 7 97 97 97 97 97 97 97 97 97  [69] 97 97 7 97 97 97 |

**Dropping of variables with very high missing values (>52% missing):**

|  |
| --- |
| drops=as.character(subset(missingdata,missingdata$percmissing>52)[,1])  > length(drops)  [1] 18  > table(drops)  drops  all\_util annual\_inc\_joint dti\_joint  1 1 1  il\_util inq\_fi inq\_last\_12m  1 1 1  max\_bal\_bc mths\_since\_last\_major\_derog mths\_since\_last\_record  1 1 1  mths\_since\_rcnt\_il open\_acc\_6m open\_il\_12m  1 1 1  open\_il\_24m open\_il\_6m open\_rv\_12m  1 1 1  open\_rv\_24m total\_bal\_il total\_cu\_tl  1 1 1 |
| > loan\_data=loan\_data[,!(names(loan\_data) %in% drops)]  > dim(loan\_data)  [1] 887379 56  > names(loan\_data)  [1] "id" "member\_id" "loan\_amnt"  [4] "funded\_amnt" "funded\_amnt\_inv" "term"  [7] "int\_rate" "installment" "grade"  [10] "sub\_grade" "emp\_title" "emp\_length"  [13] "home\_ownership" "annual\_inc" "verification\_status"  [16] "issue\_d" "loan\_status" "pymnt\_plan"  [19] "url" "desc" "purpose"  [22] "title" "zip\_code" "addr\_state"  [25] "dti" "delinq\_2yrs" "earliest\_cr\_line"  [28] "inq\_last\_6mths" "mths\_since\_last\_delinq" "open\_acc"  [31] "pub\_rec" "revol\_bal" "revol\_util"  [34] "total\_acc" "initial\_list\_status" "out\_prncp"  [37] "out\_prncp\_inv" "total\_pymnt" "total\_pymnt\_inv"  [40] "total\_rec\_prncp" "total\_rec\_int" "total\_rec\_late\_fee"  [43] "recoveries" "collection\_recovery\_fee" "last\_pymnt\_d"  [46] "last\_pymnt\_amnt" "next\_pymnt\_d" "last\_credit\_pull\_d"  [49] "collections\_12\_mths\_ex\_med" "policy\_code" "application\_type"  [52] "verification\_status\_joint" "acc\_now\_delinq" "tot\_coll\_amt"  [55] "tot\_cur\_bal" "total\_rev\_hi\_lim" |

**Sample EDA Code:**

**Bivariate plot:**

|  |
| --- |
| ggplot(data=loan\_data,aes(loan\_status,annual\_inc))+geom\_bar(stat="identity",aes(fill=loan\_status))+  + coord\_flip()+xlab("")+ylab("annual inc")+ggtitle("Annual income vs loan\_status")+xlab("loan\_status") |

**Univariate plot (Histogram)**

|  |
| --- |
| hist(loan\_data$annual\_inc,main="Annual Income",xlab="annual\_inc",col="cyan") |

**Univariate plot(barplot)**

|  |
| --- |
| barplot(table(loan\_data$emp\_length) ,main="Employment length" , xlab="emp\_length",col="cyan") |

**Conversion of factor variables into numeric:**

We have converted required factor variables into numeric, below is example of one of the variable “emp\_length”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| loan\_data$emp\_length <- gsub(" years" , "", loan\_data$emp\_length)  loan\_data$emp\_length <- gsub(" year" , "", loan\_data$emp\_length)  loan\_data$emp\_length <- ifelse(loan\_data$emp\_length == "10+", 10, loan\_data$emp\_length)  loan\_data$emp\_length <- ifelse(loan\_data$emp\_length == "< 1", 0, loan\_data$emp\_length)  loan\_data$emp\_length <- as.numeric(loan\_data$emp\_length)   |  | | --- | | table(loan\_data$emp\_length)  0 1 2 3 4 5 6 7 8 9 10  70605 57095 78870 70026 52529 55704 42950 44594 43955 34657 291569 | |  | | |  | | --- | |  | | |

**Outlier detection with boxplots:**

|  |
| --- |
| ggplot(loan\_data, aes(loan\_data$loan\_status,loan\_data$total\_rev\_hi\_lim)) +  + geom\_boxplot(aes(fill = loan\_data$loan\_status))+coord\_flip()+xlab("")+  + ylab("Total revolving high credit/credit limit")+ggtitle("Total revolving high credit/credit limit Vs Loan Status")+xlab("Loan Status") |

**Multivariate analysis using Corr plot:**

corrplot(corrplot(cor(data\_num1)), na.label = " ")

**Imputation of data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| library(mice)  md.pattern(loan\_data)  mice\_imputed\_Data <- mice(loan\_data, m=5, maxit = 5, method = 'pmm', seed = 500)  summary(mice\_imputed\_Data)   |  | | --- | | Column.0 id member\_id loan\_amnt term int\_rate  Min. : 1 Min. : 54734 Min. : 70473 Min. : 500 Min. :0.0 Min. : 5.32  1st Qu.:221846 1st Qu.: 9206643 1st Qu.:10877134 1st Qu.: 8000 1st Qu.:0.0 1st Qu.: 9.99  Median :443690 Median :34433267 Median :37095283 Median :13000 Median :0.0 Median :12.99  Mean :443690 Mean :32465133 Mean :35001825 Mean :14755 Mean :0.3 Mean :13.25  3rd Qu.:665535 3rd Qu.:54908135 3rd Qu.:58471347 3rd Qu.:20000 3rd Qu.:1.0 3rd Qu.:16.20  Max. :887379 Max. :68617057 Max. :73544841 Max. :35000 Max. :1.0 Max. :28.99  installment grade emp\_length home\_ownership annual\_inc verification\_status  Min. : 15.67 Min. :1.000 Min. : 0.000 Min. :1.000 Min. : 0 Min. :0.0000  1st Qu.: 260.70 1st Qu.:2.000 1st Qu.: 3.000 1st Qu.:1.000 1st Qu.: 45000 1st Qu.:0.0000  Median : 382.55 Median :3.000 Median : 6.000 Median :2.000 Median : 65000 Median :1.0000  Mean : 436.72 Mean :2.798 Mean : 6.009 Mean :1.902 Mean : 75028 Mean :0.6994  3rd Qu.: 572.60 3rd Qu.:4.000 3rd Qu.:10.000 3rd Qu.:3.000 3rd Qu.: 90000 3rd Qu.:1.0000  Max. :1445.46 Max. :7.000 Max. :11.000 Max. :4.000 Max. :9500000 Max. :1.0000  loan\_status purpose dti delinq\_2yrs inq\_last\_6mths mths\_since\_last\_delinq  Min. :1.000 Min. : 1.000 Min. : 0.00 Min. :0 Min. : 0.0000 Min. : 0.00  1st Qu.:1.000 1st Qu.: 3.000 1st Qu.:11.91 1st Qu.:0 1st Qu.: 0.0000 1st Qu.: 30.00  Median :1.000 Median : 3.000 Median :17.65 Median :0 Median : 0.0000 Median : 34.00  Mean :1.154 Mean : 3.571 Mean :18.13 Mean :0 Mean : 0.6946 Mean : 34.09  3rd Qu.:1.000 3rd Qu.: 3.000 3rd Qu.:23.94 3rd Qu.:0 3rd Qu.: 1.0000 3rd Qu.: 37.00  Max. :3.000 Max. :14.000 Max. :41.94 Max. :0 Max. :33.0000 Max. :188.00  open\_acc pub\_rec revol\_bal revol\_util total\_acc initial\_list\_status  Min. : 0.00 Min. : 0.0000 Min. : 0 Min. : 0.00 Min. : 1.00 Min. :0.0000  1st Qu.: 8.00 1st Qu.: 0.0000 1st Qu.: 6443 1st Qu.: 37.70 1st Qu.: 17.00 1st Qu.:0.0000  Median :11.00 Median : 0.0000 Median : 11875 Median : 56.00 Median : 24.00 Median :0.0000  Mean :11.55 Mean : 0.1953 Mean : 16921 Mean : 55.06 Mean : 25.27 Mean :0.4852  3rd Qu.:14.00 3rd Qu.: 0.0000 3rd Qu.: 20829 3rd Qu.: 73.60 3rd Qu.: 32.00 3rd Qu.:1.0000  Max. :90.00 Max. :86.0000 Max. :2904836 Max. :127.40 Max. :169.00 Max. :1.0000  out\_prncp total\_pymnt total\_pymnt\_inv total\_rec\_prncp total\_rec\_int total\_rec\_late\_fee  Min. : 0 Min. : 0 Min. : 0 Min. : 0 Min. : 0.0 Min. : 0.0000  1st Qu.: 0 1st Qu.: 1915 1st Qu.: 1900 1st Qu.: 1201 1st Qu.: 441.5 1st Qu.: 0.0000  Median : 6458 Median : 4895 Median : 4862 Median : 3215 Median : 1073.3 Median : 0.0000  Mean : 8359 Mean : 7559 Mean : 7521 Mean : 5758 Mean : 1754.8 Mean : 0.3967  3rd Qu.:13659 3rd Qu.:10617 3rd Qu.:10566 3rd Qu.: 8000 3rd Qu.: 2238.3 3rd Qu.: 0.0000  Max. :34145 Max. :57778 Max. :57778 Max. :35000 Max. :24205.6 Max. :358.6800  recoveries collection\_recovery\_fee last\_pymnt\_amnt collections\_12\_mths\_ex\_med  Min. : 0.00 Min. : 0.000 Min. : 0.0 Min. : 0.00000  1st Qu.: 0.00 1st Qu.: 0.000 1st Qu.: 280.2 1st Qu.: 0.00000  Median : 0.00 Median : 0.000 Median : 462.8 Median : 0.00000  Mean : 45.92 Mean : 4.881 Mean : 2164.2 Mean : 0.01438  3rd Qu.: 0.00 3rd Qu.: 0.000 3rd Qu.: 831.2 3rd Qu.: 0.00000  Max. :33520.27 Max. :7002.190 Max. :36475.6 Max. :20.00000  acc\_now\_delinq tot\_coll\_amt tot\_cur\_bal total\_rev\_hi\_lim time\_since\_first\_credit  Min. : 0.000000 Min. : -256 Min. :-126417 Min. : -28359 Min. : 184  1st Qu.: 0.000000 1st Qu.: 0 1st Qu.: 30362 1st Qu.: 14100 1st Qu.: 4110  Median : 0.000000 Median : 0 Median : 82953 Median : 24000 Median : 5419  Mean : 0.004991 Mean : 228 Mean : 137723 Mean : 31690 Mean : 5969  3rd Qu.: 0.000000 3rd Qu.: 0 3rd Qu.: 205284 3rd Qu.: 39400 3rd Qu.: 7336  Max. :14.000000 Max. :9152545 Max. :8000078 Max. :9999999 Max. :25933 | |  | | |  | | --- | |  | | |

