

## 

“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

## 

## 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.



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 Group Members

Place: Gurgaon Amit Jain

Gunjan Mansharamani

Naman Khurana

Nimisha Pandey

Rohit Gupta

|  |  |
| --- | --- |
| **1** | **Introduction** |
| 1.1 | Credit Risk |
| 1.2 | Credit Risk Assessment |
| 1.3 | Credit Risk Analytics |
| 1.4 | Importance Of Default Analytics In Lending |
| 1.5 | Data Source :Lending Club |
| **2** | **Project Objectives & Data Sources** |
| 2.1 | Project Objective |
| 2.2 | Data Under Study |
| 2.3 | Data Dictionary |
| **3** | **Analytical Approach** |
| 3.1 | Data Cleaning & Processing |
| 3.2 | Exploratory Data Analysis |
| **4** | **Treating Imbalanced dataset using SMOTE** |
| **5** | **Predictive Analytics** |
| 5.1 | Modelling Techniques |
| 5.1.1. | Logistic Regression |
| 5.1.2 | K-Nearest Neighbor’s Algorithm |
| 5.1.3 | Random Forest |
| 5.2 | Model Comparison |
| **6** | **Conclusions And Recommendations** |
| 6.1 | Recommendations |
| 6.2 | Future Expansions |
|  | **Appendix** |

**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:

1. Applicant’s Credit history
2. Applicant’s Capacity to repay
3. Applicant’s Capital
4. The Loan's Conditions
5. 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](https://www.lendingclub.com/info/download-data.action). 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**

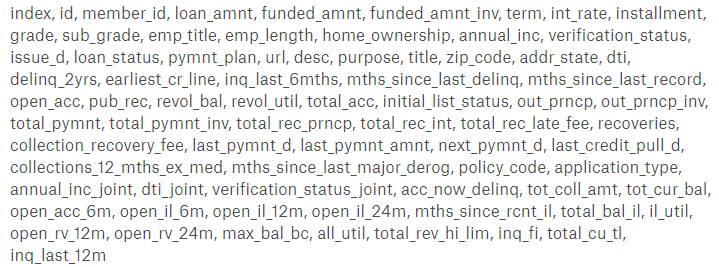
Lending Club provided us with 2 years of historical data (2016-2017). This dataset contained information pertaining to the borrower’s past credit history and Lending Club loan information. The total dataset consisted of over 500,000 records, which was sufficient for our team to conduct analysis models. 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.)

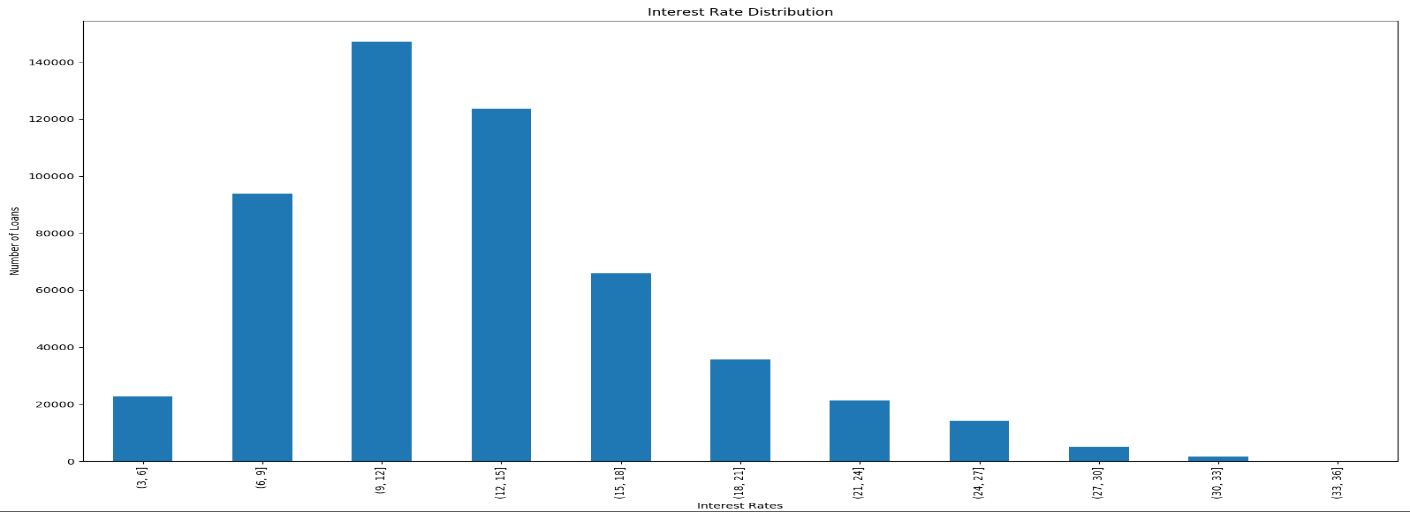
Variables after cleaning and Pre processing



