"LENDING CLUB LOAN ANALYSIS"

Submitted in Partial Fulfillment of requirements for the Award of certificate of

Post Graduate Program in Business Analytics & Business Intelligence

Capstone Project Report

Submitted to

GREAT LAKES
INSTITUTE OF MANAGEMENT, GURGAON
Global Mindset - Indian Roots

Submitted by

Amit Jain

Gunjan Mansharamani

Naman Khurana

Nimisha Pandey

Rohit Gupta

Under the guidance of

Mr. Vikas Chandra

Batch- PGPBABI.G.Jan17

January 2018

CERTIFICATE

This is to certify that the participants - Amit Jain, Gunjan Mansharamani, Naman Khurana, Nimisha Pandey and Rohit Gupta, who are the students of Great Lakes Institute of Management , has successfully completed their project on “Lending Club Loan Analysis”

This project is the record of authentic work carried out by them during the academic year 2017- 2018.

Vikash

Mentor’s Name & Sign

Mr. Vikash Chandra

Program Director

Dr.Bappaditya Mukhopadhyay

Date : 15/01/2018

Place: Gurugram

ACKNOWLEDGEMENTS

We wish to place on record our deep appreciation for the guidance and help provided to us by our Mentor Mr. Vikas Chandra from Great Lakes Institute of Management for guiding us in completing this project on time.

We would also like to place on record our appreciation for the guidance provided by Dr. Bappaditya Mukhopadhyay for giving us valuable feedback and being a source of inspiration in helping us to work on this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: 15th January, 2018

Place: Gurgaon

Group Members

Amit Jain

Gunjan Mansharamani

Naman Khurana

Nimisha Pandey

Rohit Gupta

Table of Contents

Chapter 1 Introduction

1.1 Credit Risk

Credit risk refers to the risk that a borrower may not repay a loan and that the lender may lose the principal of the loan or the interest associated with it. Credit risk arises because borrowers expect to use future cash flows to pay current debts; it's almost never possible to ensure that borrowers will definitely have the funds to repay their debts. Interest payments from the borrower or issuer of a debt obligation are a lender's or investor's reward for assuming credit risk.

1.2 Credit Risk Assessment

Credit risks are calculated based on the borrowers' overall ability to repay. To assess credit risk on a consumer loan, lenders look at the five C's:

a. Applicant’s Credit history
b. Applicant’s Capacity to repay
c. Applicant’s Capital
d. The Loan's Conditions
e. Associated Collateral

1.3 Credit Risk Analytics

In today’s hyper-competitive environment, banks are continuously looking to embrace the power of analytics to gain insights and appropriately evaluate risks and opportunities - enabling more effective Decision making in the quest to enhance wallet share. Some of the key challenges that banks grapple with in the absence of a Credit Risk analytics solution are:

- Loan Delinquency and Default
- Loan Portfolio management
- Sales/Service differentiation
- Low Loan Recovery and high Charge-off

1.4 Importance of Default Analytics in Lending

Financial institutions may inject a probability of default (PD) analysis into several steps of their credit risk processes, and each use-case provides a different benefit to the bank that directly impacts its workflow efficiency, credit decision quality, and most likely profitability.

Before any financial statements have been spread, a PD analysis can be an effective way to perform pre-screens. Lenders, for example, can analyse a business in just a few minutes and quickly see if the loan could be worthwhile (yes) or one on which they should quickly pass (no). These pre-screens could reduce strain on the credit department.

It provides management with a tool with which they may perform a deeper and more objective analysis when making credit decisions. Banks or credit unions that may have previously had very small or non-existent commercial and industrial (C&I) concentrations in their portfolio might be especially interested in tools that add depth and objectivity to the analysis performed on potential borrowers. Whether a banker is used to analysing private companies or not, a PDM is an accurate and reliable addition to an existing credit analysis process.

Similarly, a financial institution can use a probability of default model to validate their risk rating process. For a loan that may be teetering between ratings, a calculated PD for a business could provide additional clarity as to its appropriate rating. In most cases, a probability of default is an excellent component to add to an existing risk rating scorecard or similar analysis.

1.5 Data Source :Lending Club

Lending Club is a Peer-to-Peer lending company that utilizes a group of private investors to fund loan requests. LC allows for debt consolidation, home and auto loans, credit card financing loans and expense financing. With such a level of financing options available to borrowers from investors, it attracts risks also in the form of borrower default and interest loss on prepayment.

Lending Club’s model for risk assessment categorizes borrowers by assigning them a grade and a subgrade based on their credit history.

Investors are presented with a list of borrowers, along with their assigned risk assessment grades, and they have the opportunity to choose which borrowers they will fund, and the percentage of funding that they will cover.

Chapter 2. Project Objectives and Data Sources

2.1 Project Objective

The objective of this analytical study is RISK ASSESSMENT AND MITIGATION. Risk Assessment allows lenders to analyse possible risks associated with a this lending model and enables them to determine which customers will take over a credit, which one will default and which one is likely to prepay the loan. Therefore to conclude, this study involves to:-

- Conduct a set of exploratory analysis and thereby apply various machine learning techniques to predict borrower's default behaviour
- Build a classification model to predict sentiment in a product review dataset. -Analyse financial data to predict loan defaults. -Use techniques for handling missing data. –Evaluating the model using precision-recall metrics.
- Predicting Bad Loans, We're going to be using the publicly available dataset of Lending Club loan performance. It's a real world data set with a nice mix of categorical and continuous variables.

2.2 Data Under Study

All data regarding this project can be from lending club official website. So far the data is available till 2017 Q1 and can be dated back to 2007. The data consists in 4 files updated every quarter on the same day as the quarterly results of the company are released. They contain information on almost all the loans issued by LC. The only loans missing from these files are the few loans where LC was not authorized to release publicly the details of the transactions.

The information available for each loan consists of all the details of the loans at the time of their issuance as well as more information relative to the latest status of loan such as how much principal has been paid so far, how much interest, if the loan was fully paid or defaulted, or if the borrower is late on payments etc.

- LOAN DATA

These files contain complete loan data for all loans issued through the time period stated, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter.

- PERSONAL DATA

Data used in this study contains multiple personal and financial factors of a customer, which are listed below.

- Socio-demographic factors (gender, age, education, marital status...)
- Product details (credit amount, bill statement...)
- Customer behaviour (repayment status, previous payment...)
- Target variable
- Default

2.3 Data Dictionary

We have used the records which belongs to all 4 quarters of Financial Year 2016 and first Quarter of 2017 of more than 500,000+ rows and 122 variables to start with.

BrowseNotesFile	Description
acceptD	The date which the borrower accepted the offer
accNowDelinq	The number of accounts on which the borrower is now delinquent.
accOpenPast24Mths	Number of trades opened in past 24 months.
addrState	The state provided by the borrower in the loan application
all_util	Balance to credit limit on all trades
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
annualInc	The self-reported annual income provided by the borrower during registration.
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
bcOpenToBuy	Total open to buy on revolving bankcards.
bcUtil	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
chargeoff_within_12_mths	Number of charge-offs within 12 months

collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
creditPullD	The date LC pulled credit for this loan
delinq2Yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinqAmnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
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
earliestCrLine	The date the borrower's earliest reported credit line was opened
effective_int_rate	The effective interest rate is equal to the interest rate on a Note reduced by Lending Club's estimate of the impact of uncollected interest prior to charge off.
emp_title	The job title supplied by the Borrower when applying for the loan.*
empLength	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.
expD	The date the listing will expire
expDefaultRate	The expected default rate of the loan.
ficoRangeHigh	The upper boundary range the borrower’s FICO at loan origination belongs to.
ficoRangeLow	The lower boundary range the borrower’s FICO at loan origination belongs to.
fundedAmnt	The total amount committed to that loan at that point in time.
grade	LC assigned loan grade
homeOwnership	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.
il_util	Ratio of total current balance to high credit/credit limit on all install acct
ils_exp_d	wholeloan platform expiration date
initialListStatus	The initial listing status of the loan. Possible values are – W, F
inq_fi	Number of personal finance inquiries
inq_last_12m	Number of credit inquiries in past 12 months
inqLast6Mths	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.
intRate	Interest Rate on the loan
isIncV	Indicates if income was verified by LC, not verified, or if the income source was verified
listD	The date which the borrower's application was listed on the platform.
loanAmnt	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.
max_bal_bc	Maximum current balance owed on all revolving accounts
memberId	A unique LC assigned Id for the borrower member.
mo_sin_old_rev_tl_op	Months since oldest revolving account opened
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
mo_sin_rcnt_tl	Months since most recent account opened