**LOGISTIC REGRESSION MODEL FOR EXISTING CUSTOMERS**

**Divide data into test & Train:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | > set.seed(36756)  >  > inTrain <- createDataPartition(final\_loan\_data$loan\_status,p=0.70,list = FALSE)  > Training <- final\_loan\_data[inTrain,]  > Testing <- final\_loan\_data[-inTrain,]  >  > c(nrow(Training), nrow(Testing))  [1] 621166 266213  > prop.table(table(Training$loan\_status))  0 1  0.92409758 0.07590242  > prop.table(table(Testing$loan\_status))  0 1  0.92381664 0.07618336 | |  | | |  | | --- | |  | | |

**Build Logistic regression Model:**

|  |
| --- |
| > logit = glm(loan\_status ~ ., data=Training, family= binomial) |

**OVERALL SIGNIFICANCE/VALIDITY OF THE MODEL (log likelihood test):**

|  |
| --- |
| > library(lmtest)  > lrtest(logit)  Likelihood ratio test  Model 1: loan\_status ~ loan\_amnt + term + int\_rate + installment + grade +  emp\_length + home\_ownership + annual\_inc + verification\_status +  purpose + addr\_state + dti + delinq\_2yrs + inq\_last\_6mths +  mths\_since\_last\_delinq + open\_acc + pub\_rec + revol\_bal +  revol\_util + total\_acc + initial\_list\_status + out\_prncp +  total\_pymnt + total\_rec\_prncp + total\_rec\_int + total\_rec\_late\_fee +  recoveries + collection\_recovery\_fee + last\_pymnt\_amnt +  collections\_12\_mths\_ex\_med + acc\_now\_delinq + tot\_coll\_amt +  tot\_cur\_bal + total\_rev\_hi\_lim + time\_since\_first\_credit  Model 2: loan\_status ~ 1  #Df LogLik Df Chisq Pr(>Chisq)  1 104 -73256  2 1 -166874 -103 187236 < 2.2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

**MEASURE OF FIT: CALCULATE MCFadden R-Square**

|  |
| --- |
| > library(pscl)  > pR2(logit)  llh llhNull G2 McFadden r2ML r2CU  -7.325561e+04 -1.668736e+05 1.872360e+05 5.610114e-01 2.602379e-01 6.260667e-01 |

**SUMMARY OF LOGIT FUNCTION:**

> summary(logit)

Call:

glm(formula = loan\_status ~ ., family = binomial, data = Training)

Deviance Residuals:

Min 1Q Median 3Q Max

-5.7539 -0.2782 -0.1854 -0.0815 7.2934

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.703e+00 1.897e-01 -24.786 < 2e-16 \*\*\*

loan\_amnt 2.479e-04 6.817e-06 36.357 < 2e-16 \*\*\*

term 6.140e-01 4.136e-02 14.846 < 2e-16 \*\*\*

int\_rate 1.756e-01 6.467e-03 27.153 < 2e-16 \*\*\*

installment 7.785e-03 2.234e-04 34.854 < 2e-16 \*\*\*

gradeB -2.395e-01 4.070e-02 -5.884 4.00e-09 \*\*\*

gradeC -6.133e-01 5.327e-02 -11.513 < 2e-16 \*\*\*

gradeD -1.027e+00 7.054e-02 -14.557 < 2e-16 \*\*\*

gradeE -1.555e+00 8.798e-02 -17.679 < 2e-16 \*\*\*

gradeF -2.353e+00 1.120e-01 -21.017 < 2e-16 \*\*\*

gradeG -2.735e+00 1.366e-01 -20.028 < 2e-16 \*\*\*

emp\_length -5.455e-03 2.227e-03 -2.450 0.014302 \*

home\_ownershipOtherNone -7.950e-01 5.241e-01 -1.517 0.129311

home\_ownershipOWN 1.197e-02 2.762e-02 0.433 0.664805

home\_ownershipRENT 5.251e-02 2.060e-02 2.549 0.010798 \*

annual\_inc -1.524e-06 2.566e-07 -5.939 2.86e-09 \*\*\*

verification\_status -1.819e-02 1.772e-02 -1.026 0.304682

purposecredit\_card 2.149e-01 8.389e-02 2.562 0.010417 \*

purposedebt\_consolidation 3.262e-01 8.256e-02 3.952 7.76e-05 \*\*\*

purposeeducational -7.289e-01 4.088e-01 -1.783 0.074613 .

purposehome\_improvement 2.664e-01 8.812e-02 3.023 0.002503 \*\*

purposehouse 3.850e-01 1.313e-01 2.932 0.003362 \*\*

purposemajor\_purchase 1.512e-01 9.833e-02 1.537 0.124252

purposemedical 2.723e-01 1.040e-01 2.617 0.008874 \*\*

purposemoving 1.760e-01 1.120e-01 1.571 0.116083

purposeother 9.826e-02 8.739e-02 1.124 0.260810

purposerenewable\_energy 2.657e-01 2.616e-01 1.016 0.309746

purposesmall\_business 3.626e-01 1.017e-01 3.567 0.000362 \*\*\*

purposevacation 1.622e-01 1.177e-01 1.378 0.168233

purposewedding 3.440e-02 1.707e-01 0.202 0.840289

addr\_stateAL 9.801e-02 1.685e-01 0.582 0.560866

addr\_stateAR 7.315e-02 1.767e-01 0.414 0.678960

addr\_stateAZ 3.462e-02 1.642e-01 0.211 0.833003

addr\_stateCA 1.035e-01 1.576e-01 0.657 0.511433

addr\_stateCO -7.267e-02 1.659e-01 -0.438 0.661344

addr\_stateCT -8.599e-03 1.682e-01 -0.051 0.959231

addr\_stateIN 5.471e-04 1.674e-01 0.003 0.997392

addr\_stateKS -3.429e-01 1.809e-01 -1.895 0.058057 .

addr\_stateKY -1.076e-01 1.752e-01 -0.614 0.539037

addr\_stateLA 6.630e-02 1.706e-01 0.389 0.697611

addr\_stateMA 9.447e-02 1.649e-01 0.573 0.566631

addr\_stateMD 9.926e-02 1.636e-01 0.607 0.544023

addr\_stateMS -7.702e-02 1.941e-01 -0.397 0.691462

addr\_stateMT -1.930e-02 2.113e-01 -0.091 0.927244

addr\_stateNC 5.453e-02 1.627e-01 0.335 0.737518

addr\_stateND -1.345e+00 5.803e-01 -2.317 0.020482 \*

addr\_stateNE -8.166e-01 3.247e-01 -2.515 0.011892 \*

addr\_stateNH -1.549e-01 1.971e-01 -0.786 0.431895

addr\_stateNJ 7.914e-02 1.611e-01 0.491 0.623334

addr\_stateNM 5.499e-02 1.855e-01 0.296 0.766882

addr\_stateNV 3.153e-01 1.659e-01 1.900 0.057426 .