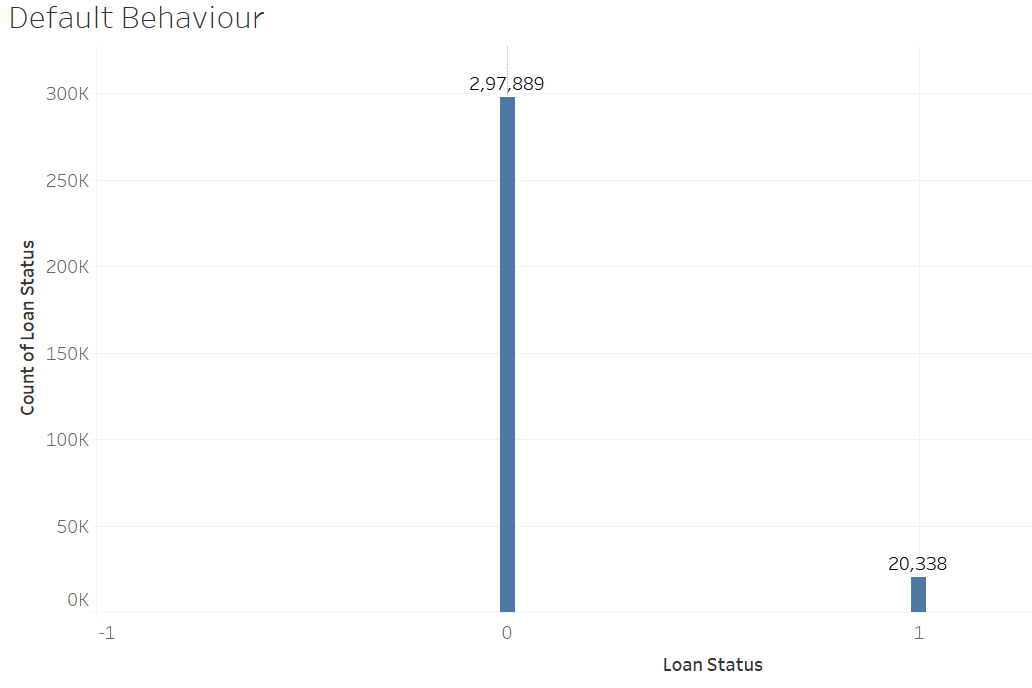
**3.1 Exploratory Data Analysis**

* **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.

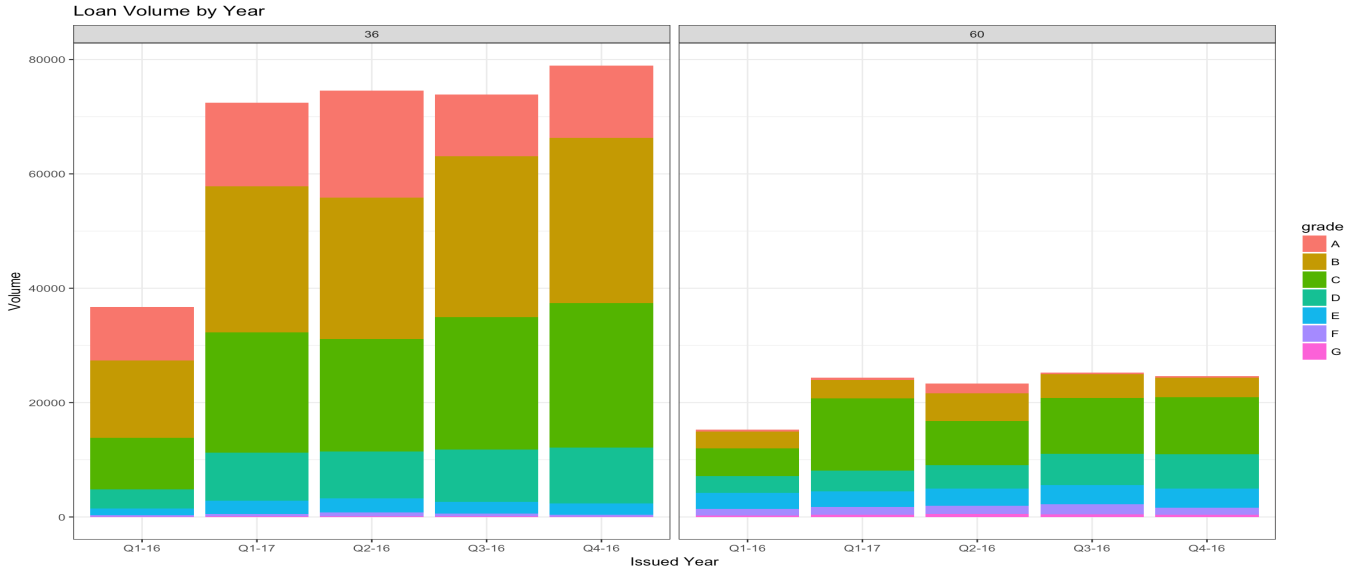


* **Default behaviour:-** Approximately 6.82 % Defaulters are there in the Train Data.