mortAcc	Number of mortgage accounts.
msa	Metropolitan Statistical Area of the borrower.
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_oldest_il_open	Months since oldest bank installment account opened
mths_since_rcnt_il	Months since most recent installment accounts opened
mthsSinceLastDelinq	The number of months since the borrower's last delinquency.
mthsSinceLastRecord	The number of months since the last public record.
mthsSinceMostRecentInq	Months since most recent inquiry.
mthsSinceRecentBc	Months since most recent bankcard account opened.
mthsSinceRecentLoanDelinq	Months since most recent personal finance delinquency.
mthsSinceRecentRevolDelinq	Months since most recent revolving delinquency.
num_accts_ever_120_pd	Number of accounts ever 120 or more days past due
num_actv_bc_tl	Number of currently active bankcard accounts
num_actv_rev_tl	Number of currently active revolving trades
num_bc_sats	Number of satisfactory bankcard accounts
num_bc_tl	Number of bankcard accounts
num_il_tl	Number of installment accounts
num_op_rev_tl	Number of open revolving accounts
num_rev_accts	Number of revolving accounts
num_rev_tl_bal_gt_0	Number of revolving trades with balance >0
num_sats	Number of satisfactory accounts
num_tl_120dpd_2m	Number of accounts currently 120 days past due (updated in past 2 months)
num_tl_30dpd	Number of accounts currently 30 days past due (updated in past 2 months)
num_tl_90g_dpd_24m	Number of accounts 90 or more days past due in last 24 months
num_tl_op_past_12m	Number of accounts opened in past 12 months
open_acc_6m	Number of open trades in last 6 months
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
open_act_il	Number of currently active installment trades
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months
openAcc	The number of open credit lines in the borrower's credit file.
pct_tl_nvr_dlq	Percent of trades never delinquent
percentBcGt75	Percentage of all bankcard accounts > 75% of limit.
pub_rec_bankruptcies	Number of public record bankruptcies
pubRec	Number of derogatory public records
purpose	A category provided by the borrower for the loan request.
reviewStatus	The status of the loan during the listing period. Values: APPROVED, NOT_APPROVED.
reviewStatusD	The date the loan application was reviewed by LC
revolBal	Total credit revolving balance
revolUtil	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
serviceFeeRate	Service fee rate paid by the investor for this loan.
subGrade	LC assigned loan subgrade
tax_liens	Number of tax liens
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
tot_coll_amt	Total collection amounts ever owed
tot_cur_bal	Total current balance of all accounts

tot_hi_cred_lim	Total high credit/credit limit
total_bal_il	Total current balance of all installment accounts
total_cu_tl	Number of finance trades
total_il_high_credit_limit	Total installment high credit/credit limit
total_rev_hi_lim	Total revolving high credit/credit limit
totalAcc	The total number of credit lines currently in the borrower's credit file
totalBalExMort	Total credit balance excluding mortgage
totalBcLimit	Total bankcard high credit/credit limit
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.
revol_bal_joint	Sum of revolving credit balance of the co-borrowers, net of duplicate balances
sec_app_fico_range_low	FICO range (high) for the secondary applicant
sec_app_fico_range_high	FICO range (low) for the secondary applicant
sec_app_earliest_cr_line	Earliest credit line at time of application for the secondary applicant
sec_app_inq_last_6mths	Credit inquiries in the last 6 months at time of application for the secondary applicant
sec_app_mort_acc	Number of mortgage accounts at time of application for the secondary applicant
sec_app_open_acc	Number of open trades at time of application for the secondary applicant
sec_app_revol_util	Ratio of total current balance to high credit/credit limit for all revolving accounts
sec_app_open_act_il	Number of currently active installment trades at time of application for the secondary applicant
sec_app_num_rev_accts	Number of revolving accounts at time of application for the secondary applicant
sec_app_chargeoff_within_12_mths	Number of charge-offs within last 12 months at time of application for the secondary applicant
sec_app_collections_12_mths_ex_med	Number of collections within last 12 months excluding medical collections at time of application for the secondary applicant
sec_app_mths_since_last_major_derog	Months since most recent 90-day or worse rating at time of application for the secondary applicant
disbursement_method	The method by which the borrower receives their loan. Possible values are: CASH, DIRECT_PAY

Chapter 3 Analytical Approach

3.1 Data Cleaning & Processing

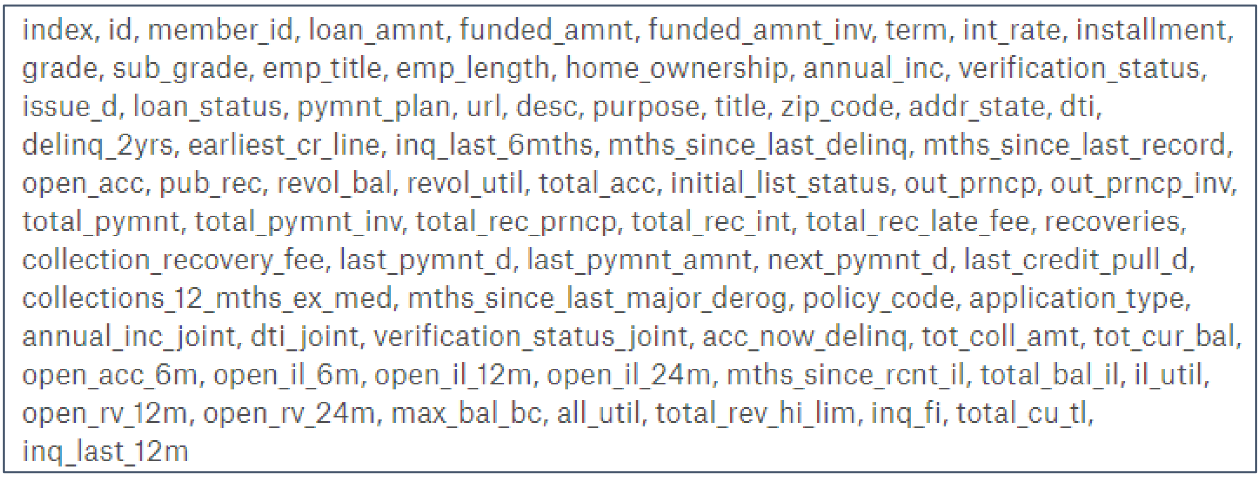
Lending Club provided us with 2 years of historical data (2016-2017). This dataset contained information pertaining to the borrower's past credit history and Lending Club loan information. The total dataset consisted of over 500,000 records, which was sufficient for our team to conduct analysis models. Variables present within the dataset provided an ample amount of information which we could use to identify relationships and gauge their effect upon the success or failure of a borrower fulfilling the terms of their loan agreement.

We required only the variables that had a direct or indirect response to a borrower's potential to default. To achieve this, we prepared the data by choosing select variables that would best fit this criteria.

Prior to data mining model analysis, the data was reviewed, cleaned and prepared as follows:

- Removed columns that obviously had no relation to the analysis in question (E.g. Applicant ID, Employee Title etc.)
- Removed columns that had bad quality data (i.e. missing values in observations, unintelligible values etc.)
- Removed columns that had identical relationships to the analysis in question (E.g. funded_amnt and funded_amnt_inv as they are always the same as loan_amt)
- Established derived columns from existing columns to facilitate model analysis (E.g. Credit_History_years, Converted binary dependent variable column called "default", mths_since_last_delinq etc.)
- Converted continuous variables to range of values to enhance interpretation of results (E.g. loan_amt, int_rate, Annual_income, credit_history_years, revol_util, total_pymnt etc.)

  Variables after cleaning and Pre processing

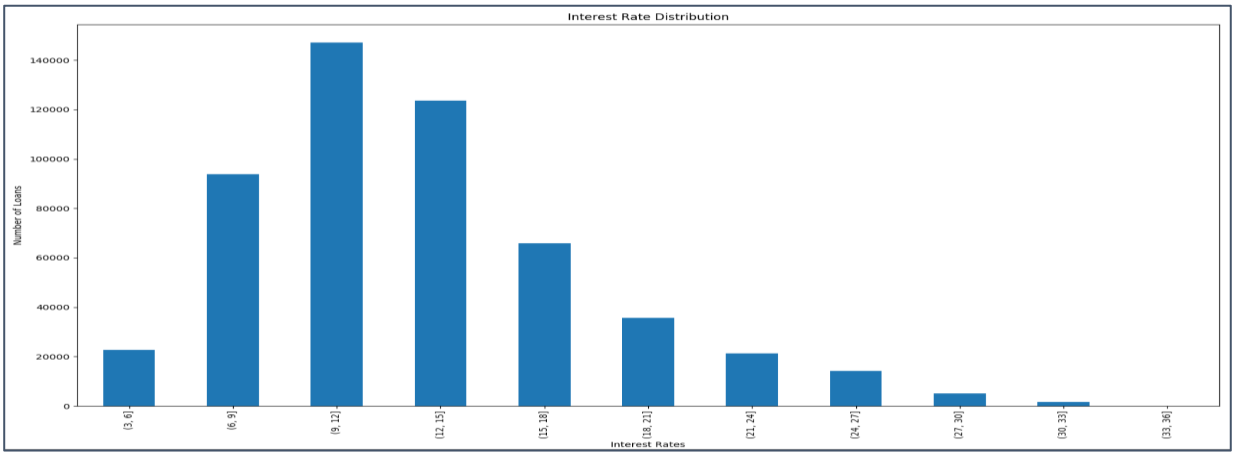


index, id, member_id, loan_amnt, funded_amnt, funded_amnt_inv, term, int_rate, installment, grade, sub_grade, emp_title, emp_length, home_ownership, annual_inc, verification_status, issue_d, loan_status, pymnt_plan, url, desc, purpose, title, zip_code, addr_state, dti, delinq_2yrs, earliest_cr_line, inq_last_6mths, mths_since_last_delinq, mths_since_last_record, open_acc, pub_rec, revol_bal, revol_util, total_acc, initial_list_status, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, next_pymnt_d, last_credit_pull_d, collections_12_mths_ex_med, mths_since_last_major_derog, policy_code, application_type, annual_inc_joint, dti_joint, verification_status_joint, acc_now_delinq, tot_coll_amt, tot_cur_bal, open_acc_6m, open_il_6m, open_il_12m, open_il_24m, mths_since_rcnt_il, total_bal_il, il_util, open_rv_12m, open_rv_24m, max_bal_bc, all_util, total_rev_hi_lim, inq_fi, total_cu_tl, inq_last_12m