addr\_stateNY 1.335e-01 1.583e-01 0.843 0.399097

addr\_stateRI 1.181e-02 1.937e-01 0.061 0.951397

addr\_stateSC -3.498e-01 1.749e-01 -2.000 0.045488 \*

addr\_stateSD 1.574e-01 2.156e-01 0.730 0.465237

addr\_stateTN 5.361e-02 1.670e-01 0.321 0.748237

addr\_stateTX -4.166e-02 1.588e-01 -0.262 0.793097

addr\_stateUT 1.254e-02 1.808e-01 0.069 0.944723

addr\_stateVA 1.120e-01 1.622e-01 0.690 0.489960

addr\_stateVT -4.253e-01 2.484e-01 -1.712 0.086830 .

addr\_stateWA -8.469e-02 1.650e-01 -0.513 0.607839

addr\_stateWI -1.151e-01 1.708e-01 -0.674 0.500300

addr\_stateWV -2.190e-02 1.899e-01 -0.115 0.908174

addr\_stateWY -2.766e-01 2.367e-01 -1.168 0.242706

dti 7.154e-03 1.038e-03 6.892 5.50e-12 \*\*\*

delinq\_2yrs 2.585e-03 9.488e-03 0.272 0.785256

inq\_last\_6mths 7.815e-02 7.260e-03 10.764 < 2e-16 \*\*\*

mths\_since\_last\_delinq -4.153e-03 5.809e-04 -7.150 8.67e-13 \*\*\*

open\_acc -5.889e-04 2.121e-03 -0.278 0.781229

pub\_rec -3.716e-02 1.385e-02 -2.684 0.007274 \*\*

revol\_bal -1.790e-06 1.071e-06 -1.671 0.094685 .

revol\_util 9.642e-04 4.402e-04 2.190 0.028507 \*

total\_acc 7.701e-03 9.438e-04 8.159 3.37e-16 \*\*\*

initial\_list\_status -1.416e-01 1.613e-02 -8.779 < 2e-16 \*\*\*

out\_prncp -5.414e-04 3.473e-06 -155.894 < 2e-16 \*\*\*

total\_pymnt 1.674e+01 9.779e+00 1.712 0.086903 .

total\_rec\_prncp -1.674e+01 9.779e+00 -1.712 0.086891 .

total\_rec\_int -1.674e+01 9.779e+00 -1.712 0.086908 .

total\_rec\_late\_fee -1.668e+01 9.779e+00 -1.705 0.088160 .

recoveries 1.220e+01 3.003e+01 0.406 0.684625

collection\_recovery\_fee -5.491e+01 5.449e+01 -1.008 0.313588

last\_pymnt\_amnt -5.123e-04 1.327e-05 -38.607 < 2e-16 \*\*\*

collections\_12\_mths\_ex\_med -6.615e-02 5.542e-02 -1.193 0.232688

acc\_now\_delinq -1.779e-01 9.549e-02 -1.863 0.062402 .

tot\_coll\_amt -9.631e-06 5.317e-06 -1.811 0.070086 .

tot\_cur\_bal -3.476e-07 8.758e-08 -3.970 7.20e-05 \*\*\*

total\_rev\_hi\_lim -1.605e-06 7.719e-07 -2.079 0.037650 \*

time\_since\_first\_credit -1.610e-05 3.220e-06 -5.000 5.74e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 333747 on 621165 degrees of freedom

Residual deviance: 146511 on 621062 degrees of freedom

AIC: 146719

Number of Fisher Scoring iterations: 25

**CALCULATING ODDS RATIOS:**

> exp(coef(logit))

(Intercept) loan\_amnt term

9.068472e-03 1.000248e+00 1.847803e+00

int\_rate installment gradeB

1.191953e+00 1.007816e+00 7.870477e-01

gradeC gradeD gradeE

5.415524e-01 3.581114e-01 2.110973e-01

gradeF gradeG emp\_length

9.504733e-02 6.489788e-02 9.945601e-01

home\_ownershipOtherNone home\_ownershipOWN home\_ownershipRENT

4.515748e-01 1.012041e+00 1.053912e+00

annual\_inc verification\_status purposecredit\_card

9.999985e-01 9.819739e-01 1.239734e+00

purposedebt\_consolidation purposeeducational purposehome\_improvement

1.385739e+00 4.824618e-01 1.305250e+00

purposehouse purposemajor\_purchase purposemedical

1.469675e+00 1.163179e+00 1.312940e+00

purposemoving purposeother purposerenewable\_energy

1.192417e+00 1.103254e+00 1.304360e+00

purposesmall\_business purposevacation purposewedding

1.437104e+00 1.176119e+00 1.035003e+00

addr\_stateAL addr\_stateAR addr\_stateAZ

1.102969e+00 1.075893e+00 1.035221e+00

addr\_stateCA addr\_stateCO addr\_stateCT

1.109022e+00 9.299049e-01 9.914382e-01

addr\_stateDC addr\_stateDE addr\_stateFL

addr\_stateID addr\_stateIL addr\_stateIN

4.282338e+00 8.153798e-01 1.000547e+00

addr\_stateKS addr\_stateKY addr\_stateLA

7.097302e-01 8.979719e-01 1.068552e+00

addr\_stateMA addr\_stateMD addr\_stateME

1.099074e+00 1.104354e+00 2.348566e-11

addr\_stateMI addr\_stateMN addr\_stateMO

9.699906e-01 1.022646e+00 9.923074e-01

addr\_stateMS addr\_stateMT addr\_stateNC

9.258739e-01 9.808891e-01 1.056043e+00

addr\_stateND addr\_stateNE addr\_stateNH

2.605702e-01 4.419103e-01 8.564871e-01

addr\_stateNJ addr\_stateNM addr\_stateNV

1.082358e+00 1.056532e+00 1.370685e+00

addr\_stateNY addr\_stateOH addr\_stateOK

1.142793e+00 9.485976e-01 9.947824e-01

addr\_stateOR addr\_statePA addr\_stateRI

1.084245e+00 1.087739e+00 1.011879e+00

addr\_stateSC addr\_stateSD addr\_stateTN

7.048299e-01 1.170493e+00 1.055069e+00

addr\_stateTX addr\_stateUT addr\_stateVA

9.591948e-01 1.012614e+00 1.118526e+00

addr\_stateVT addr\_stateWA addr\_stateWI

6.535593e-01 9.188006e-01 8.912713e-01

addr\_stateWV addr\_stateWY dti

9.783346e-01 7.583742e-01 1.007180e+00

delinq\_2yrs inq\_last\_6mths mths\_since\_last\_delinq

1.002589e+00 1.081279e+00 9.958551e-01

open\_acc pub\_rec revol\_bal

9.994113e-01 9.635190e-01 9.999982e-01

revol\_util total\_acc initial\_list\_status

1.000965e+00 1.007730e+00 8.679579e-01

out\_prncp total\_pymnt total\_rec\_prncp

9.994587e-01 1.865897e+07 5.356007e-08

total\_rec\_int total\_rec\_late\_fee recoveries

5.360736e-08 5.726819e-08 1.983156e+05

collection\_recovery\_fee last\_pymnt\_amnt collections\_12\_mths\_ex\_med

1.422774e-24 9.994878e-01 9.359936e-01

acc\_now\_delinq tot\_coll\_amt tot\_cur\_bal

8.369925e-01 9.999904e-01 9.999997e-01

total\_rev\_hi\_lim time\_since\_first\_credit

9.999984e-01 9.999839e-01

**CALCULATING PROBABLITIES:**

> exp(coef(logit))/(1+exp(coef(logit)))