****

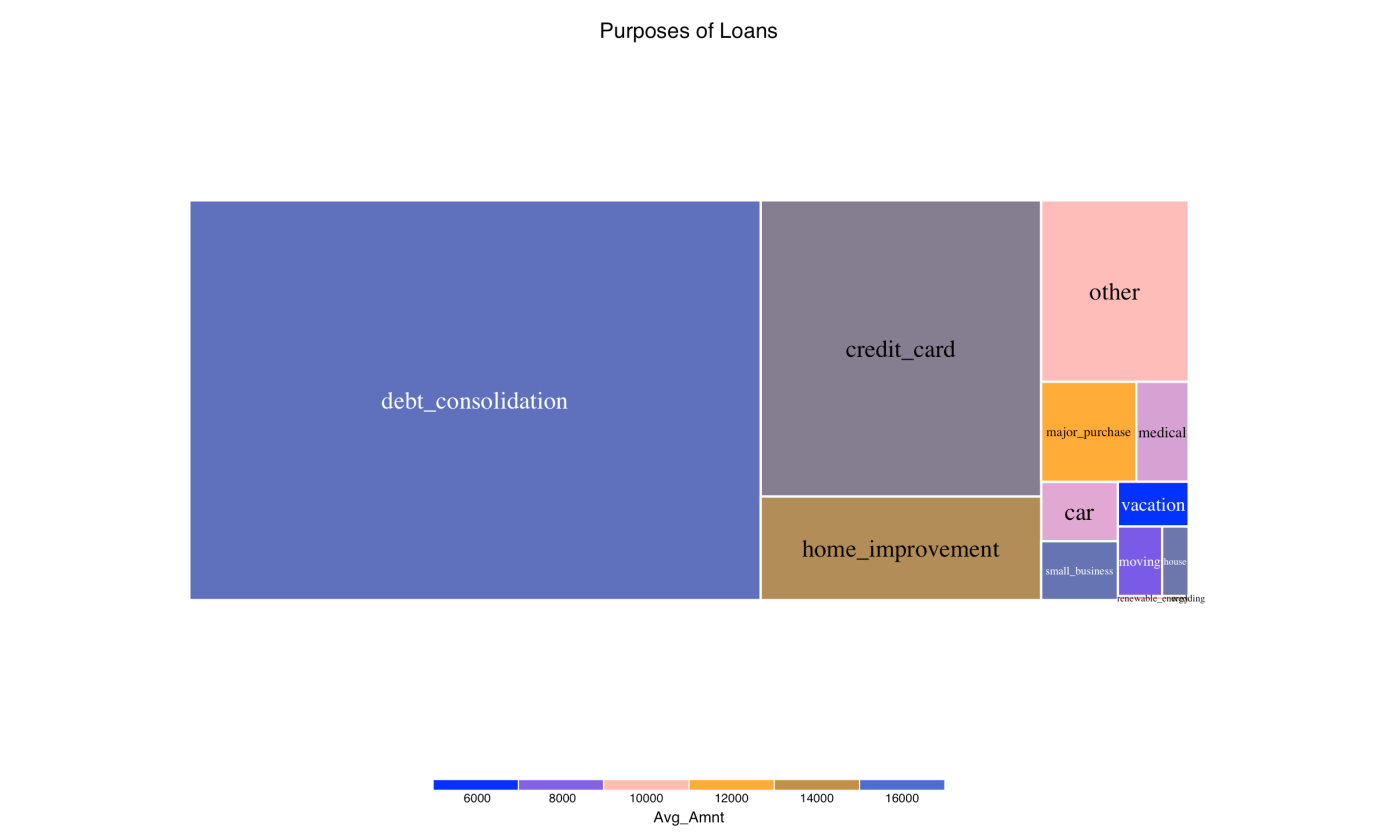
* **Loan Volume by Grade**

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.

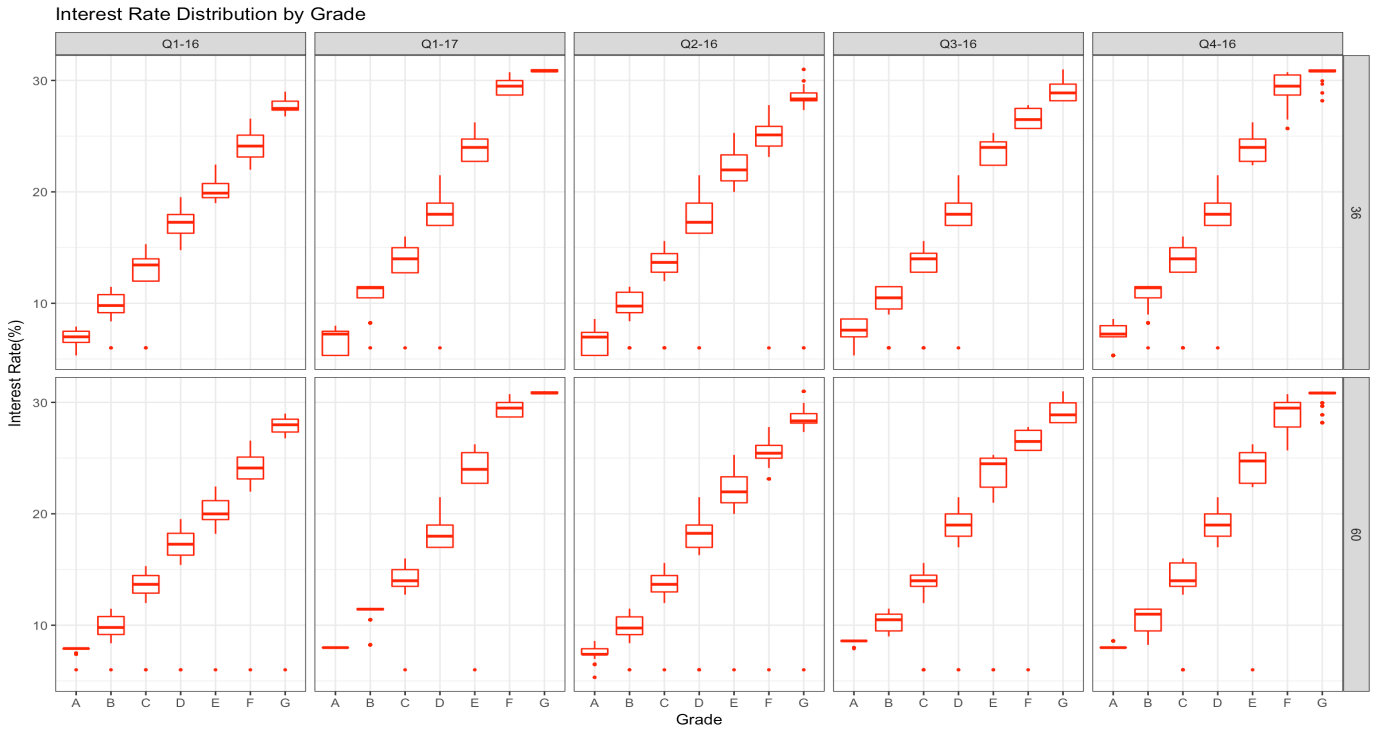


* **Volume of loans by Purpose**

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.