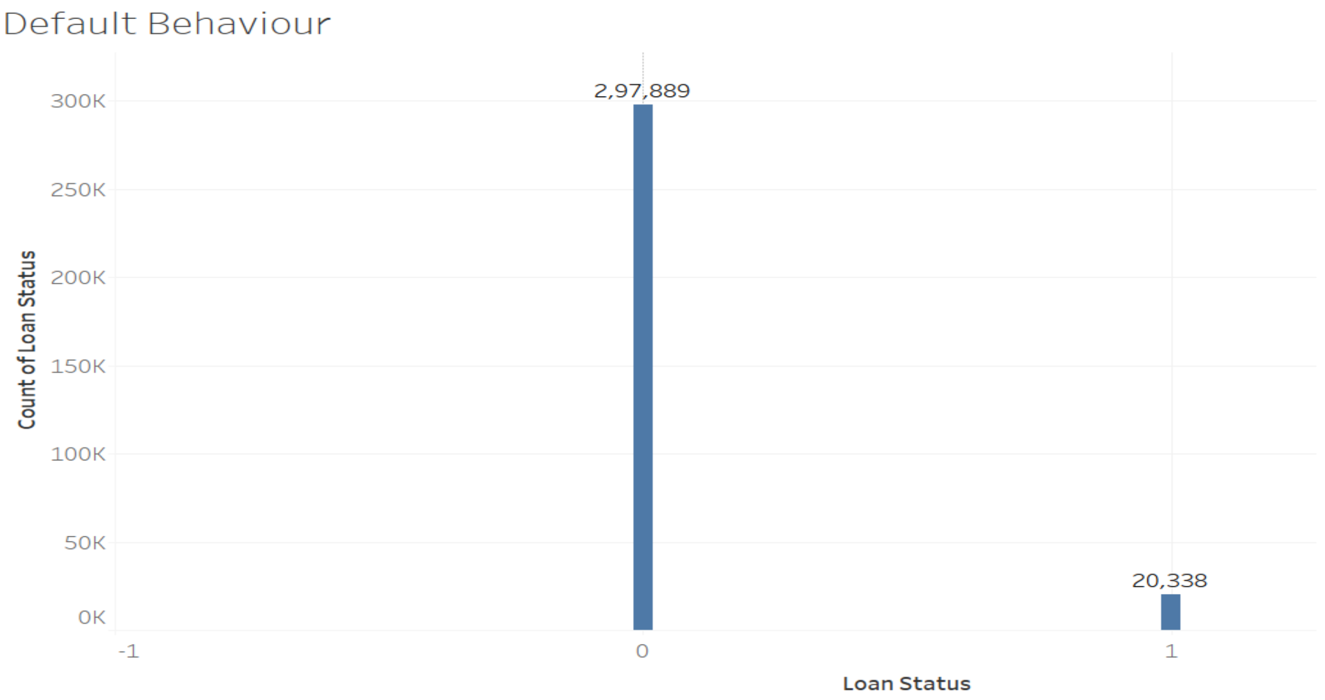
3.1 Exploratory Data Analysis

➢ Distribution of interest rates:

As per the above graphs the maximum distribution of loan is on the interests rate between 9-12% i.e. more than (1,40,000). Followed by interests rates of 12-15%. Examination of the data shows that there are people who have taken the loan at maximum interests rates i:e 30-33%. But the minimum interest rate is 3-6% which is very less & easy to pay. This bar graph is extreme right skewed. These high interest loans shall require further analysis as they could be potential candidates for defaulting.

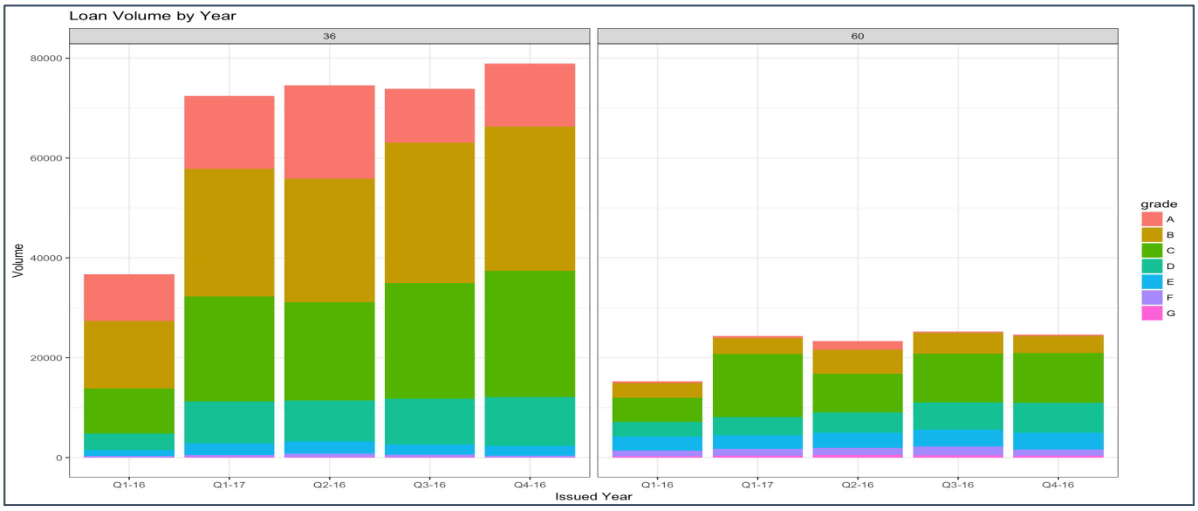


- **Default behaviour:-** Approximately 6.82 % Defaulters are there in the Train Data.



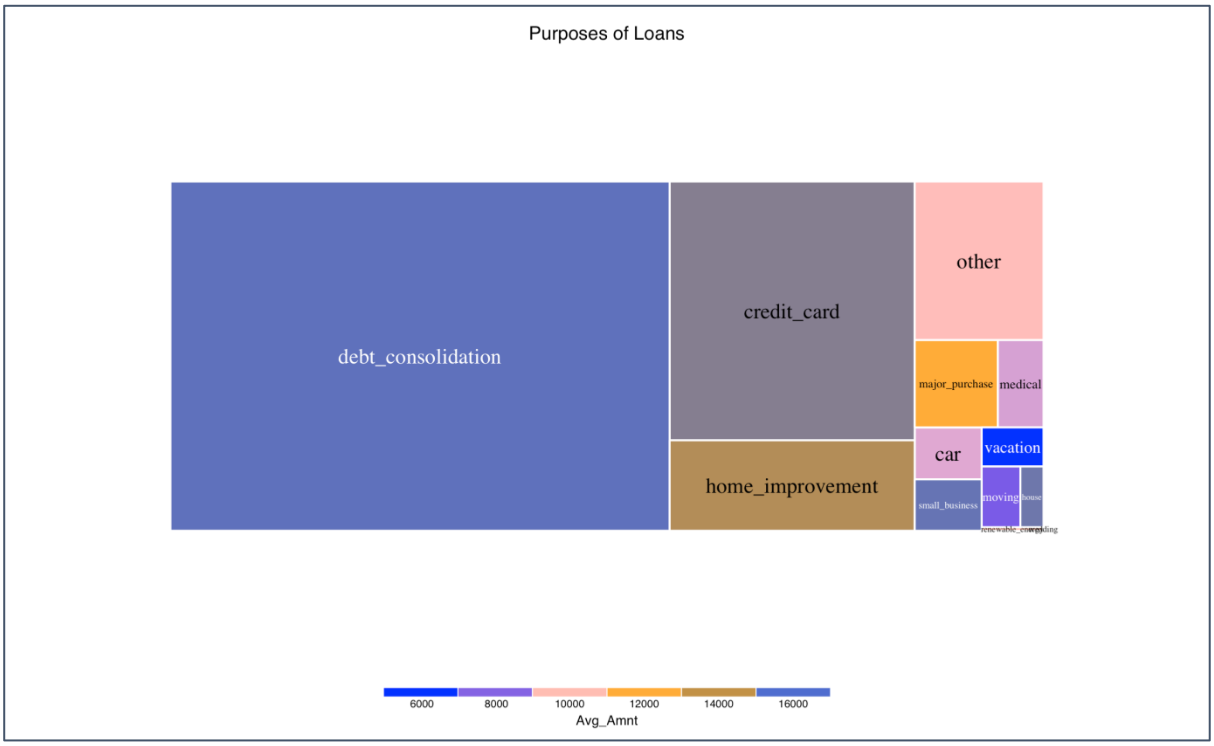
- **Loan Volume by Grade**

No of loans issued for 60 months term is invariably low as compared to loans issued for 36 months term. Grade distribution across various quarters of the year is shown in the bar chart.

