|  |  |
| --- | --- |
|  | Delinquency Prediction For A Loan In Service Using Analytics Data. |

**PROJECT REPORT**

***Delinquency prediction for a Loan in service using Analytics Data***

“submitted towards partial fulfilment of the criteria for award of PGP DSE by Great Learning Institute of Management”

**Submitted by:**

|  |  |
| --- | --- |
| **Student Name** | **Student ID number** |
| Shweta Sharma | 3UX27ZQP85 |
| Sharmita Roy | J1QT7Z5WH4 |
| Vikas M | HAVL |
| Shruti Khandelwal | A4WL791BRI |
| Jyothin K Jayan | EWKIR59GIP |

**Batch: DSE \_Bengaluru\_July A 2019**

**Mentor: Ms. Vidhya K**

****

****

**Abstract and Keywords**

**Abstract:**

The economy of the money lending market is increasing with the rate of debentures and returns. As such, for both banking and Fintechs, the ability and importance to see the future and predict the success and failure of a loan applicant to fulfil their obligation are immense. It is here Delinquency Prediction plays a major role in defining how successful a lending institution’s loan origination and subsequent repayment is going to be.

This study is about finding out the causes and predicting the loan delinquency of various financial institutions given in the dataset. The main objective was to identify the delinquency status of the loan for the next month given the delinquency status for the previous 12 months. 19 financial institutions were taken into consideration with a total of 116058 respondents who were sampled for the study. It was found from the study that, customers who had high Debt to Income Ratio seemed to have more delinquency rate, customer’s having low Insurance Percent were also more likely to be delinquent. Also, the duration of the loan taken varies significantly with the interest rate.

Further the findings indicated that with lesser interest rate levied on loan and more time given to the customers, delinquency rate can be reduced. It has therefore been recommended that, the financial institutions can levy less interest rate on customers who have longer term for loan repayment as they would not miss the payment date. Also customers who have lower credit score should not be given huge loans. Salary and the date of salary received is also a prime feature in estimating the delinquent status. More advanced techniques like use of geohash and proper research on the background of person will yield fruitful returns and lesser delinquent payments.

**Keywords:**

Delinquency status, financial institutions, Machine learning, Predictive Modelling, EDA, Oversampling.

**Acknowledgement**

At the outset, we are indebted to our Mentor Ms. Vidhya K for her time, valuable inputs and

guidance. Her experience, support and structured thought process guided us to be on the right

track towards competition of this project and the entire coursework.

We are extremely gifted and fortunate to have Teaching Assistants in DSE program that with

their in-depth knowledge coupled with passion in delivering the subjects in a lucid manner

has helped us a lot. We are thankful to them for their guidance and support.

We also thank all the course faculty of the DSE program for providing us a strong foundation

in various concepts of analytics and machine learning.

Last but not the least, we would like to sincerely thank our respective families for giving the

necessary support, space and time to complete this project.

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

original and authentic.

Shweta Sharma

Sharmita Roy

Vikas M

Shruti Khandelwal

Jyothin K Jayan

Date: 27th November 2019

Place: Bengaluru

**Certificate of Completion**

I hereby certify that the project titled “Delinquency Prediction for a Loan in Service using Analytics Data” was undertaken and completed under my guidance and supervision by Shweta Sharma, Sharmita Roy, Vikas M, Shruti Khandelwal and Jyothin K Jayan, students of July A 2019 Batch of the Post Graduate Program in Data Science & Engineering, Bengaluru.

Ms. Vidhya K

Date: 27th November, 2019

Table of Contents

[Introduction 8](#_Toc25844027)

[Industry Review – Current practices, Background Research 8](#_Toc25844028)

[Current Practices in the loan prediction: 9](#_Toc25844029)

[Delinquency prediction and machine learning: 10](#_Toc25844030)

[Literature Survey - Publications, Application, past and undergoing research: 10](#_Toc25844031)

[Applications of Loan Delinquency: 11](#_Toc25844032)

[Past Research: 12](#_Toc25844033)

[Current Research: 12](#_Toc25844034)

[Data Dictionary and Preprocessing Data Analysis: 13](#_Toc25844035)

[About Features: 14](#_Toc25844036)

[Irrelevant Columns: 16](#_Toc25844037)

[Alternate sources of data that can supplement the core dataset 16](#_Toc25844038)

[Project Statement 16](#_Toc25844039)

[Complexity Involved 16](#_Toc25844040)

[Project Outcome – Commercial 17](#_Toc25844041)

[Project Justification 17](#_Toc25844042)

[Overview of Processes: 17](#_Toc25844043)

[Data Exploration (EDA): 17](#_Toc25844044)

[Uni-variate Analysis: 18](#_Toc25844045)

[Distribution of Variables: 21](#_Toc25844046)

[Bi-variate Analysis: 22](#_Toc25844047)

[Correlation of the dataset: 26](#_Toc25844048)

[Scaling the data: 26](#_Toc25844049)

[Imbalance Data Detection: 27](#_Toc25844050)

[Treatment of Imbalance Data: 27](#_Toc25844051)

[Feature Engineering and Model Selection 28](#_Toc25844052)

[Feature Engineering 28](#_Toc25844053)

[Benefits of Feature Selection: 28](#_Toc25844054)

[Feature Selection Methods: 28](#_Toc25844055)

[Assumptions 29](#_Toc25844056)

[Logistic Regression 29](#_Toc25844057)

[Model performance measures used for evaluating models: 29](#_Toc25844058)

[Model Selection 31](#_Toc25844059)

[Classification Report 31](#_Toc25844060)

[Results obtained with oversampling: 32](#_Toc25844061)

[Feature Importance for Tree based Algorithms: 33](#_Toc25844062)

[Final Model: 35](#_Toc25844063)

[Random Forest: 35](#_Toc25844064)

[Conclusions: 36](#_Toc25844065)

[Recommendations: 37](#_Toc25844066)

[References and Bibliography: 38](#_Toc25844067)

# Introduction

## Industry Review – Current practices, Background Research

The origination of loan and interest creates a humongous share of business for majority of banks. As such, for both banking and private lending institutions, the ability and importance to see the future and forecast the success and failure of a loan applicant to fulfil their obligation are immense. It is here Delinquency Prediction plays a major role in defining how successful a bank’s loan origination and subsequent repayment is going to be.

However, there is more to lending than improving the certainty of successful loan origination through regression analysis. Many a time, borrowers who might seem to make for perfect candidates for loan origination might show erratic payment and financial behavior, once their loan is approved. This is something that the underwriters might not be able to predict at the time of loan origination. But this behavior largely increases the risk involved in lending for banks & other alternative lending fin-techs, as it reduces, or rather, jeopardizes the chances of full principle repayment along with interest. Delinquency prediction helps the lenders see the risk by observing and studying a large set of consumers and their financial behaviors using statistical models that help in removing biases and errors to give a score, close to perfection.

**One of the precocious methods used to predict the loan in 1990s to 2008 was the use of questionnaire in the banks:**

Before digitalization and technology-oriented techniques, Banks used a set of questions which mostly revolved around the customer’s collateral, his wealth management acumen and his psychology to access his behavior and his intent towards repayment. The success of predicting the intention of the customers was quite falsified at most of the cases due to less diligence of the customers stating in the questionnaire.

Hence, more adverse methods came into picture in late 2008 using various statistical and mathematical modelling. Banks and private lending institutions have started using a number of new age tools and technique to improve their lending scores while increasing and retaining their customer base and expanding along with the dynamic markets.

## Current Practices in the loan prediction:

1. Regression Analysis:

A multivariate regression analysis was used to help in considering a number of variables in a fair lending or loan origination decision. In addition to that, [comparing FICO scores and other](https://www.lendfoundry.com/use-of-third-party-data-sources-in-personal-lending/) [factors readily available from third party sources, underwriters used regression analysis](https://www.lendfoundry.com/use-of-third-party-data-sources-in-personal-lending/) to:

* Explain credit sanctioning decisions and find out if prohibitive factoring was used to

ascertain the success rate of a loan and its pricing.

* Find creditworthy applicants from the protected class section.
* Find out if the right lending rate, APR was applied to applications and the percentage variation from the tool’s expected pricing or lending rate.
* Find out why certain applications were denied or were priced higher or lower than the tool’s predictions.