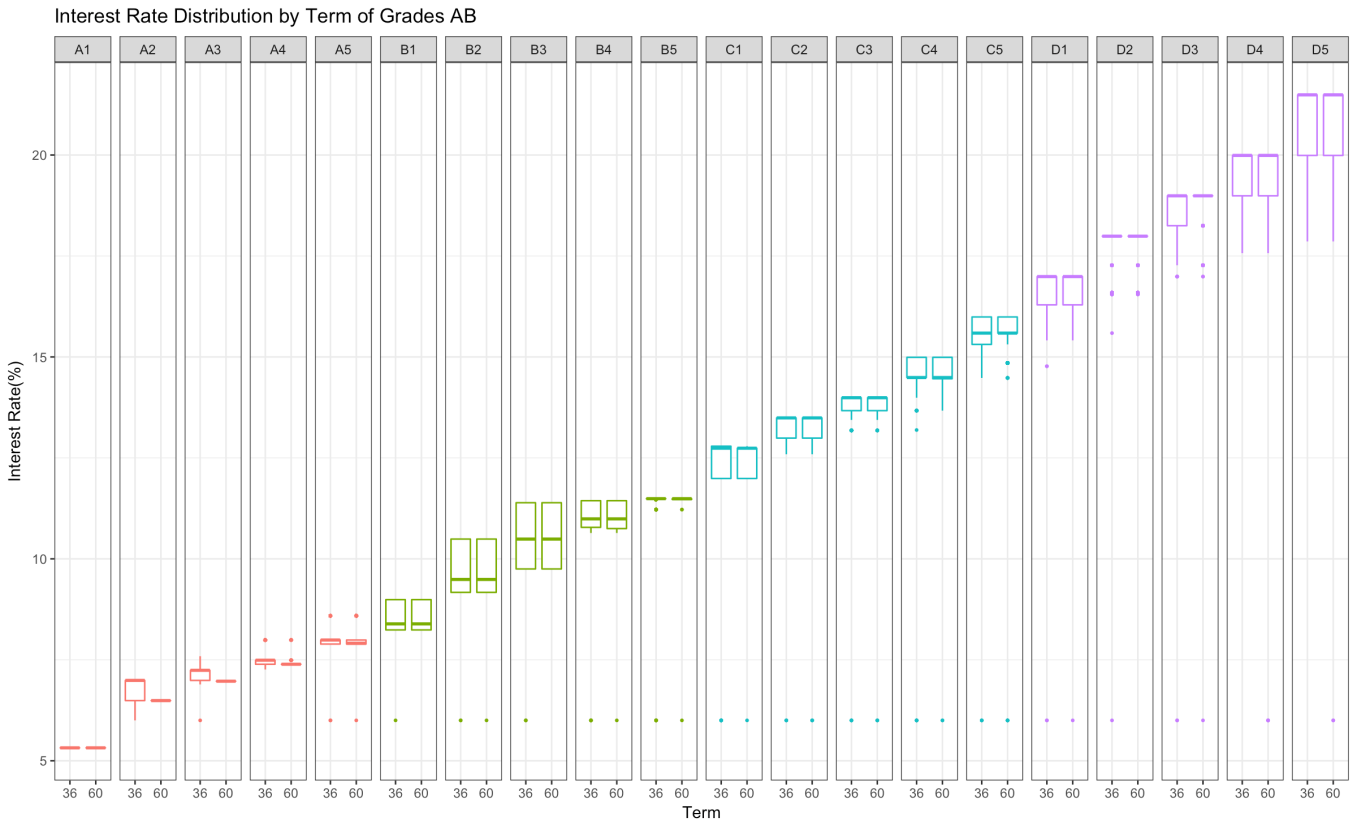
* **Interest Rate Change, Time and Term**



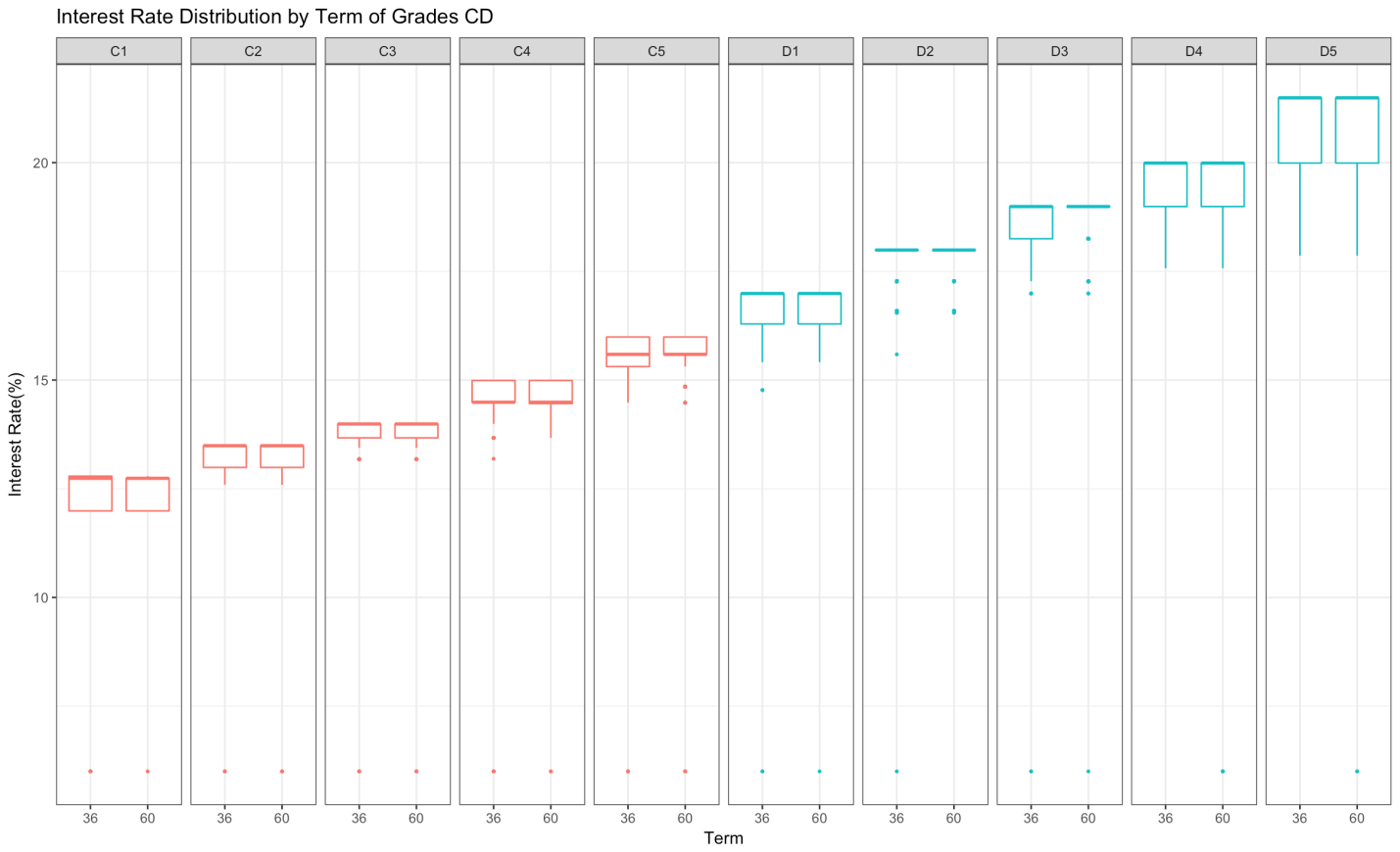
As seen above, outliers can be seen in all the quarters across grades i.e. lower rates of interest are contracted with borrowers even if they fall in grade C-G. This contradicts the policy of lending club which states that higher interest are offered to borrowers falling in higher grades.