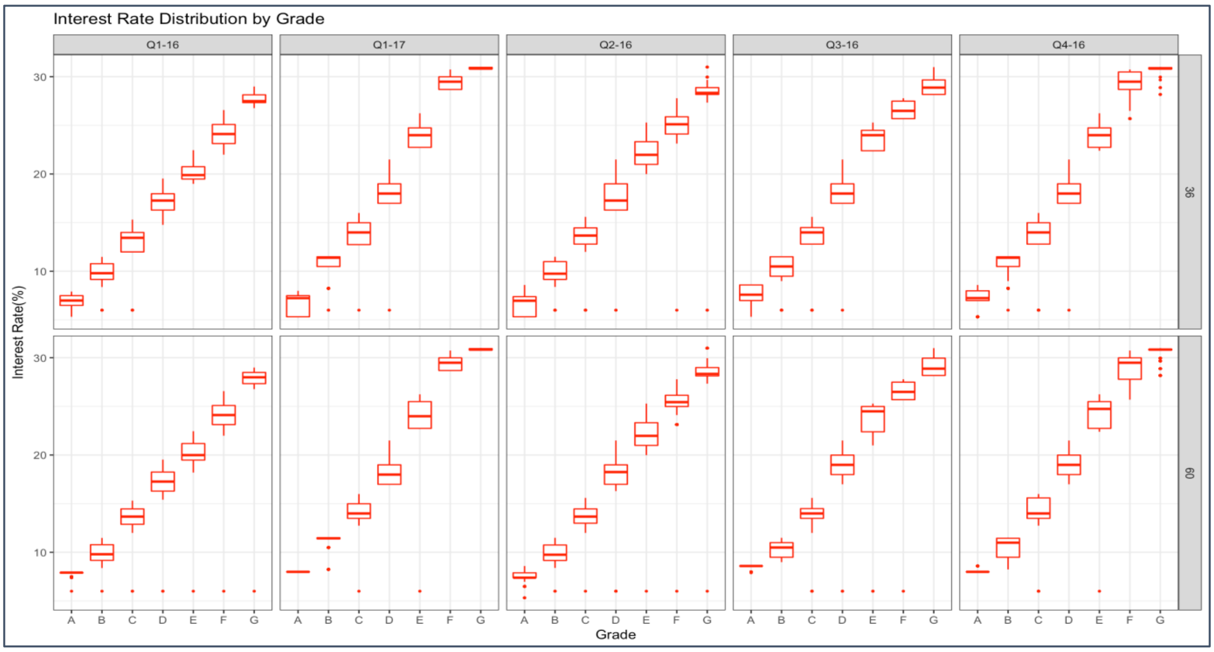


➢ Volume of loans by Purpose

Debt consolidation is the most common reason for borrowing followed by credit card and home improvement purposes. The greatest advantage of peer-to-peer lending is the low cost. Loans issued by LC usually charge lower interest rates compared with money provided by traditional banks. Most consumers choose to consolidate debt to enjoy lower borrowing costs.



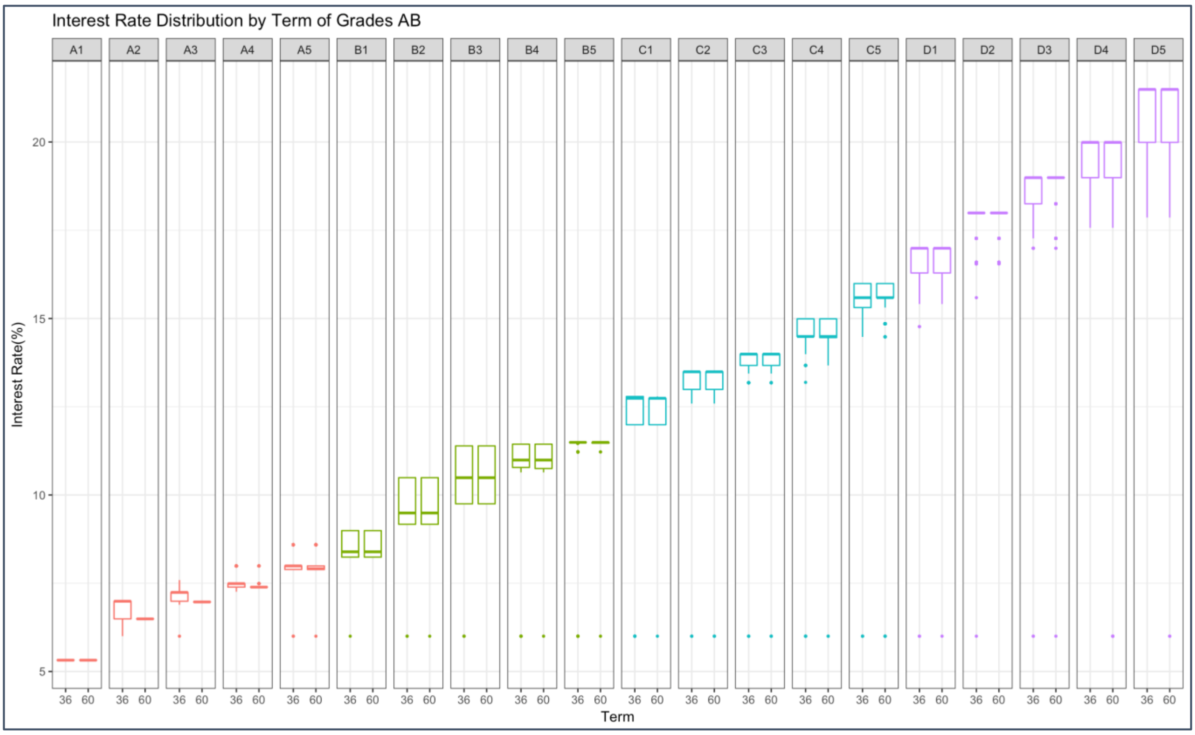
➢ Interest Rate Change, Time and Term



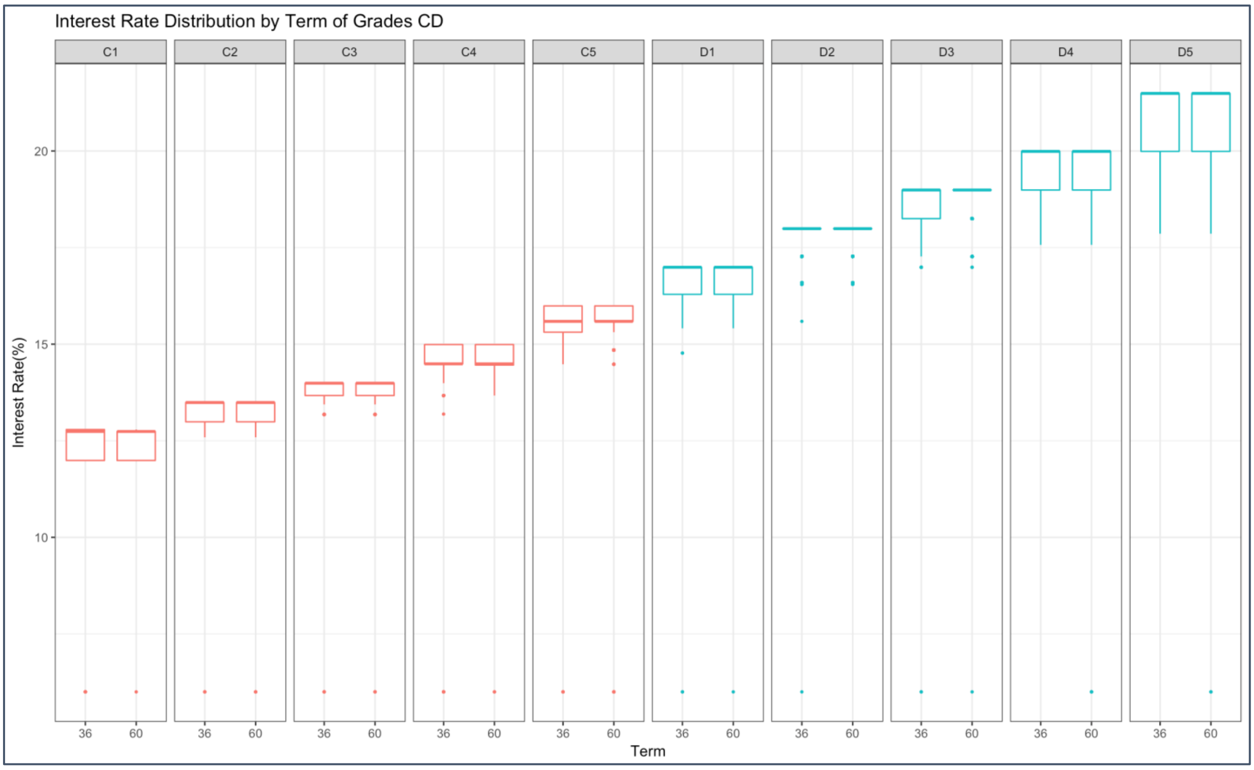
As seen above, outliers can be seen in all the quarters across grades i.e. lower rates of interest are contracted with borrowers even if they fall in grade C-G. This contradicts the policy of lending club which states that higher interest are offered to borrowers falling in higher grades.

In order to better explore the relationship among terms, grades, and interest rate, we separately analyze the distribution of interest rate for all subgrades.