(Intercept) loan\_amnt term

8.986974e-03 5.000620e-01 6.488521e-01

int\_rate installment gradeB

5.437859e-01 5.019463e-01 4.404178e-01

gradeC gradeD gradeE

3.513033e-01 2.636834e-01 1.743025e-01

gradeF gradeG emp\_length

8.679747e-02 6.094282e-02 4.986363e-01

home\_ownershipOtherNone home\_ownershipOWN home\_ownershipRENT

3.110930e-01 5.029922e-01 5.131243e-01

annual\_inc verification\_status purposecredit\_card

4.999996e-01 4.954525e-01 5.535183e-01

purposedebt\_consolidation purposeeducational purposehome\_improvement

5.808426e-01 3.254464e-01 5.662075e-01

purposehouse purposemajor\_purchase purposemedical

5.950883e-01 5.377175e-01 5.676498e-01

purposemoving purposeother purposerenewable\_energy

5.438823e-01 5.245463e-01 5.660400e-01

purposesmall\_business purposevacation purposewedding

5.896769e-01 5.404662e-01 5.086002e-01

addr\_stateAL addr\_stateAR addr\_stateAZ

5.244818e-01 5.182795e-01 5.086530e-01

addr\_stateCA addr\_stateCO addr\_stateCT

5.258465e-01 4.818398e-01 4.978503e-01

addr\_stateDC addr\_stateDE addr\_stateFL

4.170979e-01 5.082800e-01 5.249166e-01

addr\_stateGA addr\_stateHI addr\_stateIA

4.952079e-01 5.457376e-01 7.555066e-01

addr\_stateID addr\_stateIL addr\_stateIN

8.106899e-01 4.491511e-01 5.001368e-01

addr\_stateMS addr\_stateMT addr\_stateNC

4.807552e-01 4.951762e-01 5.136288e-01

addr\_stateND addr\_stateNE addr\_stateNH

2.067082e-01 3.064756e-01 4.613483e-01

addr\_stateNJ addr\_stateNM addr\_stateNV

5.197752e-01 5.137445e-01 5.781809e-01

addr\_stateNY addr\_stateOH addr\_stateOK

5.333193e-01 4.868104e-01 4.986922e-01

addr\_stateOR addr\_statePA addr\_stateRI

5.202100e-01 5.210130e-01 5.029523e-01

addr\_stateSC addr\_stateSD addr\_stateTN

4.134312e-01 5.392752e-01 5.133984e-01

addr\_stateTX addr\_stateUT addr\_stateVA

4.895862e-01 5.031338e-01 5.279736e-01

addr\_stateVT addr\_stateWA addr\_stateWI

3.952439e-01 4.788411e-01 4.712551e-01

addr\_stateWV addr\_stateWY dti

4.945243e-01 4.312928e-01 5.017886e-01

delinq\_2yrs inq\_last\_6mths mths\_since\_last\_delinq

5.006463e-01 5.195263e-01 4.989616e-01

open\_acc pub\_rec revol\_bal

4.998528e-01 4.907103e-01 4.999996e-01

revol\_util total\_acc initial\_list\_status

5.002411e-01 5.019252e-01 4.646560e-01

out\_prncp total\_pymnt total\_rec\_prncp

4.998646e-01 9.999999e-01 5.356006e-08

total\_rec\_int total\_rec\_late\_fee recoveries

5.360736e-08 5.726819e-08 9.999950e-01

collection\_recovery\_fee last\_pymnt\_amnt collections\_12\_mths\_ex\_med

1.422774e-24 4.998719e-01 4.834694e-01

acc\_now\_delinq tot\_coll\_amt tot\_cur\_bal

4.556320e-01 4.999976e-01 4.999999e-01

total\_rev\_hi\_lim time\_since\_first\_credit

4.999996e-01 4.999960e-01

**PREDICTION ON TRAIN DATA SET:**

> predicted\_status <- plogis(predict(logit, Training\_p))

>

> # Decide on optimal prediction probability cut-off for the model

> cutoff=optimalCutoff(Training\_p$loan\_status,predicted\_status)

> cutoff

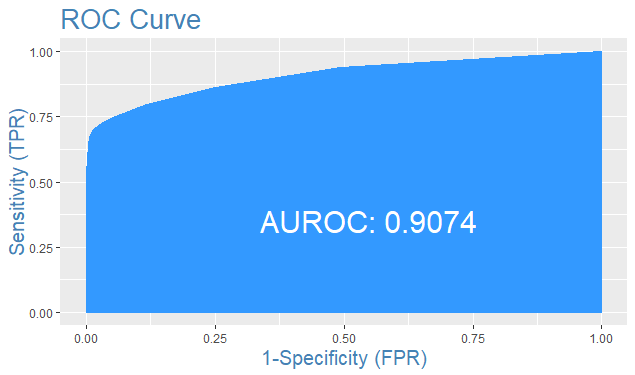
[1] 0.25

>

> # AUC-ROC for training data

> plotROC(Training\_p$loan\_status, predicted\_status)

>



> misClassError(Training\_p$loan\_status, predicted\_status, threshold = cutoff)

[1] 0.0308

>

> confusion\_Matrix<-confusionMatrix(Training\_p$loan\_status,predicted\_status,threshold = cutoff)

> confusion\_Matrix

0 1

0 570641 15782

1 3377 31366

>

> # 1's as 1's

> sensitivity(Training\_p$loan\_status,predicted\_status,threshold = cutoff)

[1] 0.6652668

>

> # 0's as 0's

> specificity(Training\_p$loan\_status,predicted\_status,threshold = cutoff)

[1] 0.9941169

>

**PREDICTION ON TEST DATA SET:**

> predict\_on\_test<- plogis(predict(logit, Testing\_p))

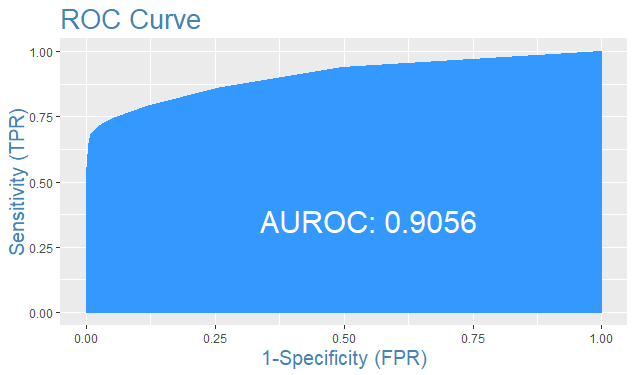
>

> cutoff1=optimalCutoff(Testing\_p$loan\_status,predict\_on\_test)

> cutoff1

[1] 0.25

> plotROC(Testing\_p$loan\_status, predict\_on\_test)



> # horizontal is actual ; vertical is predicted

> confusionMatrix(Testing\_p$loan\_status,predict\_on\_test,threshold = cutoff1)

0 1

0 244398 6806

1 1534 13475

>

> sensitivity(Testing\_p$loan\_status,predict\_on\_test,threshold = cutoff1)

[1] 0.664415

>

> specificity(Testing\_p$loan\_status,predict\_on\_test,threshold = cutoff1)

[1] 0.9937625

>

> misClassError(Testing\_p$loan\_status, predict\_on\_test, threshold = cutoff1)

[1] 0.0313

**LOGISTIC REGRESSION MODEL FOR NEW CUSTOMERS**

**VARIABLES USED FOR THE MODEL:**

> loan\_data <- loan\_data[, c(1,2,5,6,7,8,10,11,12,15,18)]

> str(loan\_data)

'data.frame': 887379 obs. of 11 variables:

$ loan\_amnt : int 5000 2500 2400 10000 3000 5000 7000 3000 5600 5375 ...

$ term : int 0 1 0 0 1 0 1 0 1 1 ...

$ grade : Factor w/ 7 levels "A","B","C","D",..: 2 3 3 3 2 1 3 5 6 2 ...

$ emp\_length : int 10 0 10 10 1 3 8 9 4 0 ...

$ home\_ownership: Factor w/ 4 levels "MORTGAGE","OtherNone",..: 4 4 4 4 4 4 4 4 3 4 ...

$ annual\_inc : num 24000 30000 12252 49200 80000 ...

$ loan\_status : int 0 1 0 0 0 0 0 0 1 1 ...

$ purpose : Factor w/ 14 levels "car","credit\_card",..: 2 1 12 10 10 14 3 1 12 10 ...

$ addr\_state : Factor w/ 51 levels "AK","AL","AR",..: 4 11 15 5 38 4 28 5 5 44 ...

$ inq\_last\_6mths: int 1 5 2 1 0 3 1 2 2 0 ...

$ pub\_rec : int 0 0 0 0 0 0 0 0 0 0 ...