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

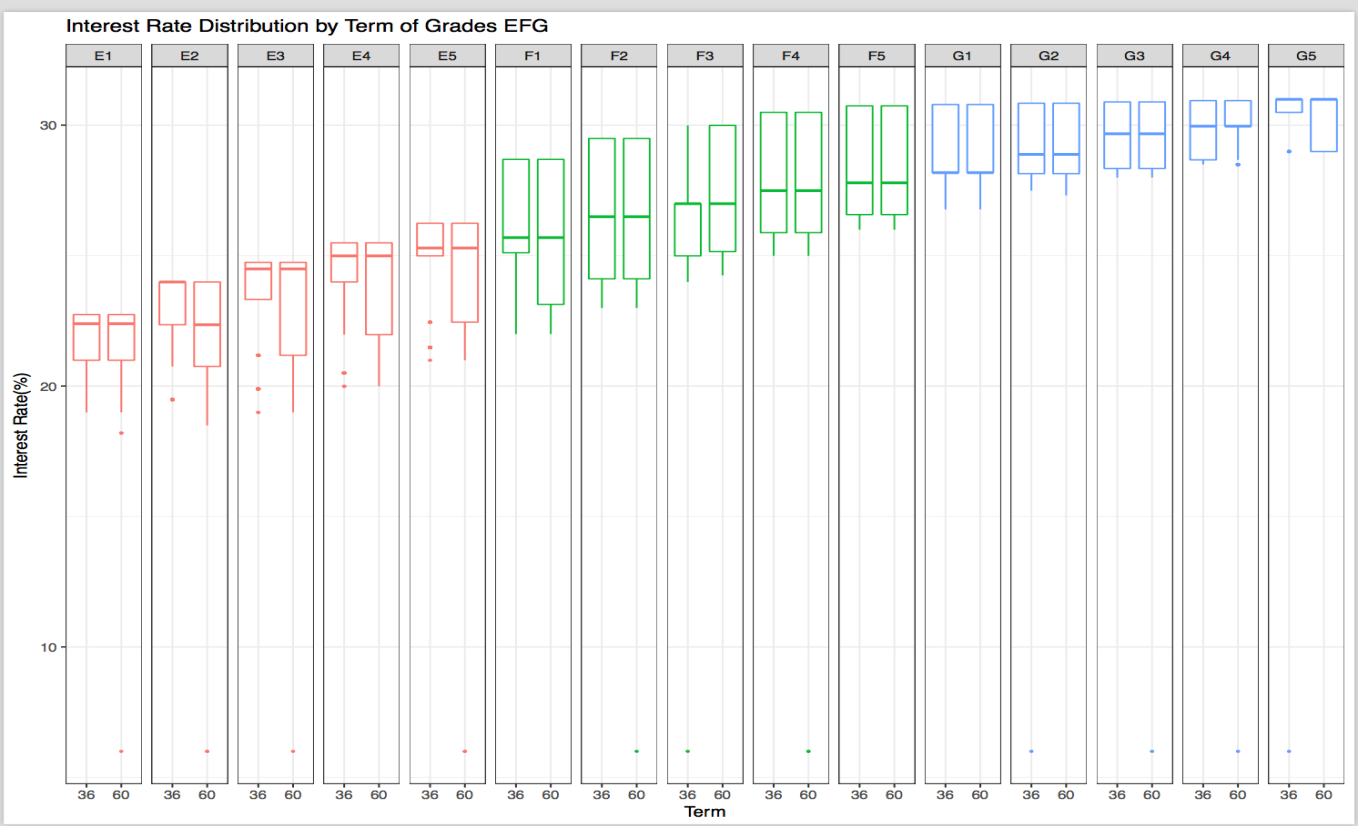
* + **Interest Rate Distribution by Term of Grades A and B**