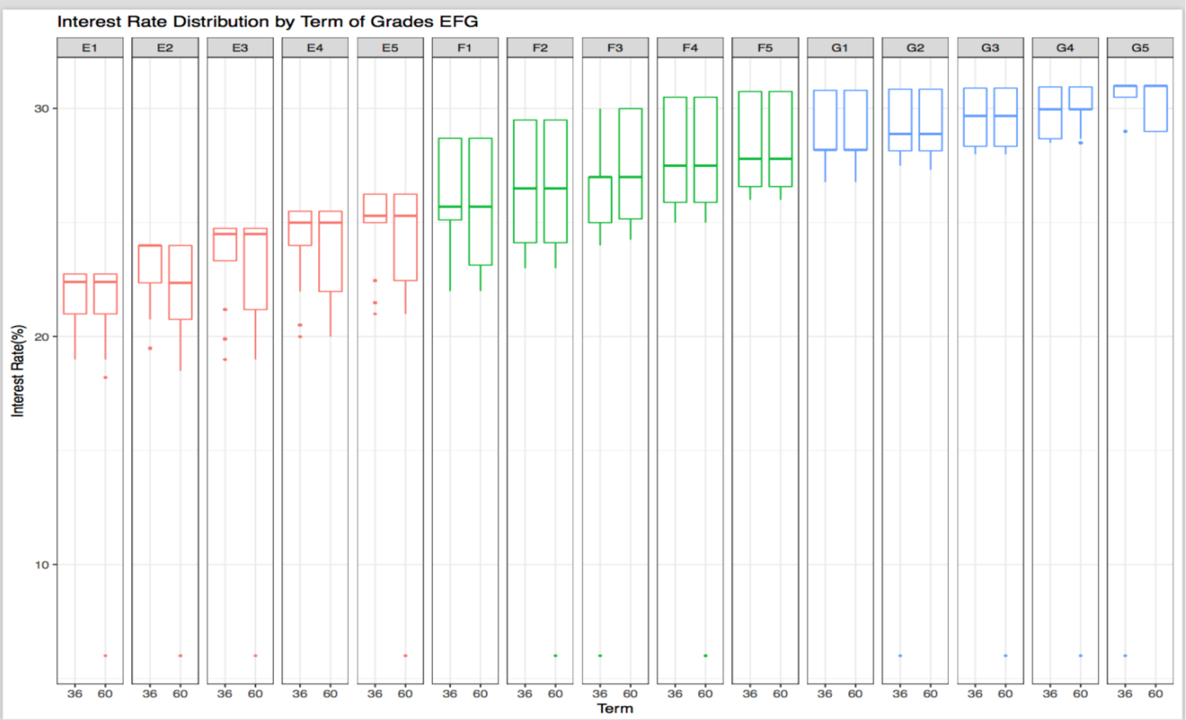
- **Interest Rate Distribution by Term of Grades A and B**



- **Interest Rate Distribution by Term of Grades C and D**



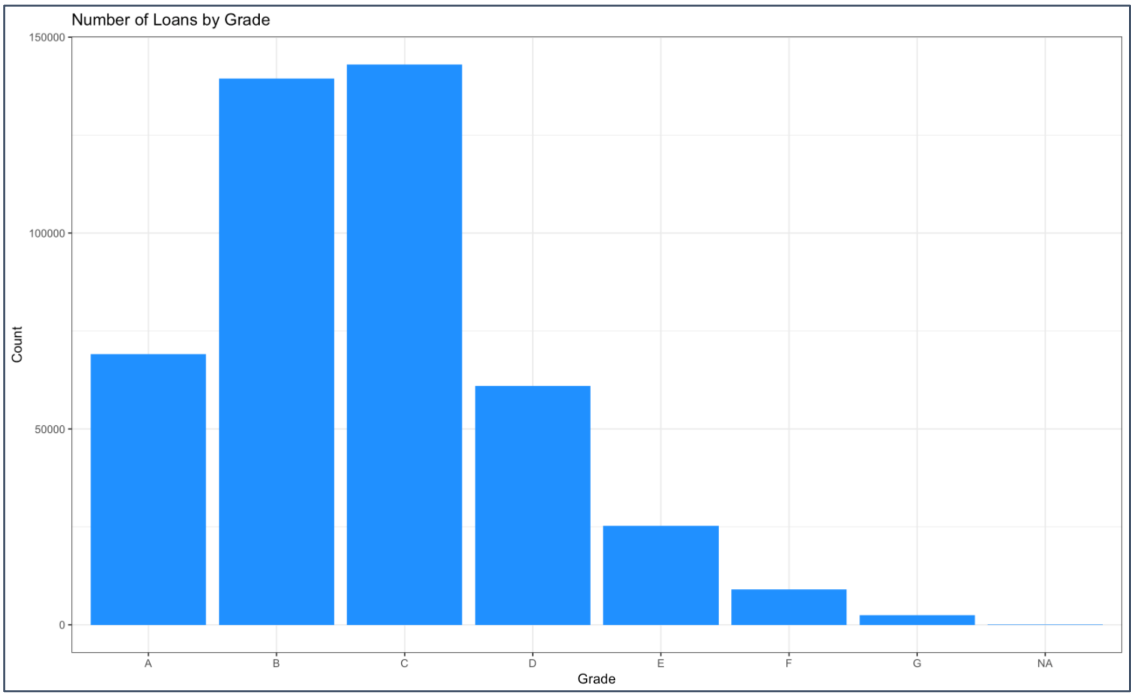
- **Interest Rate Distribution by Term of Grades E F &G**



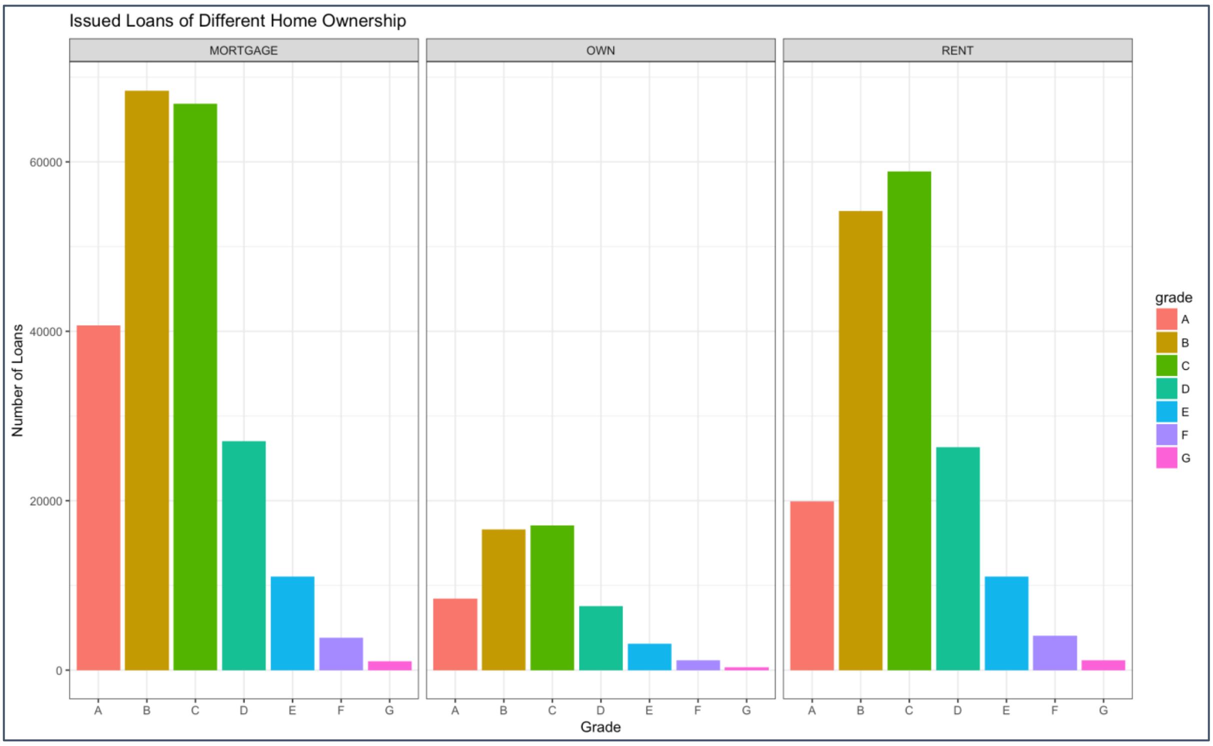
0

➢ Grade and Default

After exploring some data about evaluating grades, we continue to examine the effect of grades.