1. Use of third-party data sources in personal lending:

The personal lending market was one of the biggest markets that held a wealth of promise for bankers and lenders. This market seemed to be largely un-served and lacked the competitive edge which was usually seen in the large business lending. Banks typically wanted to lower the risk component and also keep their underwriting costs low to enjoy the benefits of lending, the traditional way. This is why banks preferred lending to larger business with better credit scores and whose underwriting costs are comfortably taken care of, given the size of the loans large businesses apply for. However, the same cannot be said for the personal lending and SMB loan market.

The otherwise overlooked personal lending market, which moved further back in the shadows post the subprime crisis of 2008, has stood neglected for a long time, with little or no promise of showing any surge. But that was until fin-tech started moving into the SMB and personal lending market. With their innovative solutions and algorithms that churn numbers in matter of seconds, access and use of big data to conduct predictive analysis, fin-tech has taken the personal lending industry by storm.

Current scenario: Outstanding balances increased about 18% in the first quarter of 2017-18 to a staggering approx. amount to [$120 Bn](http://fortune.com/2018/07/03/us-personal-loan-debt-2018/).

## Delinquency prediction and machine learning:

As mentioned above, loan behavior prediction can be a great tool in ascertaining renewals of credit lines and loans, as well as knowing if the customer can make good on their repayment schedule. The simplest way to do so is by using machine learning to assimilate all the data and apply them to statistical models and finding out scores that ascertain the probability of delinquency.

For each loan application, the bank receives plenty of information. For example, loans, transaction records, and credit cards. While these make for basic details that are required to be filled for any loan application, there are few more variables that are required to successfully study and ascertain the success rate of the loan. Some of these variables are calculated by the lender on the basis of the information shared by the borrower, supported by relevant documentation.

## Literature Survey - Publications, Application, past and undergoing research:

In recent times, a lot of focus has been directed towards understanding the challenges confronting the banking sector. One key aspect of this challenge has been analyzing the factors affecting the non-performing loans (NPLs) of banks.

The Concepts of Loan Delinquency and Loan Default are delinquent when a payment is late (CGAP, 1999). A delinquent loan becomes a defaulted loan when the chance of recovery becomes minimal. Delinquency is measured because it indicates an increased risk of loss, warnings of operational problems, and may help to predict how much of the portfolio will eventually be lost because it never gets repaid. There are three broad types of delinquency indicators: collection rates which measures amounts actually paid against amounts that have fallen due; arrears rates measures overdue amounts against total loan amounts; and portfolio at risk rates which measures the outstanding balance of loans that are not being paid on time against the outstanding balance of total loans (CGAP, 1999). Default occurs when a debtor has not met his or her legal obligations according to the debt contract. For example, a debtor has not made a scheduled payment, or has violated a loan covenant (condition) of the debt contract (Ameyaw-Amankwah, 2011). A default is the failure to pay back a loan. Default may occur if the debtor is either unwilling or unable to pay their debt. A loan default occurs when the borrower does not make required payments or in some other way does not comply with the terms of a loan. (Murray, 2011). Moreover, Pearson and Greeff (2006) defined default as a risk threshold that describes the point in the borrower’s repayment history where he or she missed at least three instalments within a 24-month period. This represents a point in time and indicator of behaviour, wherein there is a demonstrable increase in the risk that the borrower eventually will truly default, by ceasing all repayments. The definition is consistent with international standards, and was necessary because consistent analysis required a common definition. This definition does not mean that the borrower had entirely stopped paying the loan and therefore been referred to collection or legal processes; or from an accounting perspective that the loan had been classified as bad or doubtful, or actually written-off. Loan default can be defined as the inability of a borrower to fulfil his or her loan obligation as at when due (Balogun and Alimi, 1990).

Some of the publications that are taken into consideration are:

[https://cafral.org.in/sfControl/content/Speech/621201790033PMLoan\_Delinquency\_in\_Bank](https://cafral.org.in/sfControl/content/Speech/621201790033PMLoan_Delinquency_in_Banking_Systems.pdf) [ing\_Systems.pdf](https://cafral.org.in/sfControl/content/Speech/621201790033PMLoan_Delinquency_in_Banking_Systems.pdf) - Loan Delinquency in Banking Systems

<http://www.pbr.co.in/2017/2017_month/August/14.pdf>- An Empirical Analysis of Factors affecting the Loan Repayment

<https://pdfs.semanticscholar.org/1130/aa51c0d4ce6ba47c5d0d2df6ebd230d02990.pdf>- Causes and Control of Loan Default/Delinquency in Microfinance Institutions in Ghana

[https://www.lendfoundry.com/regression-analysis-as-a-fintech-tool-for-predicting-successful-](https://www.lendfoundry.com/regression-analysis-as-a-fintech-tool-for-predicting-successful-loan-applications/) [loan-applications/](https://www.lendfoundry.com/regression-analysis-as-a-fintech-tool-for-predicting-successful-loan-applications/) - Regression Analysis as a successful fin-tech tool.

## Applications of Loan Delinquency:

In today's financial services marketplace, a financial institution exists to provide a wide variety of deposit, lending and investment products to individuals, businesses or both. While some financial institutions focus on providing services and accounts for the general public, others are more likely to serve only certain consumers with more specialized offerings.

The types of financial institutions that provide loans are:

* Retail and Commercial Banks:

Traditionally, [retail banks](https://www.investopedia.com/terms/r/retailbanking.asp) offered products to individual consumers while commercial banks worked directly with businesses. Currently, the majority of large banks offer [deposit accounts,](https://www.investopedia.com/personal-finance/complete-guide-money-market-deposit-accounts/) lending and limited financial advice to both demographics. Products offered at retail and commercial banks include checking and savings accounts, certificates of deposit (CDs), personal and mortgage loans, credit cards, and business banking accounts.

* Internet Banks:

A newer entrant to the financial institution market is [internet banks](https://www.investopedia.com/terms/o/onlinebanking.asp), which work similarly to retail banks. Internet banks offer the same products and services as conventional banks, but they do so through online platforms instead of brick and mortar locations.

* Savings and Loan Associations:

Financial institutions that are mutually held and provide no more than 20% of total lending to businesses fall under the category of [savings and loan associations](https://www.investopedia.com/terms/f/federal-savings-and-loan.asp). Individual consumers use savings and loan associations for deposit accounts, personal loans, and mortgage lending.

* Mortgage Companies:

Financial institutions that originate or fund mortgage loans are [mortgage companies](https://www.investopedia.com/terms/m/mortgage-company.asp). While most mortgage companies serve the individual consumer market, some specialize in lending options for commercial real estate only.

## Past Research:

From the publishing paper found in “Evaluating the likelihood of default on delinquent loans” where the purpose of the research was to examine the characteristics of borrowers and loan contracts related to the probable resolution of past due accounts. It was found out that application information providing ex ante clues to the future value of collateral, such as location and property age in the case of mortgaged assets are consistently good indicators of the subsequent resolution of a delinquency. Credit managers should examine these items as soon as loan becomes delinquent. They should also attempt to identify more future directed items when granting credit initially.

It was also found that some findings from the research were way beyond the collateralized lending. The probability of default was found almost twice as high for the first-time delinquents. Hence, it was concluded that the first-time delinquencies should be attended quickly. Further analysis was needed with different methodologies.

## Current Research:

With the current scenario, various machine learning models are being built in relation with statistical analysis like regression analysis. The upcoming research is going on how more variables can simplify and give more valuable insights. The ongoing research has been started for following new techniques:

* Use of Geohash (a public domain code which encodes geographic location into short strings of letters and digits) in finding out the location where there are minimum and maximum delinquencies. This can help to infer various new strategies in forming different schemes for financial institutions across the locations depending on the delinquency rate.
* According to Behavioural Analysis of the customers taking loan, it is intending to find out the intention of the borrower and predict the approximate duration of the loan repayment and also the probability of the borrower to fall under the delinquency status with the help of new era technologies like iris scan, pulse rate count and facial expressions by an app while interviewing the borrower.
* Filtering clients in accord to industry wherein the financial institutions are trying to find the industries whose employers fall under lower delinquency rate and higher delinquency rate. This is because, the financial institutions can easily screen their customers based on loyalty of the repayments to be made by the customers and blacklist the rest.