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

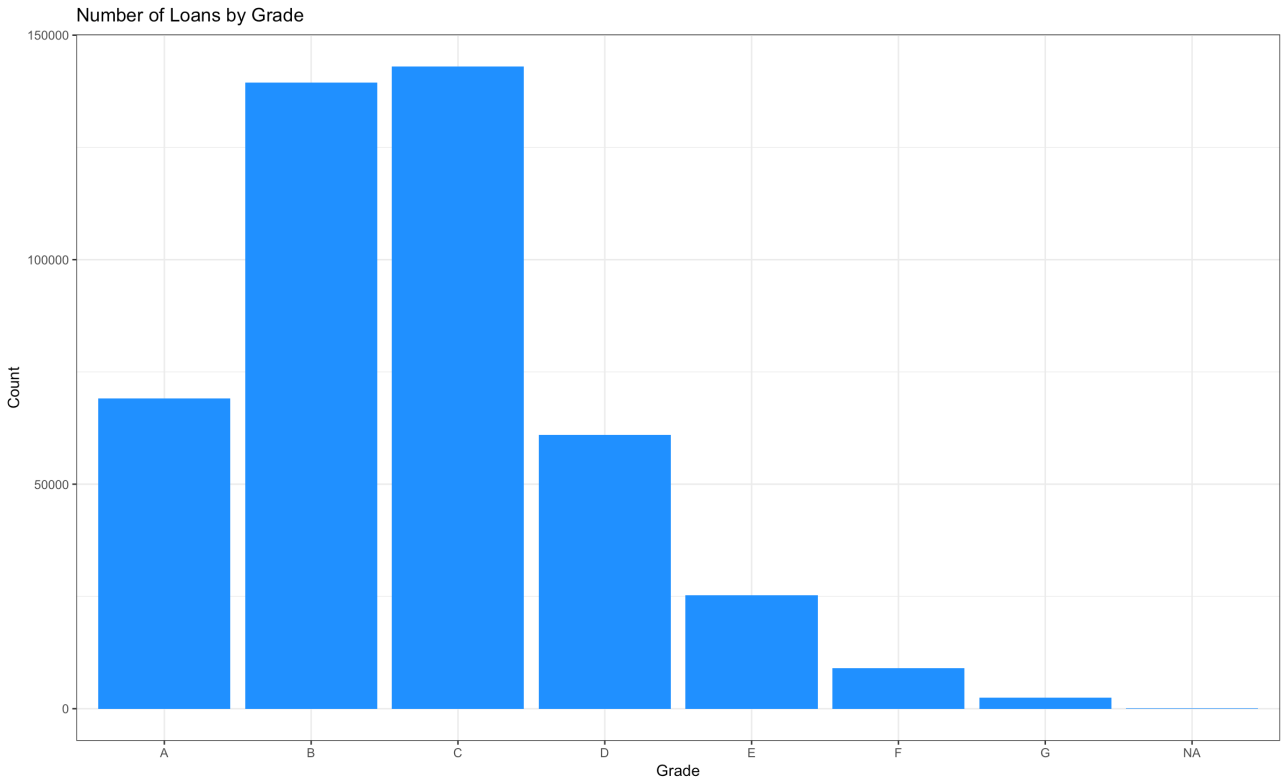


* + **Interest Rate Distribution by Term of Grades E F &G**

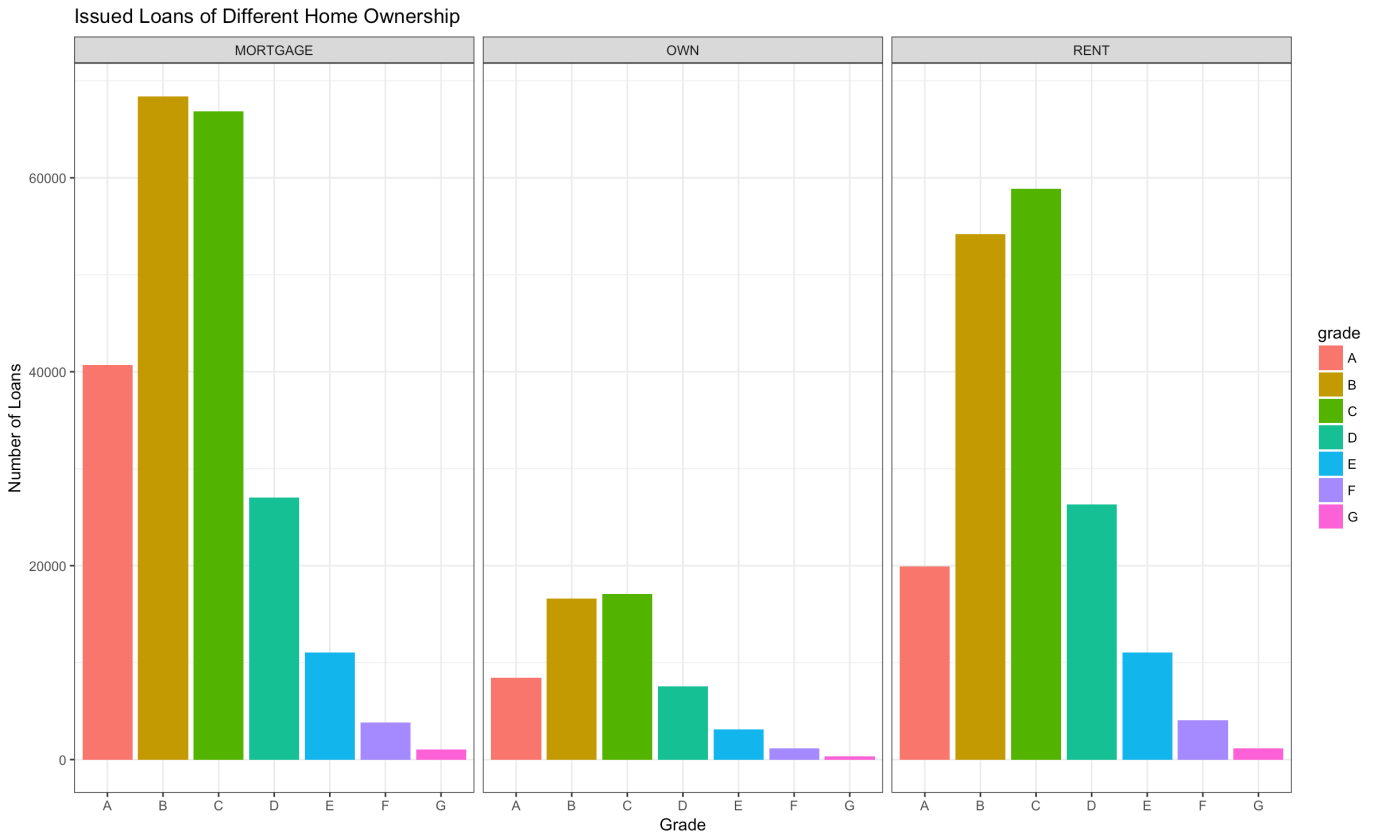
0

## Grade and Default

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