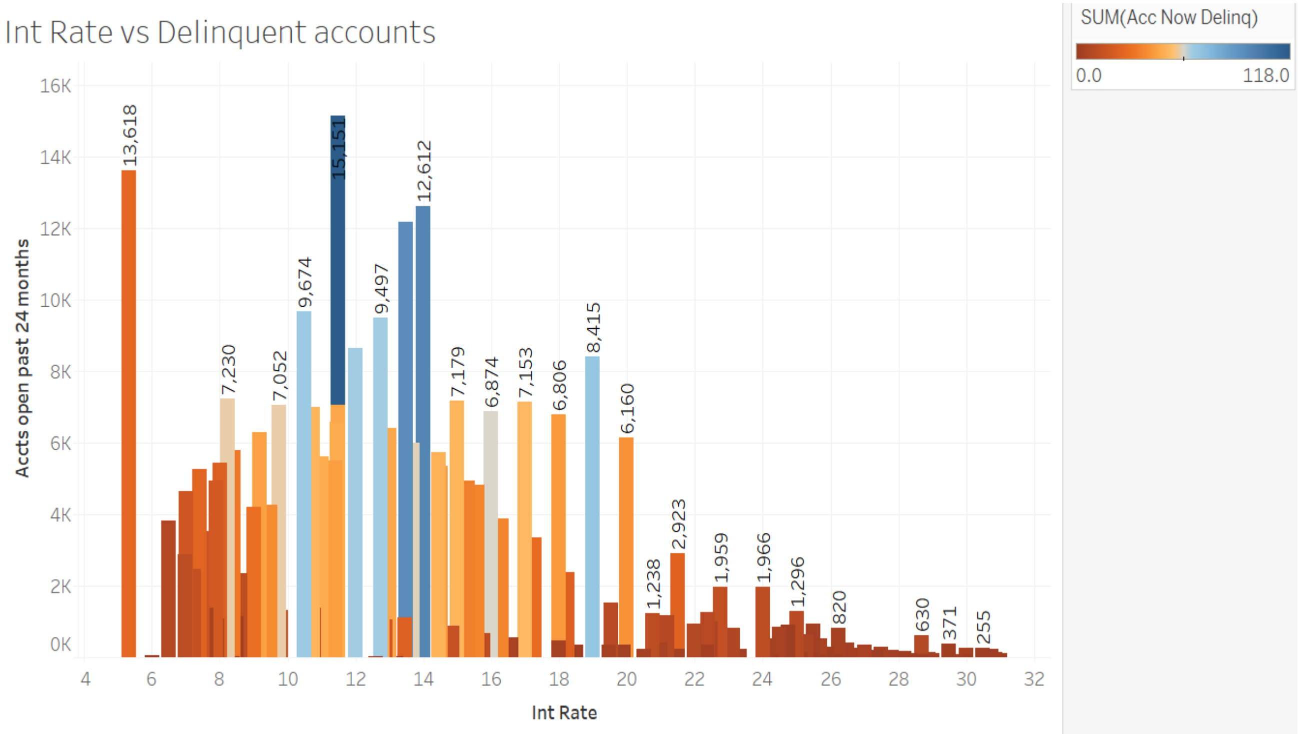


➢ Grade vs Home Ownership



Obviously, people in ‘MORTGAGE’ and ‘RENT’ have much more demands of borrowing money than people in ‘OWN’ based on the bar chart. That’s because people who own a house usually have better financial situation than others.

- **Interest Rate vs Delinquent Accounts:-** As can be seen, interest rates ranging from 12 to 14% has most number of delinquent accounts.



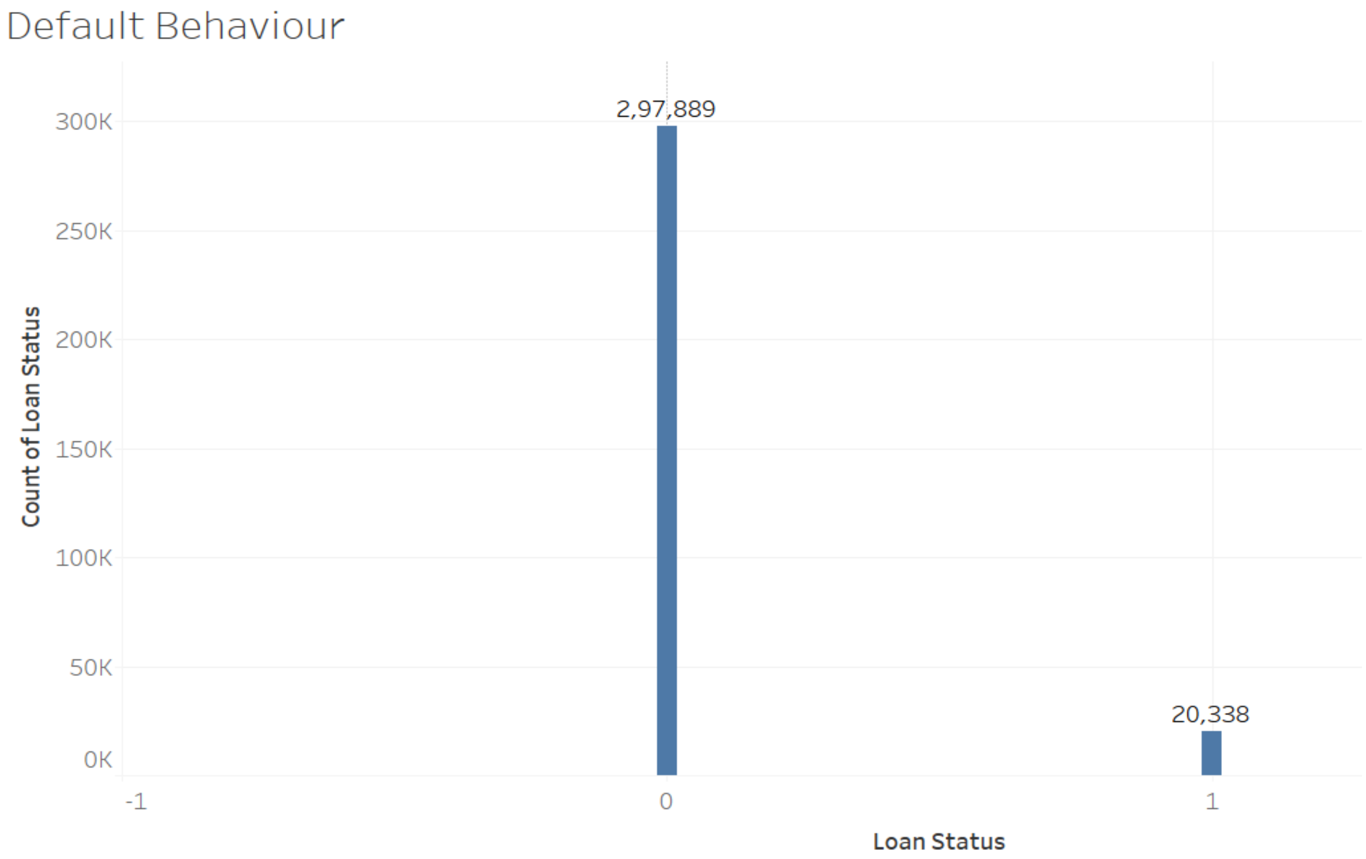
Chapter 3: Treating imbalanced dataset using SMOTE algorithm

Class Imbalance Problem- It is the problem in machine learning where the total number of a class of data (positive) is far more than the total number of another class of data (negative). When the number of instances of one class far exceeds the other, problems arise. As seen here, total no of defaulters are far less than total non defaulters.

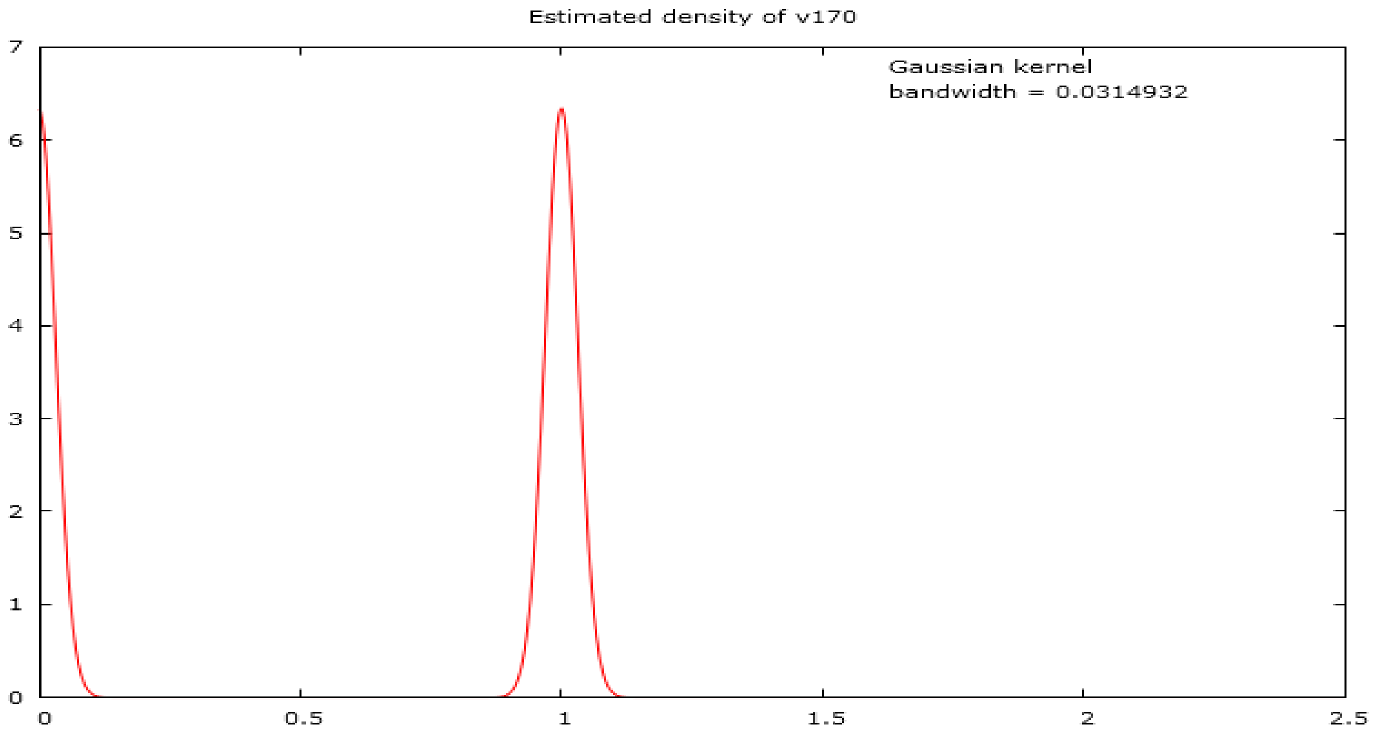
Synthetic Minority Oversampling Technique (SMOTE) is a very popular oversampling method that was proposed to improve random oversampling.