**Divide data into test & Train:**

> inTrain <- createDataPartition(final\_loan\_data$loan\_status,p=0.70,list = FALSE)

> Training <- final\_loan\_data[inTrain,]

> Testing <- final\_loan\_data[-inTrain,]

>

> c(nrow(Training), nrow(Testing))

[1] 621166 266213

> prop.table(table(Training$loan\_status))

0 1

0.92407827 0.07592173

> prop.table(table(Testing$loan\_status))

0 1

0.92386172 0.07613828

**Build Logistic regression Model:**

> logit = glm(loan\_status ~ ., data=Training, family= binomial)

**OVERALL SIGNIFICANCE/VALIDITY OF THE MODEL (log likelihood test):**

> lrtest(logit)

Likelihood ratio test

Model 1: loan\_status ~ loan\_amnt + term + grade + emp\_length + home\_ownership +

annual\_inc + purpose + addr\_state + inq\_last\_6mths + pub\_rec

Model 2: loan\_status ~ 1

#Df LogLik Df Chisq Pr(>Chisq)

1 79 -157886

2 1 -166904 -78 18036 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**MEASURE OF FIT: CALCULATE Mc-Fadden R-Square**

> library(pscl)

> pR2(logit)

llh llhNull G2 McFadden r2ML r2CU

-1.578858e+05 -1.669036e+05 1.803558e+04 5.402994e-02 2.861758e-02 6.883732e-02

**SUMMARY OF LOGIT FUCNTION**

> summary(logit)

Call:

glm(formula = loan\_status ~ ., family = binomial, data = Training)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.4765 -0.4378 -0.3531 -0.2629 8.4904

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.606e+00 1.120e-01 -32.197 < 2e-16 \*\*\*

loan\_amnt 8.036e-06 7.565e-07 10.622 < 2e-16 \*\*\*

term -1.466e-01 1.262e-02 -11.616 < 2e-16 \*\*\*

gradeB 7.465e-01 2.278e-02 32.765 < 2e-16 \*\*\*

gradeC 1.104e+00 2.256e-02 48.950 < 2e-16 \*\*\*

gradeD 1.477e+00 2.353e-02 62.746 < 2e-16 \*\*\*

gradeE 1.709e+00 2.586e-02 66.080 < 2e-16 \*\*\*

gradeF 2.052e+00 3.087e-02 66.456 < 2e-16 \*\*\*

gradeG 2.227e+00 4.634e-02 48.052 < 2e-16 \*\*\*

emp\_length -9.197e-03 1.403e-03 -6.554 5.61e-11 \*\*\*

home\_ownershipOtherNone 7.810e-01 2.124e-01 3.677 0.000236 \*\*\*

home\_ownershipOWN -2.122e-02 1.772e-02 -1.198 0.230985

home\_ownershipRENT 1.389e-01 1.141e-02 12.177 < 2e-16 \*\*\*

annual\_inc -4.071e-06 1.558e-07 -26.129 < 2e-16 \*\*\*

purposecredit\_card -7.445e-02 5.317e-02 -1.400 0.161410

purposedebt\_consolidation 5.393e-02 5.229e-02 1.031 0.302371

purposeeducational 9.040e-01 1.583e-01 5.709 1.13e-08 \*\*\*

purposehome\_improvement 4.494e-02 5.616e-02 0.800 0.423545

purposehouse 1.719e-01 8.103e-02 2.121 0.033928 \*

purposemajor\_purchase 3.443e-02 6.291e-02 0.547 0.584113

purposemedical 1.441e-01 6.854e-02 2.102 0.035545 \*

purposemoving 1.134e-01 7.376e-02 1.537 0.124295

purposeother 1.194e-01 5.548e-02 2.153 0.031344 \*

purposerenewable\_energy 3.359e-01 1.627e-01 2.065 0.038943 \*

purposesmall\_business 6.761e-01 6.107e-02 11.071 < 2e-16 \*\*\*

purposevacation 2.007e-02 8.170e-02 0.246 0.805940

purposewedding 3.301e-01 9.413e-02 3.507 0.000453 \*\*\*

addr\_stateAL 2.259e-01 1.046e-01 2.160 0.030757 \*

addr\_stateAR 1.053e-01 1.112e-01 0.948 0.343345

addr\_stateCO -1.515e-01 1.033e-01 -1.467 0.142317

addr\_stateCT 7.659e-03 1.050e-01 0.073 0.941840

addr\_stateDC -5.145e-01 1.559e-01 -3.299 0.000970 \*\*\*

addr\_stateDE 4.151e-02 1.327e-01 0.313 0.754413

addr\_stateFL 1.774e-01 9.801e-02 1.810 0.070273 .

addr\_stateGA 2.547e-02 1.004e-01 0.254 0.799668

addr\_stateHI 2.401e-01 1.139e-01 2.108 0.035005 \*

addr\_stateIA 8.218e-01 8.458e-01 0.972 0.331225

addr\_stateLA 1.729e-01 1.056e-01 1.638 0.101388

addr\_stateMA 1.177e-01 1.017e-01 1.157 0.247116

addr\_stateMD 1.696e-01 1.012e-01 1.675 0.093836 .

addr\_stateME -3.316e+00 9.957e-01 -3.330 0.000868 \*\*\*

addr\_stateMI 9.307e-02 1.010e-01 0.922 0.356750

addr\_stateMN 6.996e-02 1.033e-01 0.677 0.498200

addr\_stateMO 8.630e-02 1.037e-01 0.832 0.405491

addr\_stateMS -3.539e-01 1.299e-01 -2.724 0.006452 \*\*

addr\_stateMT -9.877e-02 1.381e-01 -0.715 0.474615

addr\_stateNC 1.369e-01 1.006e-01 1.361 0.173378

addr\_stateND -1.730e+00 5.134e-01 -3.369 0.000755 \*\*\*

addr\_stateNE -1.335e+00 2.643e-01 -5.051 4.39e-07 \*\*\*

addr\_stateNH -1.843e-01 1.245e-01 -1.480 0.138865

addr\_stateNJ 1.798e-01 9.945e-02 1.808 0.070612 .

addr\_stateNM 2.109e-01 1.147e-01 1.838 0.066055 .

addr\_stateNV 2.978e-01 1.032e-01 2.884 0.003924 \*\*

addr\_stateNY 2.004e-01 9.772e-02 2.051 0.040299 \*

addr\_stateOH 2.120e-02 1.002e-01 0.212 0.832428

addr\_stateOK 1.620e-01 1.081e-01 1.498 0.134121

addr\_stateOR 5.500e-02 1.064e-01 0.517 0.605184

addr\_stateSD 5.959e-02 1.440e-01 0.414 0.678982

addr\_stateTN 6.068e-02 1.046e-01 0.580 0.561674

addr\_stateTX -3.421e-02 9.808e-02 -0.349 0.727265

addr\_stateUT 1.885e-01 1.113e-01 1.694 0.090249 .

addr\_stateVA 1.856e-01 1.003e-01 1.851 0.064150 .

addr\_stateVT -2.818e-01 1.601e-01 -1.760 0.078455 .

addr\_stateWA 2.174e-02 1.021e-01 0.213 0.831298

addr\_stateWI -3.137e-03 1.060e-01 -0.030 0.976401

addr\_stateWV -5.870e-02 1.222e-01 -0.481 0.630815

addr\_stateWY -2.132e-01 1.516e-01 -1.406 0.159657

inq\_last\_6mths 1.726e-01 4.330e-03 39.865 < 2e-16 \*\*\*

pub\_rec -1.989e-01 1.045e-02 -19.028 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 333807 on 621165 degrees of freedom

Residual deviance: 315772 on 621087 degrees of freedom

AIC: 315930

Number of Fisher Scoring iterations: 7

**CALCULATING ODDS RATIOS**

> exp(coef(logit))

(Intercept) loan\_amnt term

0.02716192 1.00000804 0.86365660

gradeB gradeC gradeD

2.10968211 3.01672257 4.37855725

gradeE gradeF gradeG

5.52406656 7.78180681 9.27155090

emp\_length home\_ownershipOtherNone home\_ownershipOWN

0.99084489 2.18359240 0.97899973

home\_ownershipRENT annual\_inc purposecredit\_card

1.14904427 0.99999593 0.92825070

purposedebt\_consolidation purposeeducational purposehome\_improvement

1.05540870 2.46953372 1.04596683

purposehouse purposemajor\_purchase purposemedical

1.18751914 1.03503362 1.15497438

purposemoving purposeother purposerenewable\_energy

1.12005239 1.12685897 1.39925542

purposesmall\_business purposevacation purposewedding

1.96613259 1.02027380 1.39113361

addr\_stateAL addr\_stateAR addr\_stateAZ

1.25351102 1.11107213 1.14767078

addr\_stateCA addr\_stateCO addr\_stateCT

1.16309418 0.85940310 1.00768826

addr\_stateDC addr\_stateDE addr\_stateFL

0.59781572 1.04238188 1.19412527

addr\_stateGA addr\_stateHI addr\_stateIA

1.02579674 1.27134228 2.27467977

addr\_stateID addr\_stateIL addr\_stateIN

1.55395565 0.87620336 1.00638538

addr\_stateKS addr\_stateKY addr\_stateLA

0.89111288 1.01830487 1.18878655

addr\_stateMA addr\_stateMD addr\_stateME

1.12490403 1.18485132 0.03631175

addr\_stateMI addr\_stateMN addr\_stateMO

1.09754314 1.07246432 1.09012935

addr\_stateMS addr\_stateMT addr\_stateNC

0.70194579 0.90595314 1.14673883

1.19698473 1.23473711 1.34684256

addr\_stateNY addr\_stateOH addr\_stateOK

1.22188338 1.02143050 1.17585195

addr\_stateOR addr\_statePA addr\_stateRI

1.05654258 1.02729268 1.13303736

addr\_stateSC addr\_stateSD addr\_stateTN

0.89392883 1.06140131 1.06255975

addr\_stateTX addr\_stateUT addr\_stateVA

0.96637055 1.20747056 1.20396138

addr\_stateVT addr\_stateWA addr\_stateWI

0.75441013 1.02198209 0.99686826

addr\_stateWV addr\_stateWY inq\_last\_6mths

0.94298614 0.80800281 1.18841928

pub\_rec

0.81960933

**CALCULATING PROBABILITES**

> exp(coef(logit))/(1+exp(coef(logit)))