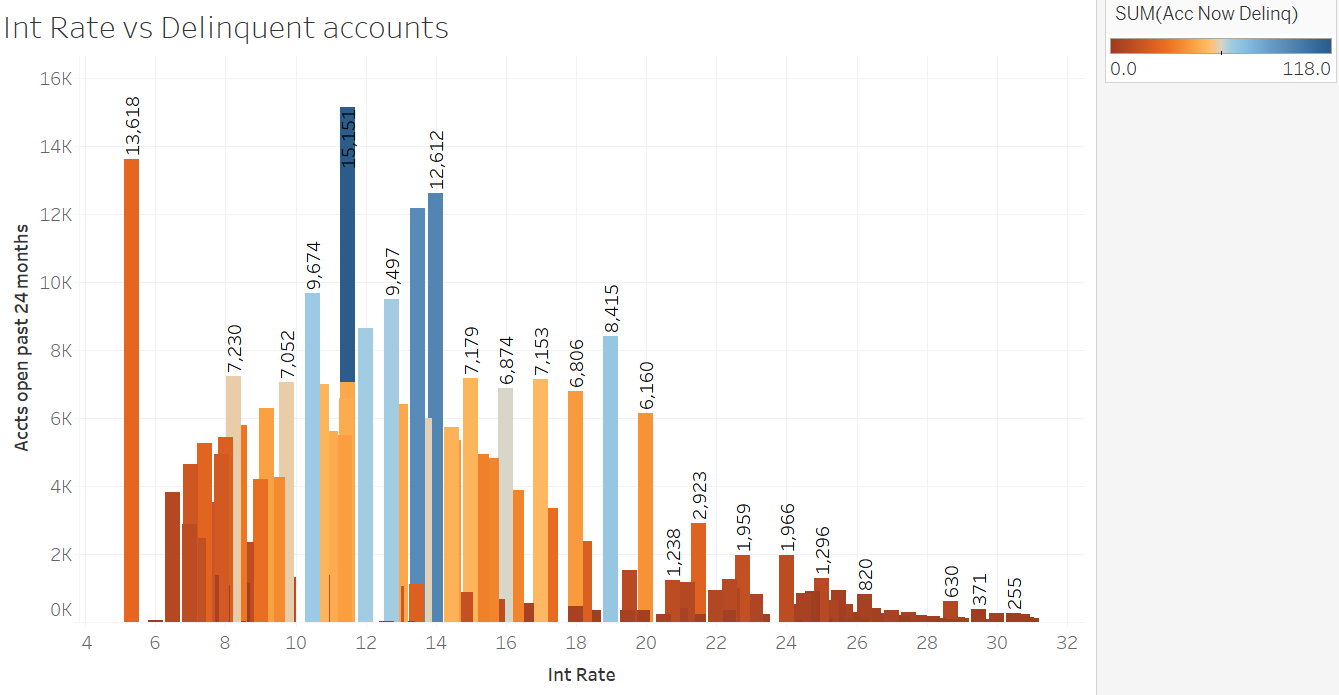


* **Grade vs Home Ownership**



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.



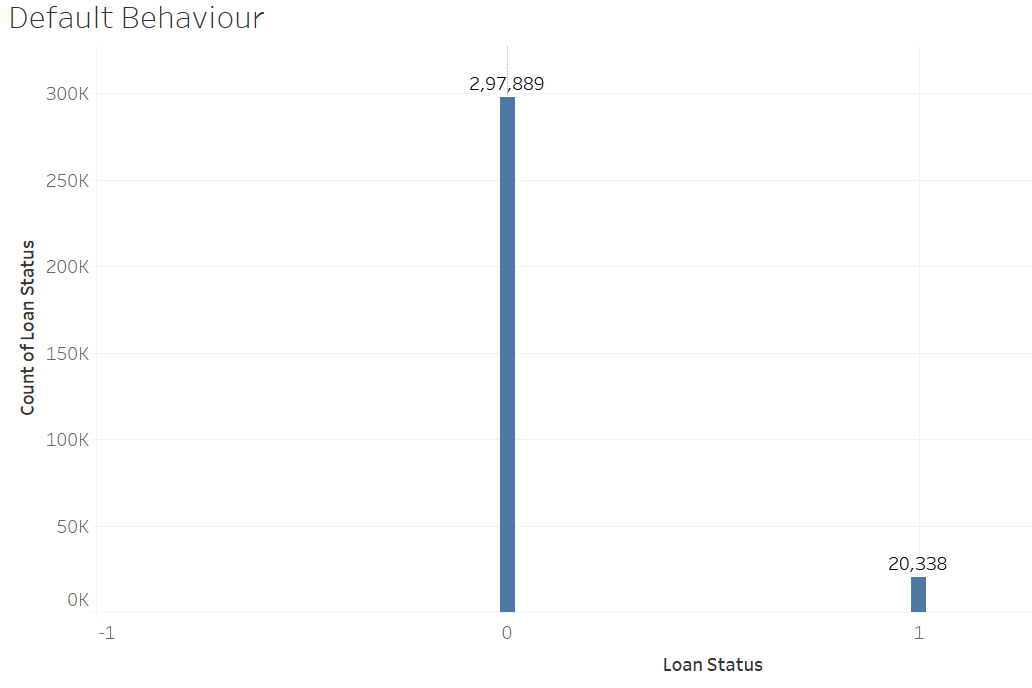
**Chapter 3: Treating imbalanced dataset using SMOTE algorithm**

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.