# Data Dictionary and Preprocessing Data Analysis:

|  |  |  |  |
| --- | --- | --- | --- |
| Sr.No | Variable Name | Category | Info |
| 1. | loan\_id | Numerical | 116058 non-null int64 |
| 2. | Source | Categorical | 116058 non-null object |
| 3. | financial\_institution | Categorical | 116058 non-null object |
| 4. | interest\_rate | Numerical | 116058 non-null float64 |
| 5. | unpaid\_principal\_bal | Numerical | 116058 non-null int64 |
| 6. | loan\_term | Numerical | 116058 non-null int64 |
| 7. | origination\_date | Categorical | 116058 non-null object |
| 8. | first\_payment\_date | Categorical | 116058 non-null object |
| 9. | loan\_to\_value | Numerical | 116058 non-null int64 |
| 10. | number\_of\_borrowers | Numerical | 116058 non-null float64 |
| 11. | debt\_to\_income\_ratio | Numerical | 116058 non-null float64 |
| 12. | borrower\_credit\_score | Numerical | 116058 non-null float64 |
| 13. | loan\_purpose | Categorical | 116058 non-null object |
| 14. | insurance\_percent | Numerical | 116058 non-null float64 |
| 15. | co-borrower\_credit\_score | Numerical | 116058 non-null float64 |
| 16. | insurance\_type | Numerical | 116058 non-null float64 |
| 17. | m1 | Numerical | 116058 non-null int64 |
| 18. | m2 | Numerical | 116058 non-null int64 |
| 19. | m3 | Numerical | 116058 non-null int64 |
| 20. | m4 | Numerical | 116058 non-null int64 |
| 21. | m5 | Numerical | 116058 non-null int64 |
| 22. | m6 | Numerical | 116058 non-null int64 |
| 23. | m7 | Numerical | 116058 non-null int64 |
| 24. | m8 | Numerical | 116058 non-null int64 |
| 25. | m9 | Numerical | 116058 non-null int64 |
| 26. | m10 | Numerical | 116058 non-null int64 |
| 27. | m11 | Numerical | 116058 non-null int64 |
| 28. | m12 | Numerical | 116058 non-null int64 |
| 29. | m13 | Numerical | 116058 non-null int64 |

* + Numerical variables – 23 (excluding target variable)

Categorical variable - 5

* + There are no null values in the dataset
  + Also, there are no redundant columns in the dataset.

## About Features:

* + - source : Loan Origination Channel
    - financial\_institution: Name of bank issued loan
    - unpaid\_principal\_bal: Amount yet to be paid
    - loan\_term : term allotted to the customer to repay loan(in days)
    - loan\_to\_value : ratio describing size of loan compared to value of property securing the loan.
    - number\_of\_borrowers: 1- A borrower taking full responsibility to pay the loan amount.

2- Co-borrower takes loan on the risk that he may have to pay

the loan if the borrower fails.

* + - debt\_to\_income\_ratio: percentage of person’s monthly income that goes towards paying debts.
    - borrower\_credit\_score and co-borrower\_credit\_score : credit score is statistical number that evaluates a consumer’s worthiness i.e. probability that the person will pay his or her debts.
    - insurance\_percent : Loan amount percentage covered by insurance.
    - Insurance\_type : 0- Premiun paid by borrower

1 -Premiun paid by lender

* + - m1 to m12 : Month-wise loan performance of an individual.
    - m13 : Target variable [ 0: Non-Delinquent , 1: Delinquent].

## Irrelevant Columns:

There are some variables that do not contribute to the target variable. Hence, we can drop them.

|  |  |
| --- | --- |
| loan\_id | We dropped this feature as it did not contributed to the dependant variable. |
| Source | Contains ‘X’, ‘Y’, ‘Z’ (no proper information given).So, we dropped it. |
| financial\_institution | since financial\_institution is irrelevant to the data set, we dropped that column as well. |

## Alternate sources of data that can supplement the core dataset

|  |  |
| --- | --- |
| Salary | Salary is a good predictor for classification model as in case of delinquent and non-delinquent. |
| Salary pay date | Since, every customer has different pay dates, financial institutions can imply different strategies to reduce the delinquency rates. |
| Number of dependents | Financial institutions can concentrate on customers having a greater number of dependents while issuing loans. |

## Project Statement

Loan Delinquency Prediction is one of the most critical and crucial problem faced by financial institutions and organizations as it has a noteworthy effect on the profitability of these institutions. In recent years, there is a tremendous increase in the volume of non – performing loans which results in a jeopardizing effect on the growth of these institutions. Therefore, to maintain a healthy portfolio, the banks put stringent monitoring and evaluation measures in place to ensure timely repayment of loans by borrowers. Despite these measures, a major proportion of loans become delinquent. Delinquency occurs when a borrower misses a payment against his/her loan.

Given the information like mortgage details, borrowers related details and payment details, *our objective is to identify the delinquency status of loans for the next month given the delinquency status for the previous 12 months (in number of months).*

## Complexity Involved

Not many complications are involved with respect to this dataset, but more insights could have drawn in if Salary date and Salary range would have been there. In that way we could have more efficiently predicted the reason why the borrower is missing the payment date and his intention also.

## Project Outcome – Commercial

The project outcome is entirely commercial. The prediction will help various financial

institutions to leverage and efficiently build models to predict the risk of giving out

loans. The loan industry gains huge profit but at the same time, too many wrong

decisions would land a financial institution at stake. Hence, proper modelling and good

methodologies can help the financial institutions to be at a safer side.

## Project Justification

● This is a fictional data set containing details of customers took loan from banks.;

● The Data set is about Loan Delinquency Prediction using Analytics Data.

● This is a classification problem. The dependant variable is ​Delinquency

(0: Non-Delinquent, 1: Delinquent).

● We can use Classification model algorithms like Logistic Regression, K-NN, Decision Tree,

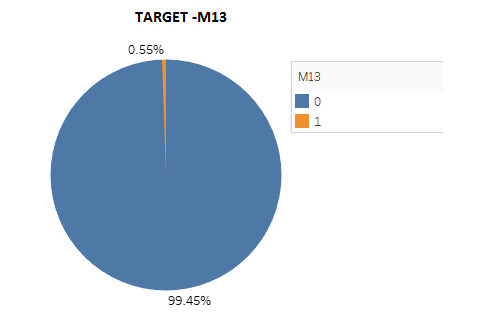
Random Forest, etc.

# Overview of Processes:

## Data Exploration (EDA):

The purpose of exploratory data analysis is two-fold:

* To understand the data in terms of delinquency rate and the features contributing more on predicting that.
* Get insights on various features.
  + 1. **Understand data distribution:**



m13 – target variable

0: Non- Delinquent

1: Delinquent

* This indicates that number of observations belonging to class 1 (Delinquent) is significantly lower than those belonging to class 0 (Non- Delinquent).

The conventional accuracy of the predictive models is not a relevant measure of model

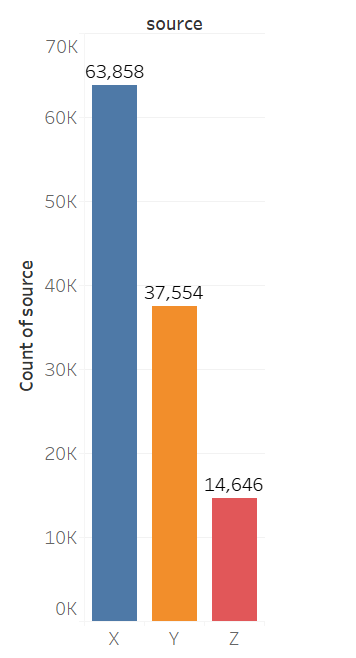
performance because machine learning algorithms are usually designed to improve accuracy by

reducing the error. Thus, they do not take into account the class distribution / proportion or

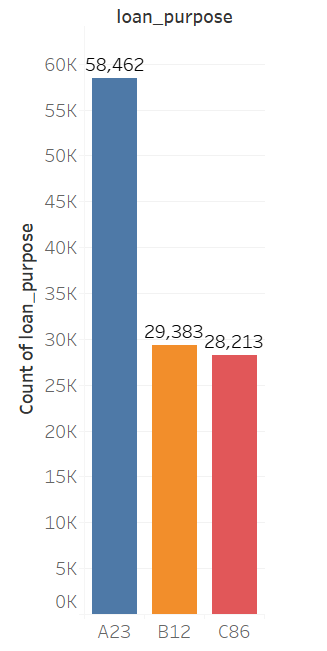
balance of classes.