Effects of undersampling of majority class and oversampling of minority class can be seen below.

Before SMOTE



After SMOTE


Estimated density of v170
Gaussian kernel
bandwidth = 0.0314932
7
6
5
4
3
2
1
0
0
0.5
1
1.5
2
2.5

Chapter 5 Predictive Analytics

Predictive analytics encompasses a variety of statistical techniques from predictive modelling, machine learning, and data mining that analyze current and historical facts to make predictions about future or otherwise unknown events

5.1 Modelling Techniques

We have used four modelling techniques depicted on the below diagram to build the predictive models and there by predict the customers who are likely to default

1. Logistic Regression
2. K Nearest Neighbors
3. Random Forest

5.1.1 Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where p is the probability of presence of the characteristic of interest.
The logit transformation is defined as the logged odds:

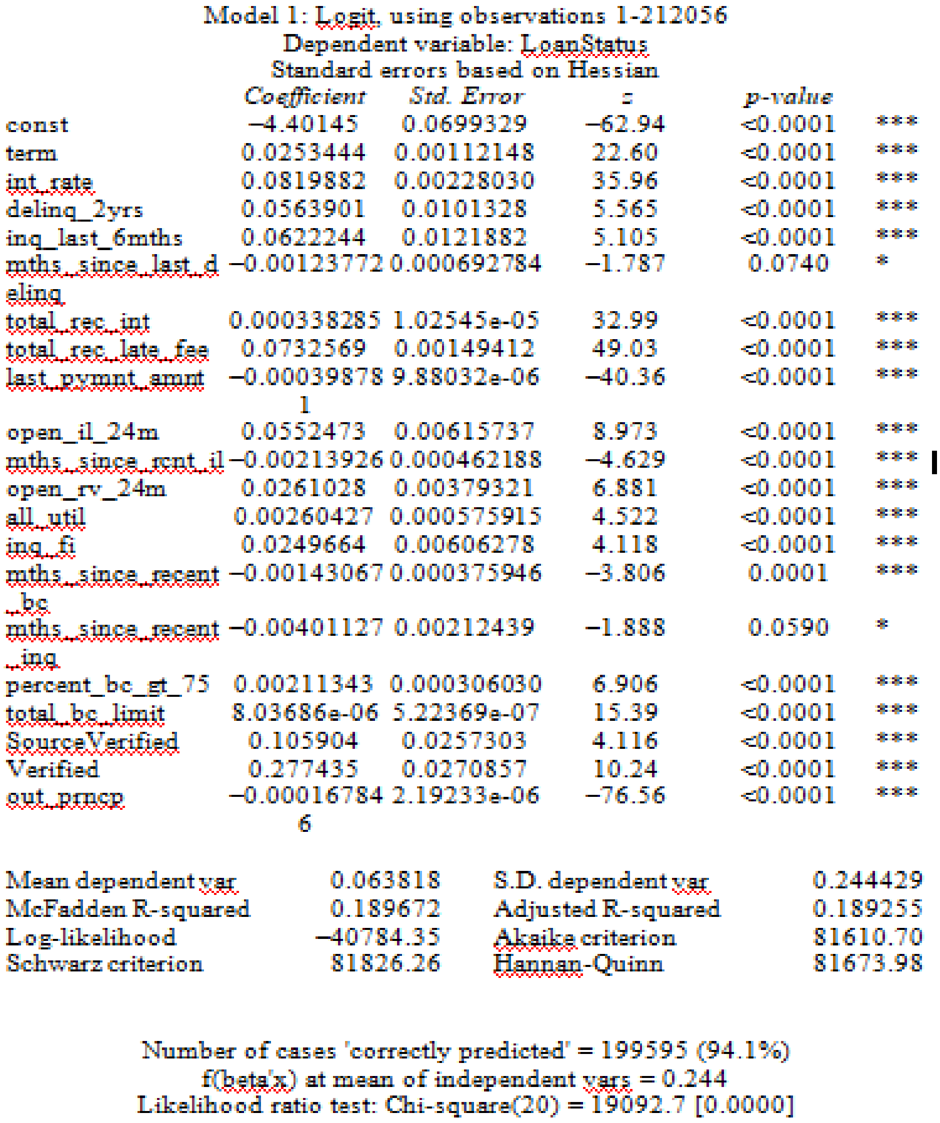
$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

and

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

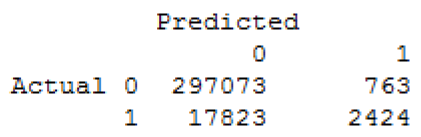
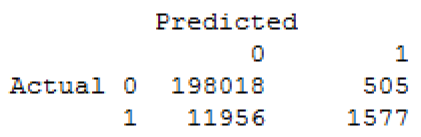
Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

Model Output using Gretl:



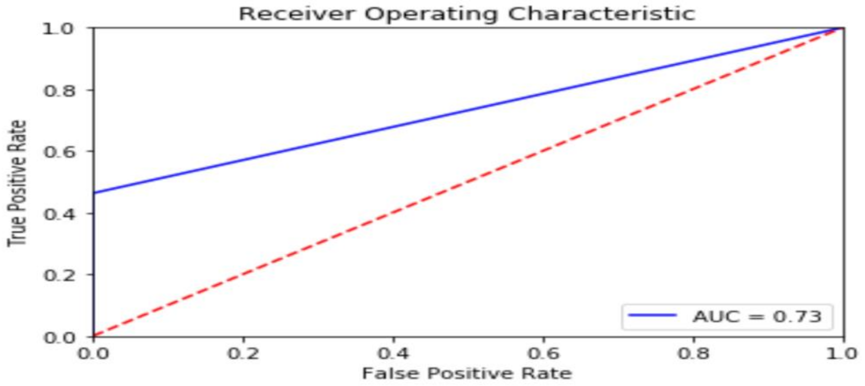
```
                 Model 1: Logit, using observations 1-212056
                      Dependent variable: LoanStatus
                      Standard errors based on Hessian
                    Coefficient     Std. Error       z          p-value
const                −4.40145      0.0699329      −62.94       <0.0001    ***
term                 0.0253444     0.00112148      22.60       <0.0001    ***
int_rate             0.0819882     0.00228030      35.96       <0.0001    ***
delinq_2yrs          0.0563901     0.0101328       5.565       <0.0001    ***
inq_last_6mths       0.0622244     0.0121882       5.105       <0.0001    ***
mths_since_last_d  −0.00123772  0.000692784     −1.787        0.0740     *
elinq
total_rec_int       0.000338285  1.02545e-05      32.99       <0.0001    ***
total_rec_late_fee   0.0732569     0.00149412      49.03       <0.0001    ***
last_pymnt_amnt    −0.00039878  9.88032e-06     −40.36       <0.0001    ***
                        1
open_il_24m          0.0552473     0.00615737      8.973       <0.0001    ***
mths_since_rcnt_il −0.00213926  0.000462188     −4.629       <0.0001    ***
open_rv_24m          0.0261028     0.00379321      6.881       <0.0001    ***
all_util            0.00260427    0.000575915      4.522       <0.0001    ***
inq_fi               0.0249664     0.00606278      4.118       <0.0001    ***
mths_since_recent  −0.00143067  0.000375946     −3.806        0.0001     ***
_bc
mths_since_recent  −0.00401127   0.00212439     −1.888        0.0590     *
_inq
percent_bc_gt_75    0.00211343    0.000306030      6.906       <0.0001    ***
total_bc_limit      8.03686e-06   5.22369e-07      15.39       <0.0001    ***
SourceVerified       0.105904      0.0257303       4.116       <0.0001    ***
Verified             0.277435      0.0270857       10.24       <0.0001    ***
out_prncp          −0.00016784  2.19233e-06     −76.56       <0.0001    ***
                        6

Mean dependent var         0.063818     S.D. dependent var        0.244429
McFadden R-squared         0.189672     Adjusted R-squared        0.189255
Log-likelihood            −40784.35     Akaike criterion          81610.70
Schwarz criterion          81826.26     Hannan-Quinn              81673.98


          Number of cases 'correctly predicted' = 199595 (94.1%)
             f(beta'x) at mean of independent vars = 0.244
          Likelihood ratio test: Chi-square(20) = 19092.7 [0.0000]
```