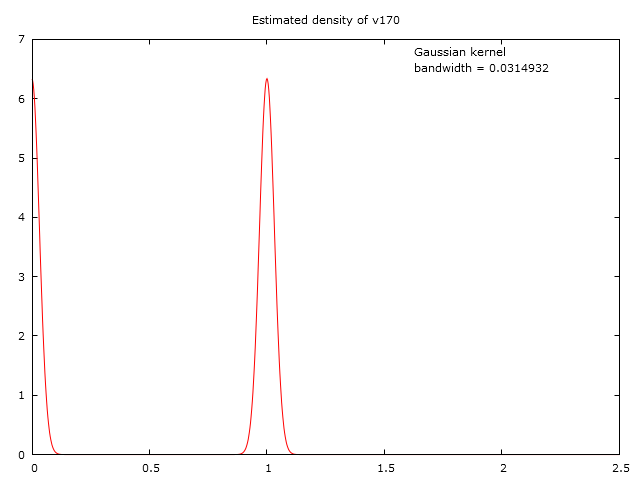
Synthetic Minority Oversampling Technique (SMOTE) is a very popular oversampling method that was proposed to improve random oversampling.

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

Before SMOTE



After SMOTE

****

**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:



where p is the probability of presence of the characteristic of interest.

The logit transformation is defined as the logged odds:

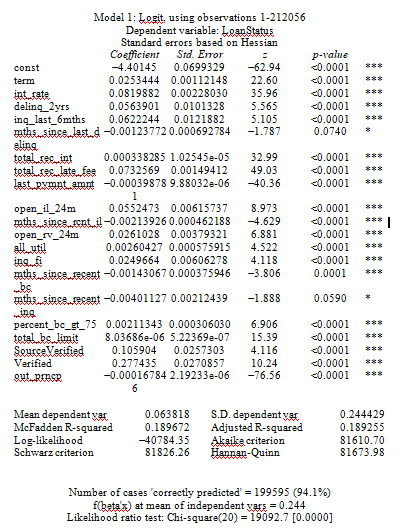


and



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:**

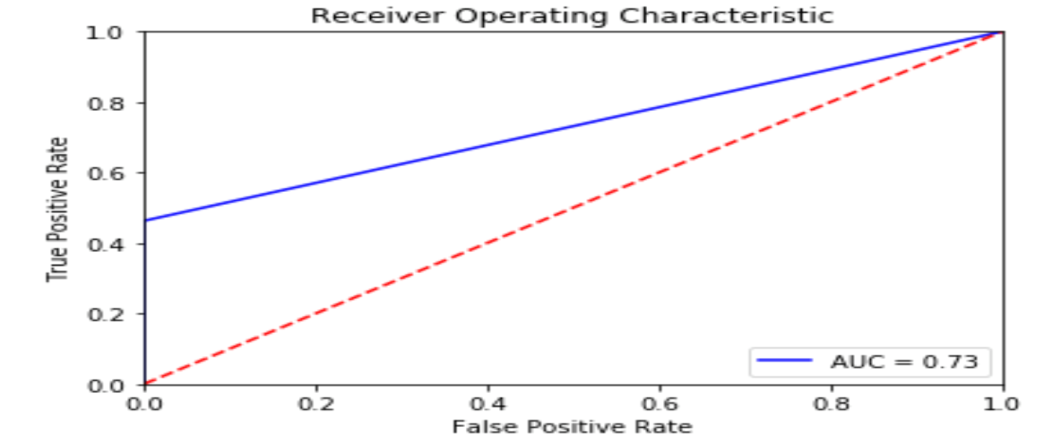
****

**Model Performance**

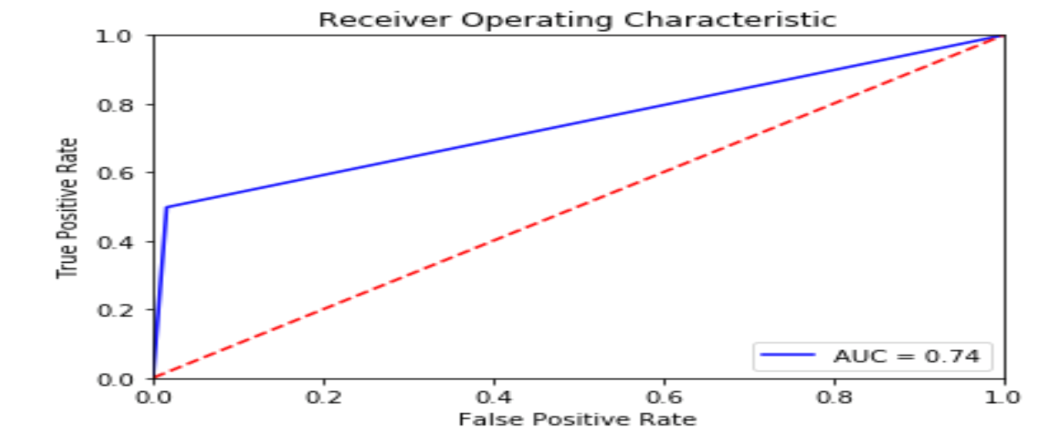
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Performance metrics | On Train Data | | | On Test Data | | |
| Accuracy | 94.1% | | | 94.1% | | |
| Confusion Matrix | C:\Users\Gunjan Ramani\Desktop\unnamed (1).png | | | C:\Users\Gunjan Ramani\Desktop\unnamed.png | | |
| Logarithmic 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](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Instance-based_learning), or [lazy learning](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/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]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-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](https://en.wikipedia.org/wiki/K-means), another popular [machine learning](https://en.wikipedia.org/wiki/Machine_learning) technique.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Performance metrics | On Train Data | | | On Test Data | | |
| Accuracy | 94.68% | | | 92.47% | | |
| Confusion Matrix |  | | |  | | |
| 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](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks, that operate by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_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](https://www.analyticsvidhya.com/blog/2015/02/data-exploration-preparation-model/), 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.



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Performance metrics | On Train Data | | | On Test Data | | |
| Accuracy | 94.37% | | | 94.31% | | |
| Confusion Matrix |  | | |  | | |
| 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:**

1. **Tools :**
2. **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.