Hence, we will consider other model performance measures to evaluate a model, keeping in mind the class imbalance problem.

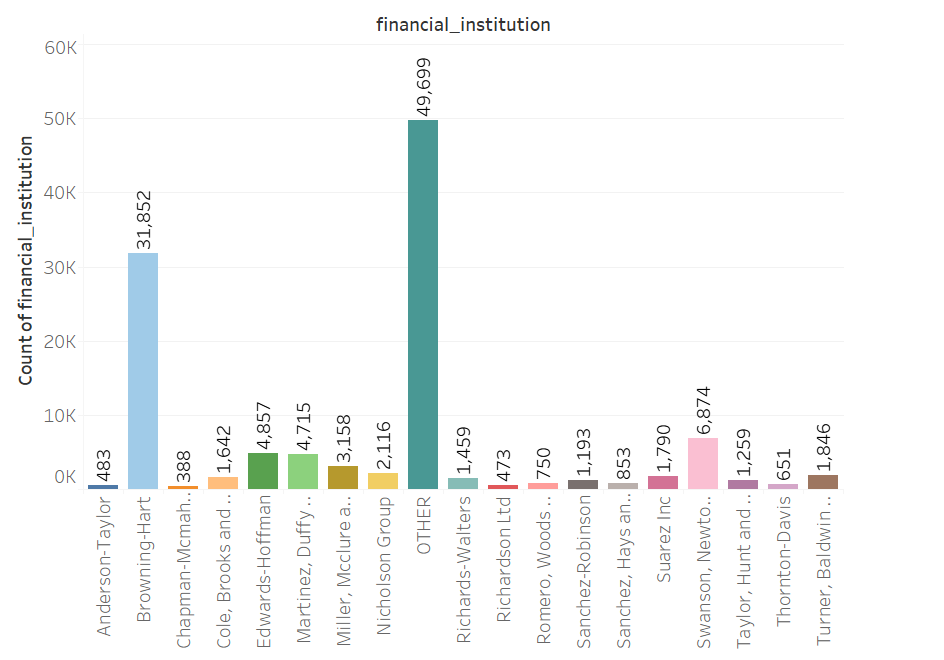
## Uni-variate Analysis:

* + - 1. Count plot of ‘source’ variable:
* Here, the count of X source is the highest among X, Y, Z.

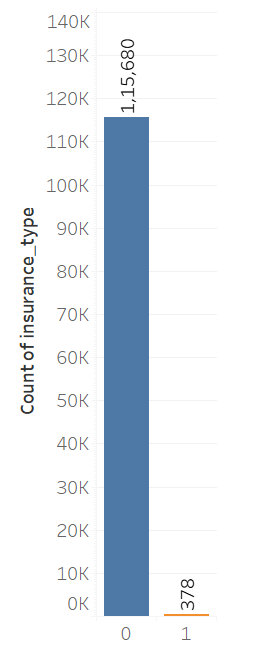
* + - 1. Count plot of *loan\_purpose* variable:



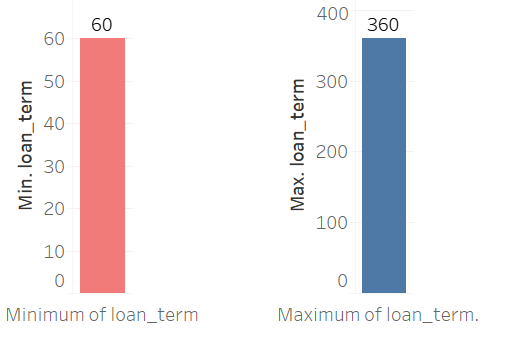
* The highest no of loans is given for A23 category.
  + - 1. *financial institutions* variable:



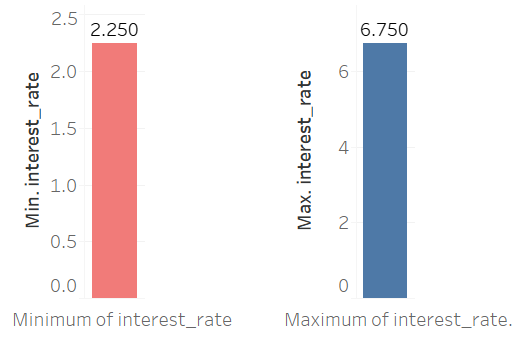
* ‘Others’ followed by ‘Browning Hart’ has the highest no of loans disbursed.
  + - 1. Count plot of *insurance\_type* variable:



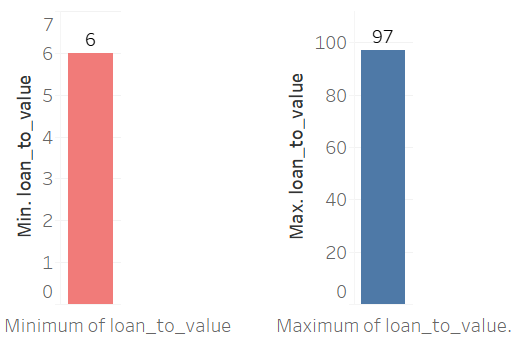
* Majority of the loan have no insurance.
  + - 1. *Loan\_term*:
* Minimum loan term is 60 days and the maximum term is 360 days.

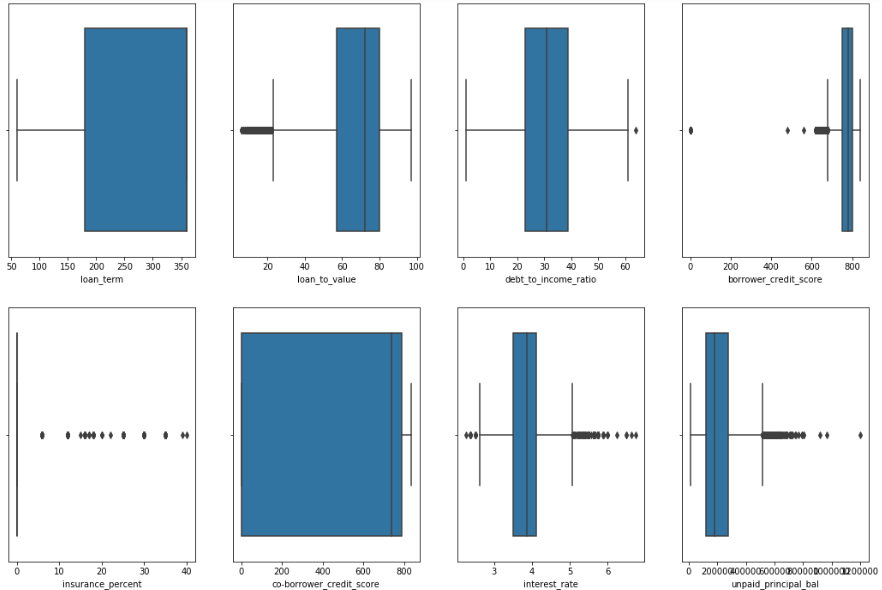


* + - 1. *Interest \_rate*:



* Minimum interest is 2.25 and maximum rate is 6.75
  + - 1. *Loan\_to\_value*:

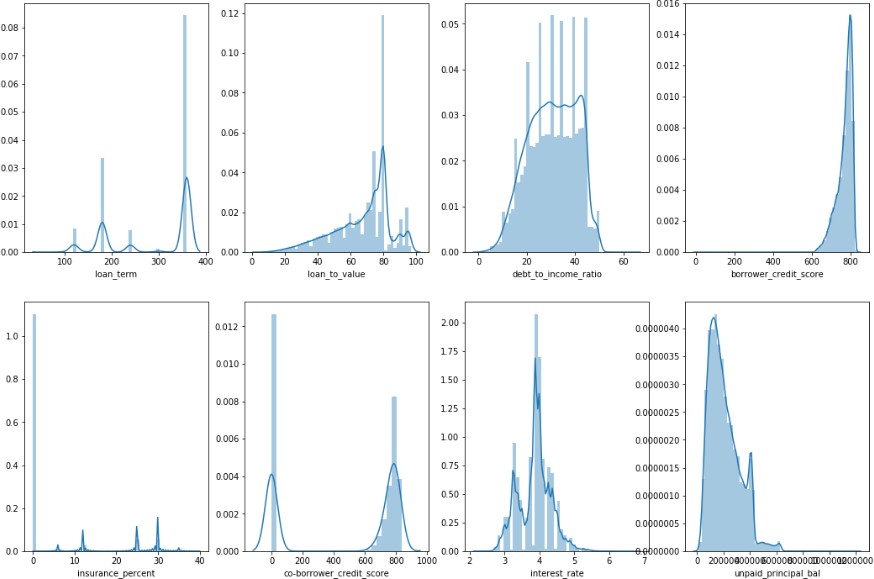


* A loan with a lower LTV ratio is less of a risk for the lender. The min LTV is 6 and the max is 97.
  + 1. **Outlier Detection through Boxplot:**
* Except Loan term and Co borrower credit score there are extreme values and outliers in remaining features.
* To reduce the extreme values, we are using Standard Scalar.
* Majority of the loan borrowers has 0 interest rate.

Box plot is drawn for independent features against Target variable and outlier had been detected.

Since, the outliers are legitimate, we have decided to retain them in dataset.

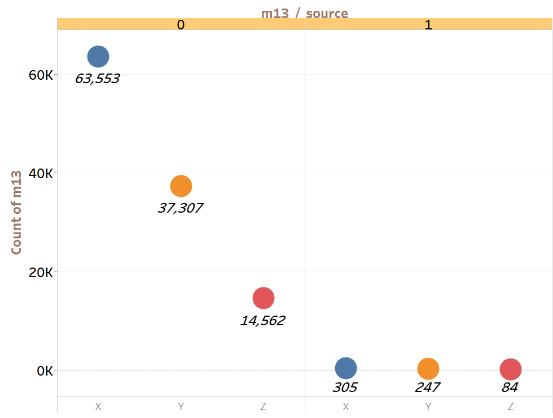
## Distribution of Variables:



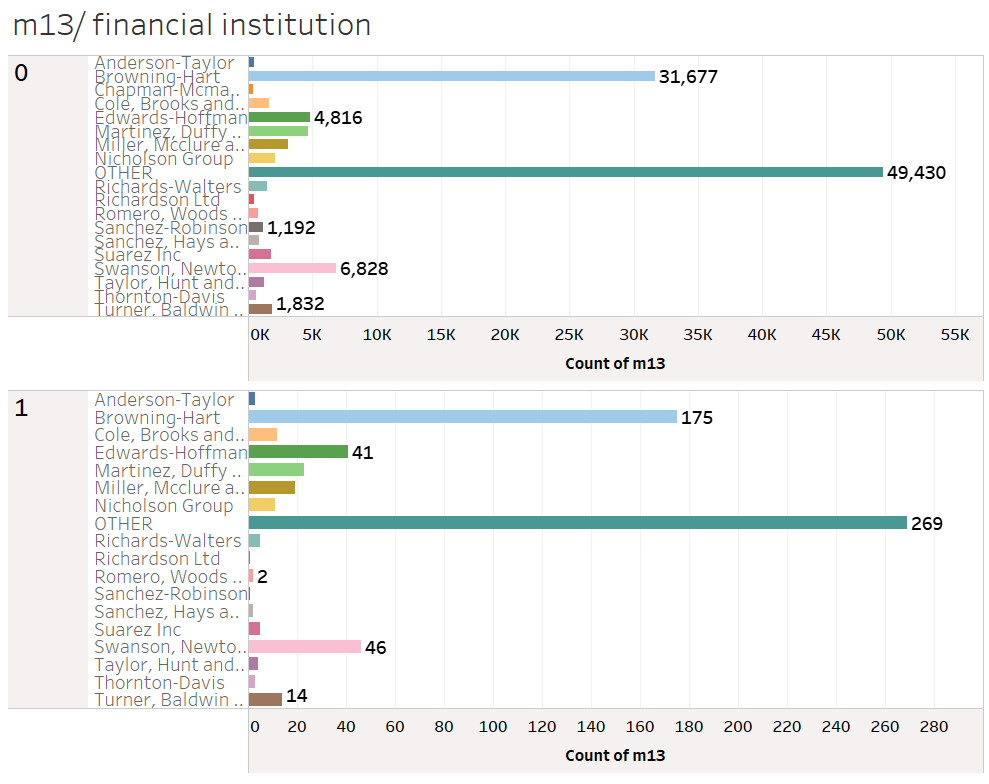
* Interest Rate is significantly normally distributed.
* Loan term is normally significantly distributed.
* Loan to Value is left skewed.
* Debt to Income is significantly normally distributed.
* Borrower credit score is highly left skewed.
* Insurance percent is not normally distributed.
* Co Borrower is not normally distributed.
* Un-paid Principal balance is highly right skewed.

## Bi-variate Analysis:

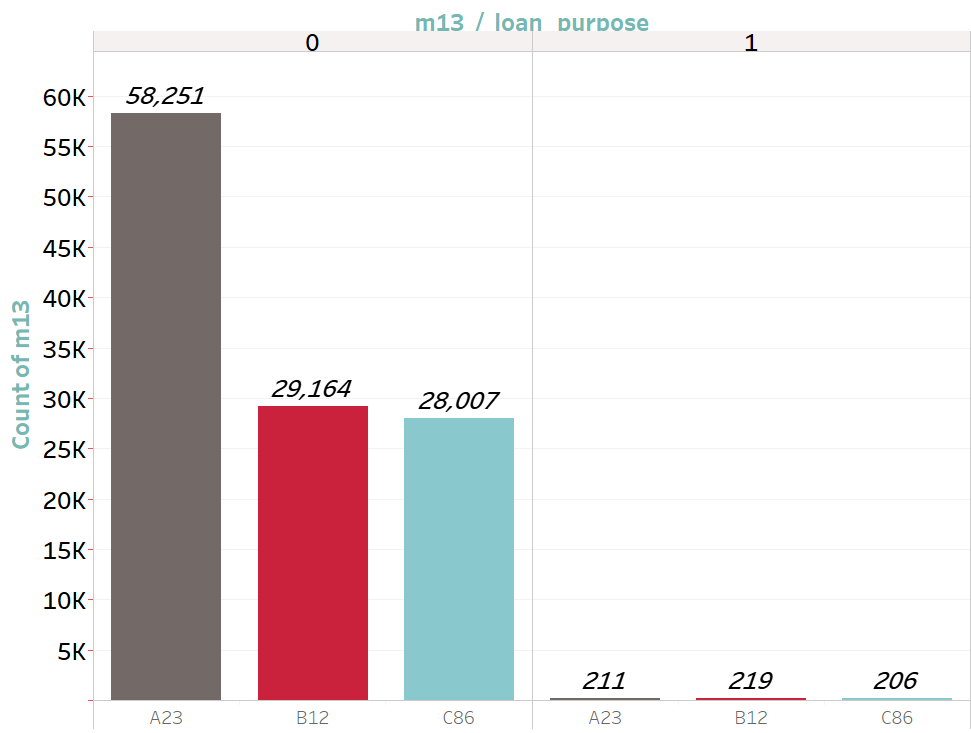
* + - 1. source vs target variable:



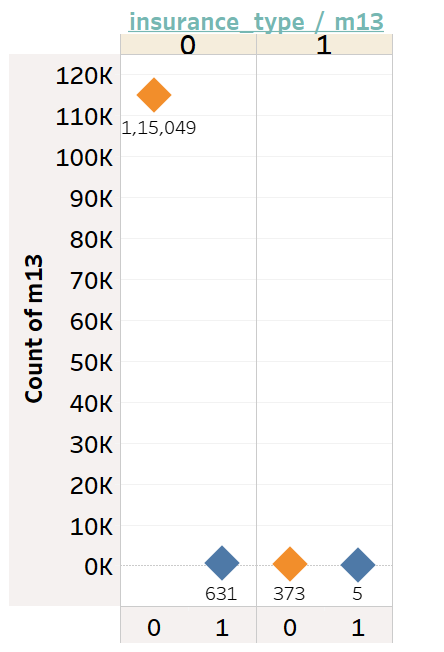
* Source **X** seems to be highest defaulter with 305 cases.
* Source **Z** is the lowest defaulter with 84 cases.
  + - 1. Financial institution vs target variable:



* + - From Financial Institutions we infer that the category called **‘Others’** is the highest defaulter with 269 cases.
    - **Browning-Hart** seems to be second highest defaulter with 175 cases.
    - **Chapman-Mcmahon** has zero cases of defaulting.
      1. Loan purpose vs target variable:



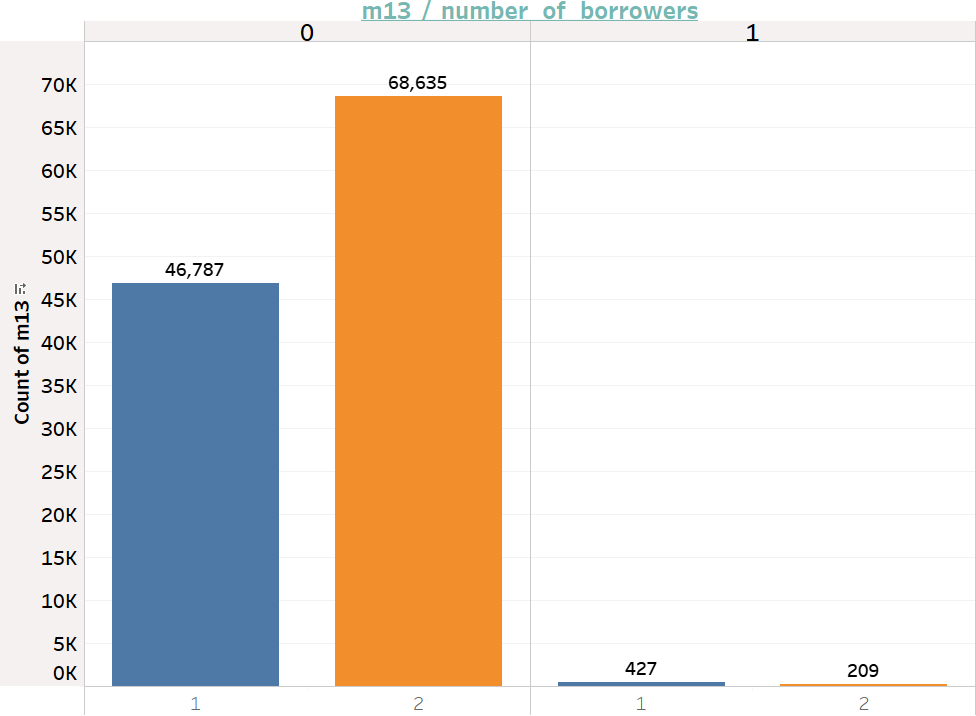
* **B12** loan purpose seems to be highest defaulter with 290 cases.
* **C86** loan purpose seems to be the least defaulter with 206 cases.
  + - 1. Insurance type vs target variable:

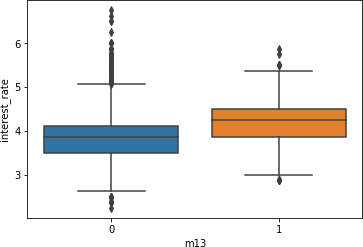


* Premium

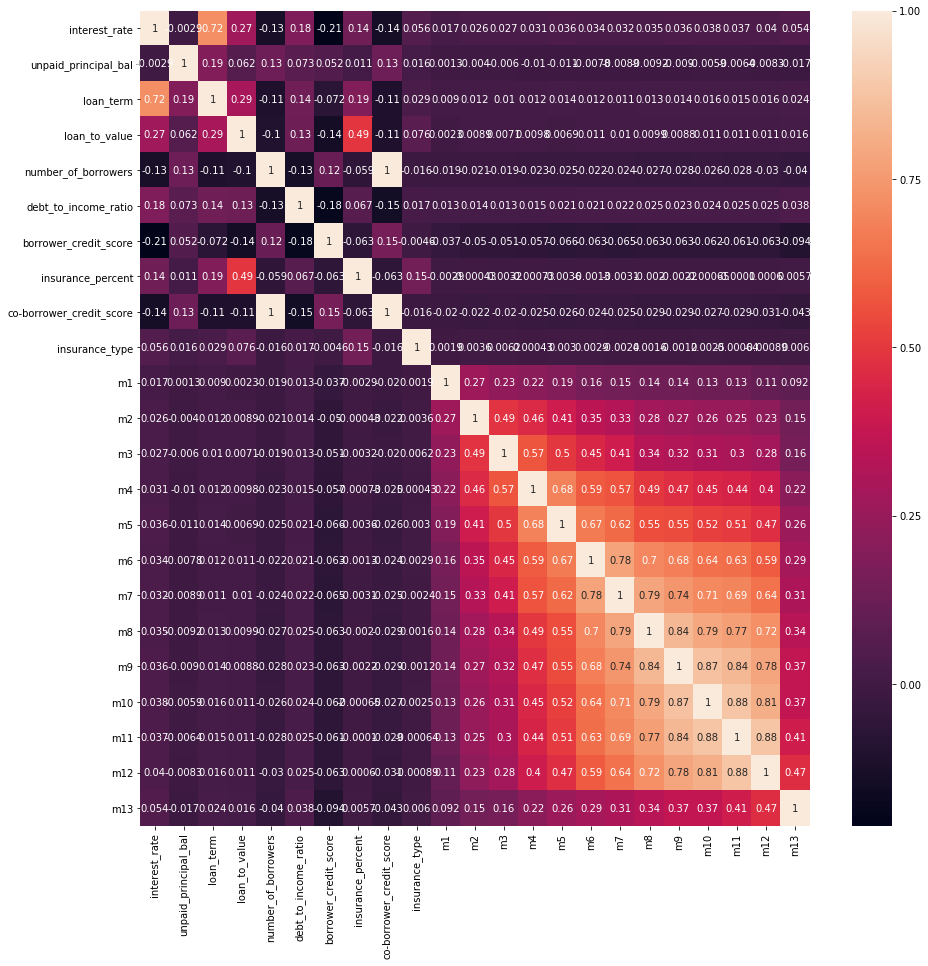
paid by borrower seems to have high delinquency rate with 631 cases.

* And on the other hand, premium paid by Lender has 5 cases of delinquency.
  + - 1. No of borrowers vs target variable:



* Delinquency rate also depends on number of borrowers.
* Number of single borrowers are more delinquent than multiple borrowers with Single borrowers = 427 cases Multiple borrowers = 209 cases.
  + - 1. Interest rate vs target variable:
* Financial Institutions having high interest rate seems to have high delinquency cases.

## Correlation of the dataset:



* From the correlation matrix we can infer that co linearity between feature variables is low, except number of borrowers, Insurance type and co-borrower credit score as they have high co linearity.
* Loan term and Interest rate are also correlated.

## Scaling the data:

Standard Scaler standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation.

Standard Scaler results in a distribution with a standard deviation equal to 1. The variance is equal to 1 also, because variance = standard deviation squared.

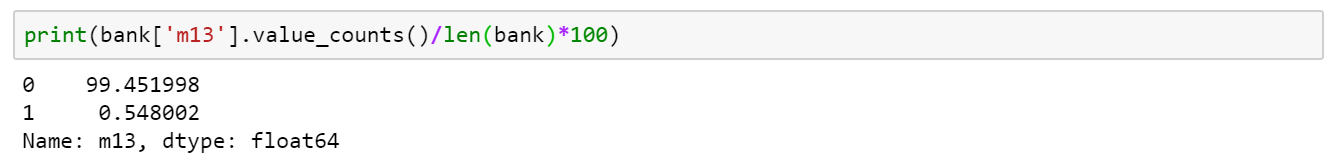
## Imbalance Data Detection:

In our dataset, the target feature is ‘Delinquency’(m13). Delinquency field has two values ‘0’ and '1'.