Model Performance

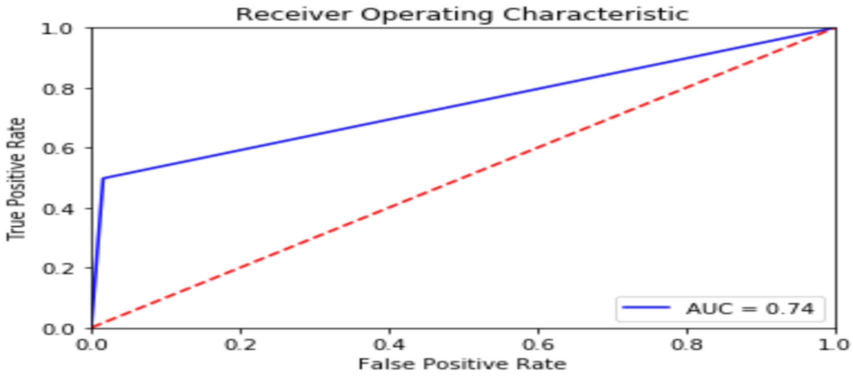
Performance metrics	**On Train Data**			**On Test Data**		
Accuracy	94.1%			94.1%		
Confusion Matrix	 Predicted 0 1 Actual 0 297073 763 1 17823 2424			 Predicted 0 1 Actual 0 198018 505 1 11956 1577		
Logarithmic Loss	1.186			1.624		
AUC	73.15%			74.09%		
Classification report	0	1	Avg/total	0	1	Avg/total
precision	0.96	1.00	0.97	0.97	0.68	0.95
Recall	1	0.46	0.97	0.98	0.50	0.95
F1-score	0.98	0.63	0.96	0.98	0.57	0.95
support	297889	20338	318227	198694	13458	212152

ROC-AUC Curve

On Train Data



On Test Data



5.1.2 k-Nearest Neighbour's Algorithm

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

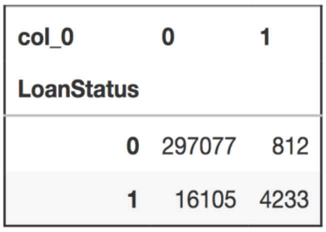
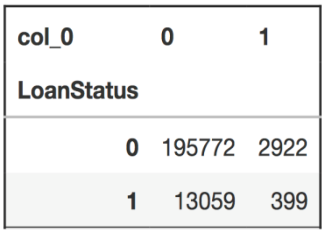
In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

Both for classification and regression, a useful technique can be to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.[2]

The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data. The algorithm is not to be confused with k-means, another popular machine learning technique.

Performance metrics	**On Train Data**			**On Test Data**		
Accuracy	94.68%			92.47%		
Confusion Matrix	 col_0 0 1 LoanStatus 0 297077 812 1 16105 4233			 col_0 0 1 LoanStatus 0 195772 2922 1 13059 399		
Logarithmic Loss	1.836			2.602		
AUC	0.603			0.507		
Classification report	0	1	Avg/total	0	1	Avg/total
precision	0.95	0.84	0.94	0.94	0.12	0.89
Recall	1	0.21	0.95	0.99	0.03	0.92
F1-score	0.97	0.33	0.93	0.96	0.05	0.9
support	297889	20338	318227	198694	13458	212152

5.1.3 Random Forest

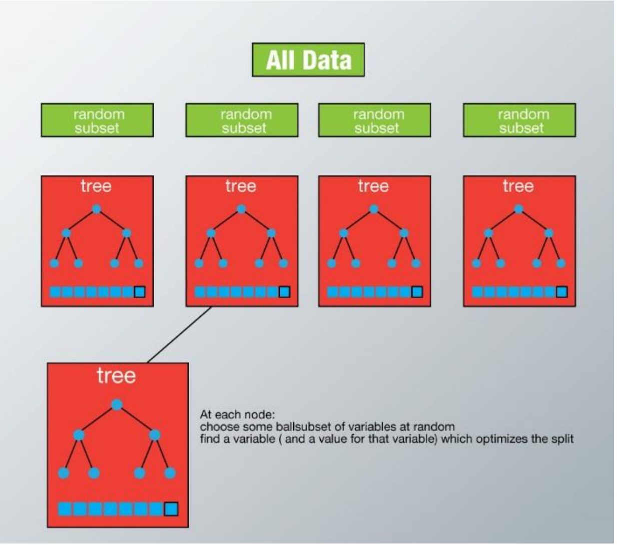
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Random forests can be used to rank the importance of variables in a regression or classification problem in a natural way

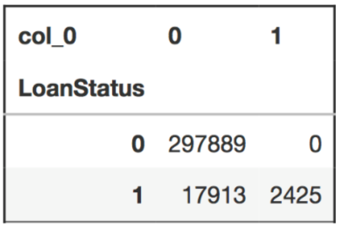
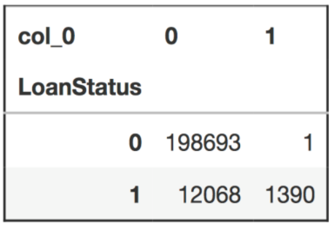
Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration, and does a fairly good job. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.

In Random Forest, we grow multiple trees as opposed to a single tree in CART model

To classify a new object based on attributes, each tree gives a classification and we say the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest) and in case of regression, it takes the average of outputs by different trees.



At each node:
choose some ballsubset of variables at random
find a variable (and a value for that variable) which optimizes the split

Performance metrics	**On Train Data**			**On Test Data**		
Accuracy	94.37%			94.31%		
Confusion Matrix	 col_0 0 1 LoanStatus 0 297889 0 1 17913 2425			 col_0 0 1 LoanStatus 0 198693 1 1 12068 1390		
Logarithmic Loss	1.944			1.964		
AUC	0.560			0.552		
Classification report	0	1	Avg/total	0	1	Avg/total
precision	0.94	1	0.95	0.94	1	0.95
Recall	1	0.12	0.94	1	0.1	0.94
F1-score	0.97	0.21	0.92	0.97	0.19	0.92
support	297889	20338	318227	198694	13458	212152

5.2 Model comparison:

	On Train Data		**On Test Data**	
	Sensitivity	Specificity	Sensitivity	Specificity
Logistic Regression	0.98	0.46	0.99	0.49
kNN	0.98	0.21	0.98	0.03
Random Forest	0.99	0.12	0.99	0.01

As seen in the above comparison, all the models perform similar on the given data set. Though, for any organizations the success of the model is more important to them than the accuracy of classification. The organization believe in identifying good customers with a conservative approach.

Chapter 6 Conclusions & Recommendations

6.1 Conclusions:

As can be seen from the comparison results, the logistic regression performed better than the other models in terms of Fit measures, Area-under-curve from the ROC, Confusion Matrix as well as the Lift Chart.

Random forest followed closely with comparable results.

Other notable findings through our analysis was that the independent variables relating to the total amount of payments that a borrower had made and the amount of the loan had the strongest relationship with the response variable. Throughout all analysis models, these two variables returned the highest degree of correlation to the response variable than any other model.

6.2 Recommendations:

- The higher the loan amount, the higher the likelihood of default Choose loans that $9000 or less.
- Loans with term of 36 months tended to be defaulted a lot more than loans with term of 60 months. Choose loans with 60 month terms.
- Certain sub-grades were almost certain to default compared to other sub-grades.
- Selecting loans of subgrade B5 and higher will result in a 90% chance of repayment.

6.3 Future expansion:

The Lending Club data contains a few fields which are not immediately usable in learning models, namely zip code and loan description. Since there are many zip codes, expanding them into boolean columns as we did for categorical features will not be effective. Instead, we joined census data with the Lending Club data. The description of the loan contains freeform text input by the loan requestor, and thus may contain keywords that correlate with defaulting or non-defaulting loans.

Appendix:

I. Tools :

- R Studio
- Python
- Gretl
- M.S. Excel

II. Model Performance Metrics

- **Confusion Matrix**: The confusion matrix is a handy presentation of the accuracy of a model with two or more classes. The table presents predictions on the x-axis and accuracy outcomes on the y-axis. The cells of the table are the number of predictions made by a machine learning algorithm.
- **Overall Accuracy**: Classification accuracy is the number of correct predictions made as a ratio of all predictions made.
- **Classification Report:**
  - **Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labelled as survived, how many actually survived? High precision relates to the low false positive rate.

    Precision = TP/TP+FP

  - **Recall** (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The question recall answers is: Of all the passengers that truly survived, how many did we label? We have got recall of 0.631 which is good for this model as it’s above 0.5.

    Recall = TP/TP+FN

  - **F1 score -** F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost.

    F1 Score = 2*(Recall * Precision) / (Recall + Precision)

- **Logarithmic Loss**: Logarithmic loss (or log loss) is a performance metric for evaluating the predictions of probabilities of membership to a given class.

  The scalar probability between 0 and 1 can be seen as a measure of confidence for a prediction by an algorithm. Predictions that are correct or incorrect are rewarded or punished proportionally to the confidence of the prediction.
- **AUC** Area under ROC Curve (or AUC for short) is a performance metric for binary classification problems. The AUC represents a model's ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random.
- **Sensitivity** is the true positive rate also called the recall. It is the number instances from the positive (first) class that actually predicted correctly.
- **Specificity** is also called the true negative rate. Is the number of instances from the negative class (second) class that were actually predicted correctly.