(Intercept) loan\_amnt term

0.02644366 0.50000201 0.46342046

gradeB gradeC gradeD

0.67842372 0.75104081 0.81407654

gradeE gradeF gradeG

0.84672137 0.88612822 0.90264372

emp\_length home\_ownershipOtherNone home\_ownershipOWN

0.49770070 0.68588944 0.49469422

home\_ownershipRENT annual\_inc purposecredit\_card

0.53467687 0.49999898 0.48139524

purposedebt\_consolidation purposeeducational purposehome\_improvement

0.51347875 0.71177683 0.51123352

purposehouse purposemajor\_purchase purposemedical

0.54286114 0.50860763 0.53595736

purposemoving purposeother purposerenewable\_energy

0.52831354 0.52982308 0.58320403

purposesmall\_business purposevacation purposewedding

0.66286065 0.50501759 0.58178832

addr\_stateAL addr\_stateAR addr\_stateAZ

0.55624801 0.52630704 0.53437929

addr\_stateCA addr\_stateCO addr\_stateCT

0.53769928 0.46219300 0.50191470

addr\_stateDC addr\_stateDE addr\_stateFL

0.37414560 0.51037560 0.54423751

addr\_stateGA addr\_stateHI addr\_stateIA

0.50636706 0.55973170 0.69462663

addr\_stateID addr\_stateIL addr\_stateIN

0.60845052 0.46700874 0.50159127

addr\_stateKS addr\_stateKY addr\_stateLA

0.47121084 0.50453471 0.54312585

addr\_stateMA addr\_stateMD addr\_stateME

0.54483070 0.55252007 0.57389558

addr\_stateNY addr\_stateOH addr\_stateOK

0.54993137 0.50530082 0.54040991

addr\_stateOR addr\_statePA addr\_stateRI

0.51374700 0.50673131 0.53118496

addr\_stateSC addr\_stateSD addr\_stateTN

0.47199706 0.51489310 0.51516556

addr\_stateTX addr\_stateUT addr\_stateVA

0.49144885 0.54699283 0.54627154

addr\_stateVT addr\_stateWA addr\_stateWI

0.43000785 0.50543578 0.49921584

addr\_stateWV addr\_stateWY inq\_last\_6mths

0.48532829 0.44690352 0.54304917

pub\_rec

0.45043148

**PREDICTION ON TRAIN DATA**

> predicted\_status <- plogis(predict(logit, Training\_p))

> # Decide on optimal prediction probability cut-off for the model

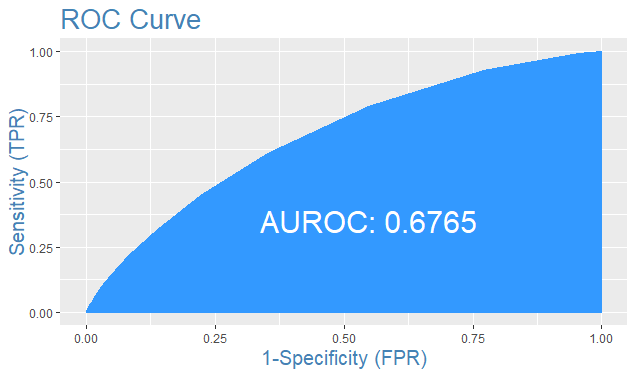
> cutoff=optimalCutoff(Training\_p$loan\_status,predicted\_status)

> cutoff

[1] 0.4834187

> plotROC(Training\_p$loan\_status, predicted\_status)

>



> misClassError(Training\_p$loan\_status, predicted\_status, threshold = cutoff)

[1] 0.0759

> confusionMatrix(Training\_p$loan\_status,predicted\_status,threshold = cutoff)

0 1

0 573946 47115

1 60 45

> sensitivity(Training\_p$loan\_status,predicted\_status,threshold = cutoff)

[1] 0.0009541985

> specificity(Training\_p$loan\_status,predicted\_status,threshold = cutoff)

[1] 0.9998955

**PREDICTION ON TEST DATA**

>

> Testing\_p <- Testing

> library(InformationValue)

> predict\_on\_test<- plogis(predict(logit, Testing\_p))

>

> cutoff1=optimalCutoff(Testing\_p$loan\_status,predict\_on\_test)

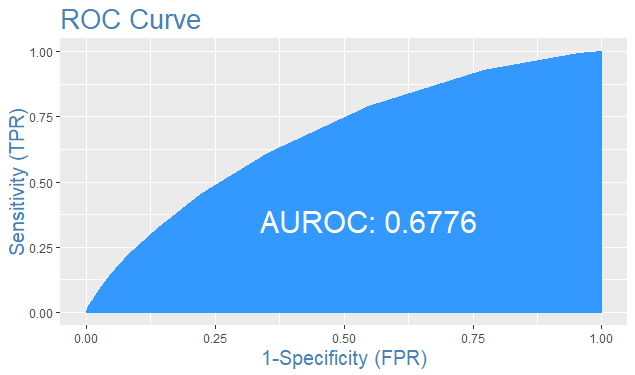
> cutoff1

[1] 0.4569768

> #

> plotROC(Testing\_p$loan\_status, predict\_on\_test)

>



> # horizontal is actual ; vertical is predicted

> confusionMatrix(Testing\_p$loan\_status,predict\_on\_test,threshold = cutoff1)

0 1

0 245920 20247

1 24 22

>

> (245920+22)/nrow(Testing\_p)

[1] 0.9238542

> (151605+158184)/nrow(Testing\_p)

[1] 1.163688

> ( 72000+82108)/nrow(Testing\_p)

[1] 0.5788898

>

> # 1's as 1's

> sensitivity(Testing\_p$loan\_status,predict\_on\_test,threshold = cutoff1)

[1] 0.001085401

>

> # 0's as 0's

> specificity(Testing\_p$loan\_status,predict\_on\_test,threshold = cutoff1)

[1] 0.9999024

**SVM PREDICTION MODEL FOR NEW CUSTOMERS**

**Variables used for the model:**

> str(loan\_data)

'data.frame': 887379 obs. of 11 variables:

$ loan\_amnt : int 5000 2500 2400 10000 3000 5000 7000 3000 5600 5375 ...

$ term : int 0 1 0 0 1 0 1 0 1 1 ...

$ grade : Factor w/ 7 levels "A","B","C","D",..: 2 3 3 3 2 1 3 5 6 2 ...

$ emp\_length : int 10 0 10 10 1 3 8 9 4 0 ...

$ home\_ownership: Factor w/ 4 levels "MORTGAGE","OtherNone",..: 4 4 4 4 4 4 4 4 3 4 ...

$ annual\_inc : num 24000 30000 12252 49200 80000 ...

$ loan\_status : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 2 2 ...

$ purpose : Factor w/ 14 levels "car","credit\_card",..: 2 1 12 10 10 14 3 1 12 10 ...

$ addr\_state : Factor w/ 51 levels "AK","AL","AR",..: 4 11 15 5 38 4 28 5 5 44 ...

$ inq\_last\_6mths: int 1 5 2 1 0 3 1 2 2 0 ...

$ pub\_rec : int 0 0 0 0 0 0 0 0 0 0 ...