1: denoting the customers not paying loan amount on time.

0: denoting the customers paying loan amount on time.

From the records, let us see the count of the categories:



This clearly states that the data is highly imbalance.

99.5% are non-delinquent cases and 0.5% are delinquent cases.

## Treatment of Imbalance Data:

There are three major techniques to treat imbalance data:

* Over Sampling

When one class of data is the underrepresented minority class in the data sample, over sampling techniques maybe used to duplicate these results for a more balanced amount of positive results in training. Over sampling is used when the amount of data collected is insufficient.

* Under Sampling

When a class of data is the overrepresented majority class, under sampling may be used to balance it with the minority class. Under sampling is used when the amount of collected data is sufficient.

* SMOTE Analysis

SMOTE (Synthetic Minority Over-sampling Technique), which creates synthetic samples by

randomly sampling the characteristics from occurrences in the minority class.

# Feature Engineering and Model Selection

## Feature Engineering

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable.

Having irrelevant features in data can decrease the accuracy of the models and make the model learn based on irrelevant features.

## Benefits of Feature Selection:

The major benefits of performing feature selection before modelling is as follows:

* Reduces Overfitting
* Improves Accuracy
* Reduces Training Time

## Feature Selection Methods:

Feature selection methods can be divided into three major buckets:

* **Filter based:**We specify some metric and based on that filter features.
* An example of such a metric could be correlation (Pearson Correlation)/chi-square (K-best method).
* VIF - Collinearity is the state where two variables are highly correlated and contain similar information about the variance within a given dataset. To detect collinearity among variables, simply create a correlation matrix and find variables with large absolute values.
* **Wrapper-based:**A Wrapper method needs one machine learning algorithm and uses its performance as evaluation criteria. It is an iterative and computationally expensive process but it is more accurate than the filter method.
* Forward Selection
* Backward Selection
* Recursive Feature Elimination (RFE)
* **Embedded method:** Embedded methods use algorithms that have built-in feature selection methods. Regularization methods are the most commonly used embedded methods which penalize a feature given a coefficient threshold.
* Lasso - it performs L1 regularization which adds penalty equivalent to absolute value of the magnitude of coefficients.
* Ridge – it performs L2 regularization which adds penalty equivalent to square of the magnitude of coefficients.

## Assumptions

### Logistic Regression

* Logistic regression does not require a linear relationship between the dependent and independent variables.
* Second, the error terms (residuals) do not need to be normally distributed.
* Third, homoscedasticity is not required.
* Finally, the dependent variable in logistic regression is not measured on an interval or ratio scale.

However, some other assumptions still apply.

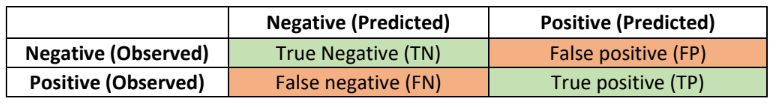
* Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
* Logistic regression requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data.
* Logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other.
* Logistic regression assumes linearity of independent variables and log odds. although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.
* Finally, logistic regression typically requires a large sample size. A general guideline is that you need at a minimum of 10 cases with the least frequent outcome for each independent variable in your model. For example, if you have 5 independent variables and the expected probability of your least frequent outcome is .10, then you would need a minimum sample size of 500 (10\*5 / .10).

## Model performance measures used for evaluating models:

The various models built, must be evaluated based on certain model performance measures to

identify the most robust models. The choice of the right model performance measures is highly critical since the dataset is a highly imbalanced dataset and the conversion rate is 15.47%.

Model accuracy alone may not be enough to evaluate a model. Hence the following model performance measures have been used to evaluate the models, based on the confusion matrix built for the predictions on the training and test datasets:



*Accuracy*

Accuracy is the number of correct predictions made by the model by the total number of

records. The best accuracy is 100% indicating that all the predictions are correct.

Considering the response rate (conversion rate) of our dataset which is ~16%, accuracy is not a

valid measure of model performance. Even if all the records are predicted as 0, the model will still have an accuracy of 84%. Hence other model performance measures need to be evaluated.

*Sensitivity or recall*

Sensitivity (Recall or True positive rate) is calculated as the number of correct positive

predictions divided by the total number of positives. It is also called recall or true positive rate (TPR).

For our dataset, it gives the ratio of actual customers who generated revenue by the total

number of customers predicted who will generate the revenue.

*Specificity*

Specificity (true negative rate) is calculated as the number of correct negative predictions

divided by the total number of negatives.

For our dataset, specificity gives the ratio of actual customers who will not generate revenue by

the number of customers who are predicted who will not generate revenue.

*Precision*

Precision (Positive predictive value) is calculated as the number of correct positive predictions

divided by the total number of positive predictions.

Precision tells us, what proportion of customers who generated revenue as customers actually

generated revenue. If precision is low, it implies that the model has lot of false positives.

*F1-Score*

F1 is an overall measure of a model’s accuracy that combines precision and recall A good F1

score means that you have low false positives and low false negatives, so you’re correctly identifying real threats and you are not disturbed by false alarms.

An F1 score is considered perfect when it’s 1, while the model is a total failure when it’s 0.

*ROC chart & Area under the curve (AUC)*

ROC chart is a plot of 1-specificity in the X axis and sensitivity in the Y axis. Area under the

ROC curve is a measure of model performance. The AUC of a random classifier is 50% and that of a perfect classifier is 100%. For practical situations, an AUC of over 70% is desirable.

*Level of significance*

For all the hypothesis tests in the project, the level of significance is assumed as 5% unless

specified otherwise.

## Model Selection

### Classification Report

One of the purposes of this project is to get the analysis of predicting the number of delinquents for an year. The dataset is fed to Logistic Regression, Naïve Bayes, Random Forest, Gradient Boosting and Ada boost classifiers. The Accuracy, Precision, Bias Error and Variance Error and F1-Score are presented for each classifier. Results on class imbalanced dataset:

Tables below show the results obtained with Logistic Regression, Naïve Bayes, Random Forest, Gradient Boosting and Ada boost classifiers respectively. However, a class imbalance problem arises since the number of negative class instances in the data set is much higher than that of the positive class instances, and the imbalanced success rates on positive (TPR) and negative (TNR) samples show that the classifiers tend to label the test samples as the majority class. This class imbalance problem is a natural situation to most of the financial institutions.

|  |  |  |
| --- | --- | --- |
| Algorithms | ROC-AUC Score | F1-Score |
| Logistic Regression | 0.81378 | 0.00 |
| Naïve Bayes | 0.83766 | 0.11 |
| Ada boost | 0.88242 | 0.51 |
| Gradient Boosting | 0.89113 | 0.45 |
| Random Forest | 0.89807 | 0.49 |

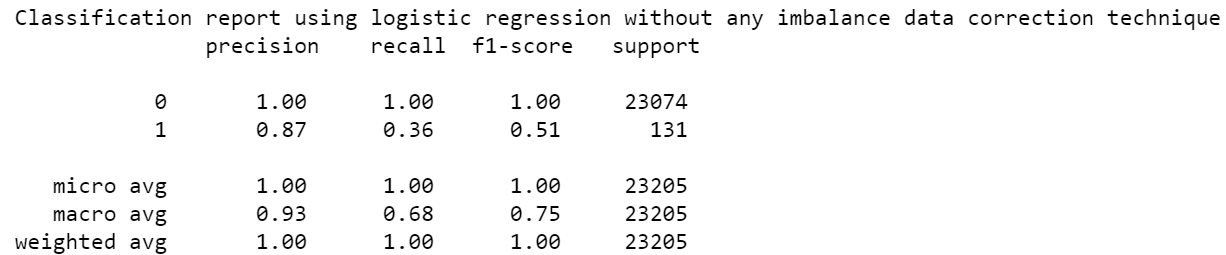
Also from the literature survey, it was found out that all the mentioned features were correlating with each other and also were contributing significant amount of inference in the dataset(target variable: m13)

Feature selection

Performed feature selection using RFE.