**Divide data into test & Train:**

> set.seed(786449)

>

> inTrain <- createDataPartition(strat\_loan\_data$loan\_status,p=0.70,list = FALSE)

> Training <- strat\_loan\_data[inTrain,]

> Testing <- strat\_loan\_data[-inTrain,]

>

> c(nrow(Training), nrow(Testing))

[1] 155292 66553

> prop.table(table(Training$loan\_status))

0 1

0.92401412 0.07598588

> prop.table(table(Testing$loan\_status))

0 1

0.92401545 0.07598455

**Build SVM Model**

library(e1071)

model\_svm <- svm(loan\_status~. , data=Training, method="C-classification",gamma=4, cost=5)

summary(model\_svm)

**SUMMARY OF LOGIT FUCNTION**

> summary(model\_svm)

Call:

svm(formula = loan\_status ~ ., data = Training, method = "C-classification", gamma = 4, cost = 5)

Parameters:

SVM-Type: C-classification

SVM-Kernel: radial

cost: 5

gamma: 4

Number of Support Vectors: 120174

( 108412 11762 )

Number of Classes: 2

Levels:

0 1

**PREDICTION ON TRAIN DATA**

> pred <- predict(model\_svm, Training)

> confusionMatrix(Training$loan\_status,pred)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 143428 64

1 2147 9653

Accuracy : 0.9858

95% CI : (0.9852, 0.9863)

No Information Rate : 0.9374

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.8897

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9853

Specificity : 0.9934

Pos Pred Value : 0.9996

Neg Pred Value : 0.8181

Prevalence : 0.9374

Detection Rate : 0.9236

Detection Prevalence : 0.9240

Balanced Accuracy : 0.9893

'Positive' Class : 0

**PREDICTION ON TEST DATA**

> prediction\_on\_test <- predict(model\_svm, Testing)

> confusionMatrix(Testing$loan\_status,prediction\_on\_test)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 60102 1394

1 4892 165

Accuracy : 0.9055

95% CI : (0.9033, 0.9078)

No Information Rate : 0.9766

P-Value [Acc > NIR] : 1

Kappa : 0.0146

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.92473

Specificity : 0.10022

Pos Pred Value : 0.97733

Neg Pred Value : 0.03263

Prevalence : 0.97658

Detection Rate : 0.90307

Detection Prevalence : 0.92402

Balanced Accuracy : 0.51528

'Positive' Class : 0

**plot(model\_svm,data=Training,loan\_amnt ~ annual\_inc )**



**CLUSTERING FOR EXISTING CUSTOMERS**

**# IMPORTING THE DATA**

setwd("C:\\Users\\devlina\\Downloads\\capstone")

> loan\_data<-read.csv("Post Imputation finalV2.csv",header=TRUE)

> library(splitstackshape)

> loan=loan\_data

**# DATA DESCRIPTION**

> describe(loan\_data)

vars n mean sd min max range se

X 1 887379 443690.00 256164.40 1.00 887379.00 887378.00 271.93

loan\_amnt 2 887379 14755.26 8435.46 500.00 35000.00 34500.00 8.95

term 3 887379 0.30 0.46 0.00 1.00 1.00 0.00

int\_rate 4 887379 13.25 4.38 5.32 28.99 23.67 0.00

installment 5 887379 436.72 244.19 15.67 1445.46 1429.79 0.26

emp\_length 6 887379 6.01 3.58 0.00 11.00 11.00 0.00

home\_ownership 7 887379 1.90 0.94 1.00 4.00 3.00 0.00

annual\_inc 8 887379 75027.58 64698.15 0.00 9500000.00 9500000.00 68.68

verification\_status 9 887379 0.70 0.46 0.00 1.00 1.00 0.00

loan\_status 10 887379 1.15 0.52 1.00 3.00 2.00 0.00

purpose 11 887379 3.57 2.28 1.00 14.00 13.00 0.00

dti 12 887379 18.13 8.30 0.00 41.94 41.94 0.01

inq\_last\_6mths 13 887379 0.69 1.00 0.00 33.00 33.00 0.00

mths\_since\_last\_delinq 14 887379 34.09 15.37 0.00 188.00 188.00 0.02

open\_acc 15 887379 11.55 5.32 0.00 90.00 90.00 0.01

pub\_rec 16 887379 0.20 0.58 0.00 86.00 86.00 0.00

revol\_bal 17 887379 16920.79 22426.79 0.00 2904836.00 2904836.00 23.81

revol\_util 18 887379 55.06 23.80 0.00 127.40 127.40 0.03

total\_acc 19 887379 25.27 11.84 1.00 169.00 168.00 0.01

initial\_list\_status 20 887379 0.49 0.50 0.00 1.00 1.00 0.00

out\_prncp 21 887379 8359.26 8387.67 0.00 34145.16 34145.16 8.90

total\_pymnt 22 887379 7558.83 7871.24 0.00 57777.58 57777.58 8.36

total\_rec\_prncp 23 887379 5757.71 6625.44 0.00 35000.03 35000.03 7.03

total\_rec\_int 24 887379 1754.80 2095.36 0.00 24205.62 24205.62 2.22

total\_rec\_late\_fee 25 887379 0.40 4.09 0.00 358.68 358.68 0.00

recoveries 26 887379 45.92 409.69 0.00 33520.27 33520.27 0.43

collection\_recovery\_fee 27 887379 4.88 63.13 0.00 7002.19 7002.19 0.07

last\_pymnt\_amnt 28 887379 2164.15 4794.78 0.00 36475.59 36475.59 5.09

collections\_12\_mths\_ex\_med 29 887379 0.01 0.13 0.00 20.00 20.00 0.00

acc\_now\_delinq 30 887379 0.00 0.08 0.00 14.00 14.00 0.00

tot\_coll\_amt 31 887379 227.87 9894.73 -256.00 9152545.00 9152801.00 10.50

tot\_cur\_bal 32 887379 137723.34 149659.69 -126417.00 8000078.00 8126495.00 158.87

total\_rev\_hi\_lim 33 887379 31690.13 36262.35 -28359.00 9999999.00 10028358.00 38.49

time\_since\_first\_credit 34 887379 5969.46 2726.19 184.00 25933.00 25749.00 2.89

grade 35 887379 NaN NA Inf -Inf -Inf NA

addr\_state 36 887379 NaN NA Inf -Inf -Inf NA

delinq\_2yrs 37 887350 0.31 0.86 0.00 39.00 39.00 0.00

Warning messages:

1: In FUN(newX[, i], ...) : no non-missing arguments to min; returning Inf

2: In FUN(newX[, i], ...) : no non-missing arguments to min; returning Inf

3: In FUN(newX[, i], ...) :

no non-missing arguments to max; returning -Inf

4: In FUN(newX[, i], ...) :

no non-missing arguments to max; returning -Inf

>

>

>

#CONVERTING GRADE INTO ORDINAL NUMERICAL VARIABLES DOR THE PURPOSE OF CLUSTERING

> loan$grade<- as.character(loan$grade)

> loan$grade[loan$grade=="A"] <- "7" #Convert to 0

> loan$grade[loan$grade=="B"] <- "6" #Convert to 1

> loan$grade[loan$grade=="C"] <- "5"

> loan$grade[loan$grade=="D"] <- "4"

> loan$grade[loan$grade=="E"] <- "3"

> loan$grade[loan$grade=="F"] <- "2"

> loan$grade[loan$grade=="G"] <- "1"

> loan$grade<- as.numeric(loan$grade)

>

**# CHOOSING ONLY THE SIGNIFICANT VARIABLES WE GOT FROM LOGISTIC MODEL**

> loan\_new= loan[,c(2,3,4,5,6,8,12,14,16,18,19,20,21,28,32,34,22,13,35)]

**# SCALING OF THE DATA**

> loan\_num <- scale(loan\_new)

> describe(loan\_num)

vars n mean sd min max range se

loan\_amnt 1 887379 0 1 -1.69 2.40 4.09 0

term 2 887379 0 1 -0.65 1.53 2.18 0

int\_rate 3 887379 0 1 -1.81 3.59 5.40 0

installment 4 887379 0 1 -1.72 4.13 5.86 0

emp\_length 5 887379 0 1 -1.68 1.39 3.07 0

annual\_inc 6 887379 0 1 -1.16 145.68 146.84 0

dti 7 887379 0 1 -2.18 2.87 5.05 0

mths\_since\_last\_delinq 8 887379 0 1 -2.22 10.01 12.23 0

pub\_rec 9 887379 0 1 -0.34 147.41 147.75 0

revol\_util 10 887379 0 1 -2.31 3.04 5.35 0

total\_acc 11 887379 0 1 -2.05 12.14 14.19 0

initial\_list\_status 12 887379 0 1 -0.97 1.03 2.00 0

out\_prncp 13 887379 0 1 -1.00 3.07 4.07 0

last\_pymnt\_amnt 14 887379 0 1 -0.45 7.16 7.61 0

tot\_cur\_bal 15 887379 0 1 -1.76 52.53 54.30 0

time\_since\_first\_credit 16 887379 0 1 -2.12 7.32 9.45 0

total\_pymnt 17 887379 0 1 -0.96 6.38 7.34 0

inq\_last\_6mths 18 887379 0 1 -0.70 32.36 33.05 0

grade 19 887379 0 1 -3.20 1.37 4.57 0

**#FINDING OPTIMAL NUMBER OF CLUSTERS**

**# USE MAP\_DBL TO RUN MANY MODELS WITH VARYING VALUE OF K**

> library(purrr)

> tot\_withinss <- map\_dbl(1:10, function(k){

+ model <- kmeans(x = loan\_num, centers = k)

+ model$tot.withinss

+ })

Warning messages:

1: Quick-TRANSfer stage steps exceeded maximum (= 44368950)

2: Quick-TRANSfer stage steps exceeded maximum (= 44368950)

3: Quick-TRANSfer stage steps exceeded maximum (= 44368950)

4: Quick-TRANSfer stage steps exceeded maximum (= 44368950)

5: Quick-TRANSfer stage steps exceeded maximum (= 44368950)

# GENERATE A DATA FRAME CONTAINING BOTH K AND TOT\_WITHINSS

> elbow\_df <- data.frame(

+ k = 1:10 ,

+ tot\_withinss = tot\_withinss

+ )

>

> # Plot the elbow plot

> print(elbow\_df)

k tot\_withinss

1 1 16860182

2 2 14743035

3 3 13622239

4 4 12855925

5 5 12250851

6 6 11913470

7 7 11803302

8 8 11294866

9 9 11122520

10 10 10908698

**#GENERATING THE ELBOW PLOT**

> ggplot(elbow\_df, aes(x = k, y = tot\_withinss)) +

+ geom\_line() +

+ scale\_x\_continuous(breaks = 1:10)

**# BUILD A KMEANS MODEL FOR HIGH-RISK, LOW RISK, AND MEDIUM RISK CUSTOMERS**

> km\_loan <- kmeans(loan\_num, center = 10)

> clust\_km\_loan <- km\_loan $cluster

> loan\_km2 <- mutate(loan\_new, cluster = clust\_km\_loan)

> count(loan\_km2,cluster)

# A tibble: 10 x 2

cluster n

*<int>* *<int>*

1 1 112967

2 2 82906

3 3 77190

4 4 36411

5 5 100318

6 6 64398

7 7 72818

8 8 152285

9 9 142782

10 10 45304

**# CALCULATE THE MEAN FOR EACH CATEGORY**

> loan\_km2 %>%

+ group\_by(cluster) %>%

+ summarise\_all(funs(mean(.)))

# A tibble: 10 x 20

cluster loan\_amnt term int\_rate installment emp\_length annual\_inc dti mths\_since\_last~ pub\_rec revol\_util total\_acc initial\_list\_st~ out\_prncp last\_pymnt\_amnt tot\_cur\_bal time\_since\_firs~ total\_pymnt inq\_last\_6mths grade

*<int>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 1 9304. 0.0151 11.9 306. 7.69 67003. 23.6 29.5 0.123 58.8 32.3 0.474 5406. 893. 152568. 7619. 4404. 0.720 5.60

2 2 18171. 0.0996 12.1 583. 6.30 84143. 16.0 34.3 0.0752 56.1 26.6 0.214 1630. 5610. 158200. 5945. 18611. 0.757 5.77

3 3 15678. 0.865 20.0 434. 6.01 61437. 21.2 33.3 0.166 62.5 24.5 0.368 9098. 1280. 106542. 5287. 7408. 1.19 3.13

4 4 25300. 0.625 16.2 728. 6.61 103956. 17.3 33.5 0.105 58.0 30.1 0.394 26.2 20272. 219752. 6366. 28964. 0.938 4.43

5 5 17524. 0.967 13.3 408. 6.45 71276. 19.9 35.2 0.127 55.7 26.2 0.825 14894. 461. 139144. 6280. 3996. 0.370 5.11

6 6 30011. 0.765 17.4 829. 6.87 105703. 20.6 31.1 0.119 65.1 30.5 0.552 24013. 814. 217728. 6880. 9384. 0.715 3.89

7 7 24108. 0.112 9.28 738. 6.50 140677. 16.4 32.3 0.0812 54.3 30.6 0.682 18775. 755. 307039. 7543. 6397. 0.452 6.28

8 8 7830. 0.0265 14.8 267. 4.18 47828. 17.6 35.4 0.0871 63.2 16.3 0.284 3392. 1103. 56732. 3975. 4969. 0.747 4.80

9 9 9233. 0.0206 8.79 290. 5.21 63226. 13.4 35.8 0.123 36.8 21.4 0.575 5339. 980. 91060. 5357. 4318. 0.393 6.47

10 10 9887. 0.120 13.9 318. 6.46 66071. 16.0 41.8 1.75 44.5 26.8 0.509 6231. 1233. 80733. 6868. 4336. 1.34 4.98

**#SUBSETTING HIGH RISK CLUSTER i.e CLUSTER 3 OBTAINED FROM PREVIOUS CLUSTRING EXCERCISE**

> Riskycust = subset(loan\_km2, cluster==3)

> Risk= Riskycust [,-c(20)]

> str(Risk)

'data.frame': 77190 obs. of 19 variables:

$ loan\_amnt : int 7000 16000 10850 20000 10500 8000 12000 25000 25000 16225 ...

$ term : int 0 0 0 0 0 0 0 0 0 0 ...

$ int\_rate : num 16.1 17 16.7 15 15.3 ...

$ installment : num 246 571 385 693 366 ...

$ emp\_length : int 10 7 0 2 0 10 1 3 2 2 ...

$ annual\_inc : num 189500 85400 55200 80000 62000 ...

$ dti : num 22.47 25.32 26.52 3.11 1.72 ...

$ mths\_since\_last\_delinq : int 0 0 31 2 43 33 24 12 27 2 ...

$ pub\_rec : int 0 0 0 0 0 0 0 0 0 0 ...

$ revol\_util : num 92.3 54.5 65.5 59.3 95.8 99.3 83.7 46.1 29.8 6 ...

$ total\_acc : int 31 51 18 6 7 23 4 47 33 20 ...

$ initial\_list\_status : int 0 0 0 0 0 0 0 0 0 0 ...

$ out\_prncp : num 0 0 0 0 0 0 0 0 0 0 ...

$ last\_pymnt\_amnt : num 246 1026 1545 693 366 ...

$ tot\_cur\_bal : int 206252 106149 33506 138179 101412 69034 115694 255416 231298 121668 ...

$ time\_since\_first\_credit: int 7214 4169 2677 3590 1492 6240 1156 3621 5021 1156 ...

$ total\_pymnt : num 1232 21125 13828 8294 2559 ...

$ inq\_last\_6mths : int 4 3 4 6 13 1 12 2 5 0 ...

$ grade : num 2 2 2 2 2 1 3 1 2 1 ...

>

|  |
| --- |
| **# FINDING OPTIMAL NUMBER OF SUB CLUSTERS FOR HIGH RISK CUSTOMERS**  **# USE MAP\_DBL TO RUN MANY MODELS WITH VARYING VALUE OF K**  > library(purrr)  > tot\_withinss <- map\_dbl(1:10, function(k){  + model <- kmeans(x = Risk, centers = k)  + model$tot.withinss  + })  >  **# GENERATE A DATA FRAME CONTAINING BOTH K AND TOT\_WITHINSS**  > elbow\_df <- data.frame(  + k = 1:10 ,  + tot\_withinss = tot\_withinss  + )  >  > # Plot the elbow plot  > print(elbow\_df)  k tot\_withinss  1 1 8.509142e+14  2 2 3.091624e+14  3 3 1.827293e+14  4 4 1.383938e+14  5 5 1.145883e+14  6 6 9.778637e+13  7 7 8.510301e+13  8 8 7.806224e+13  9 9 7.024641e+13  10 10 6.485845e+13  **#GENERATING THE ELBOW PLOT**  > ggplot(elbow\_df, aes(x = k, y = tot\_withinss)) +  + geom\_line() +  + scale\_x\_continuous(breaks = 1:10)  **# BUILD A KMEANS MODEL FOR SUBSETTING HIGH RISK CUSTOMERS**  > km\_risk<- kmeans(Risk, center = 3)  > clust\_km\_risk <- km\_risk$cluster  > risk\_km2 <- mutate(Risk, cluster = clust\_km\_risk)  > count(risk\_km2,risk\_km2$cluster)  # A tibble: 3 x 2  `risk\_km2$cluster` n  *<int>* *<int>*  1 1 5816  2 2 48579  3 3 22795  **> # CALCULATE THE MEAN FOR EACH CATEGORY**  > risk\_km2 %>%  + group\_by(cluster) %>%  + summarise\_all(funs(mean(.)))  cluster loan\_amnt term int\_rate installment emp\_length annual\_inc dti mths\_since\_last\_~ pub\_rec revol\_util total\_acc initial\_list\_sta~ out\_prncp last\_pymnt\_amnt tot\_cur\_bal  <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  1 1 15858. 0.876 19.9 435. 6.29 92725. 19.4 30.7 0.158 65.7 30.0 0.351 8768. 1255. 352345. 5786. 7435. 1.62 3.18  2 2 15450. 0.854 20.1 431. 5.80 56167. 21.5 34.2 0.173 61.3 22.6 0.391 9421. 1269. 43545. 5170. 7032. 1.08 3.09  3 3 16117. 0.883 19.7 440. 6.38 64684. 21.0 32.1 0.151 64.3 27.1 0.325 8494. 1308. 178083. 5410. 8202. 1.30 3.22 |
|  |
| |  | | --- | |  | |