* RFE method works by recursively removing attributes and building a model on those attributes that remain.
* It uses accuracy metric to rank the feature according to their importance.
* The RFE method takes the model to be used and the number of required features as input.
* It then gives the ranking of all the variables, 1 being most important.
* It also gives its support, True being relevant feature and False being irrelevant feature.

Got 18 features as the most relevant one for predicting target variable correctly.



Here, as we can observe f1-score is 1 for predicting Non-delinquency and 0.51 for predicting Delinquency.

We can infer that the model is not learning the information properly and hence tends to overfit.

Therefore, we opted for model building without any feature selection.

### Results obtained with oversampling:

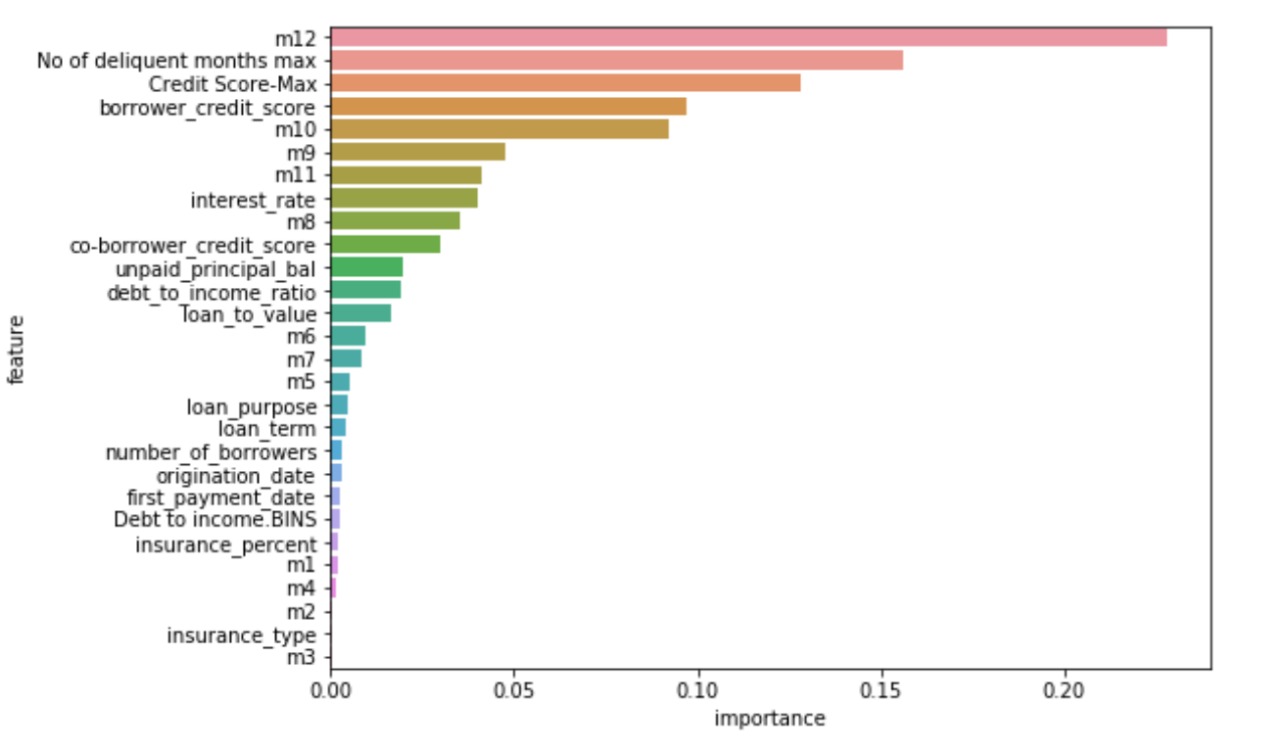
The results presented in this section show that the classifiers tend to minimize their errors on majority class samples, which leads to an imbalance between the accuracy rates of the positive and negative classes. However, in a real-time user behaviour analysis model, correctly identifying defaulters, which are represented with positive class in our dataset, is as important as identifying negative class samples. Therefore, a balanced classifier is needed to increase the conversion rates in Financial Institutions. To deal with class imbalance problem, we use oversampling method, in which a uniform distribution over the classes is aimed to be achieved by adding more of the minority (positive class in our dataset) class instances. Since this dataset is created by selecting multiple instances of the minority class more than once, first oversampling the dataset and then dividing it into training and test sets may lead to biased results due to the possibility that the same minority class instance may be used both for training and test. For this reason, in our study, we are matching with majority class through over sampling, by taking 40% as n\_samples to overcome highly imbalanced data i.e., 69234 as defaulters.

The results obtained on the balanced dataset are shown in the table below. Since the number of samples belonging to positive and negative classes is equalized with oversampling, both accuracy and F1-score metrics can be used to evaluate the results.

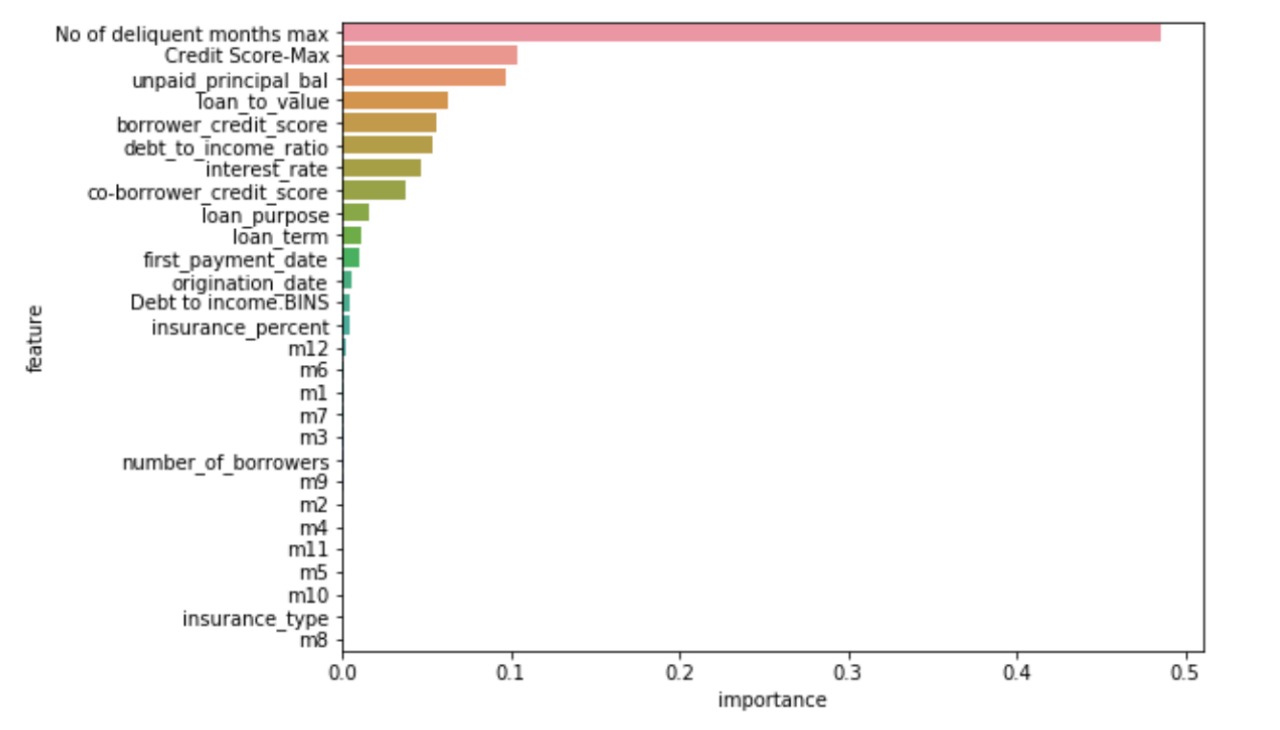
|  |  |  |
| --- | --- | --- |
| Algorithms | ROC-AUC Score | F1-Score |
| Logistic Regression | 0.81357 | 0.66 |
| Naïve Bayes | 0.83457 | 0.68 |
| Ada boost | 0.91625 | 0.78 |
| Gradient Boosting | 0.94741 | 0.81 |
| Random Forest | 0.97318 | 0.86 |

### Feature Importance for Tree based Algorithms:

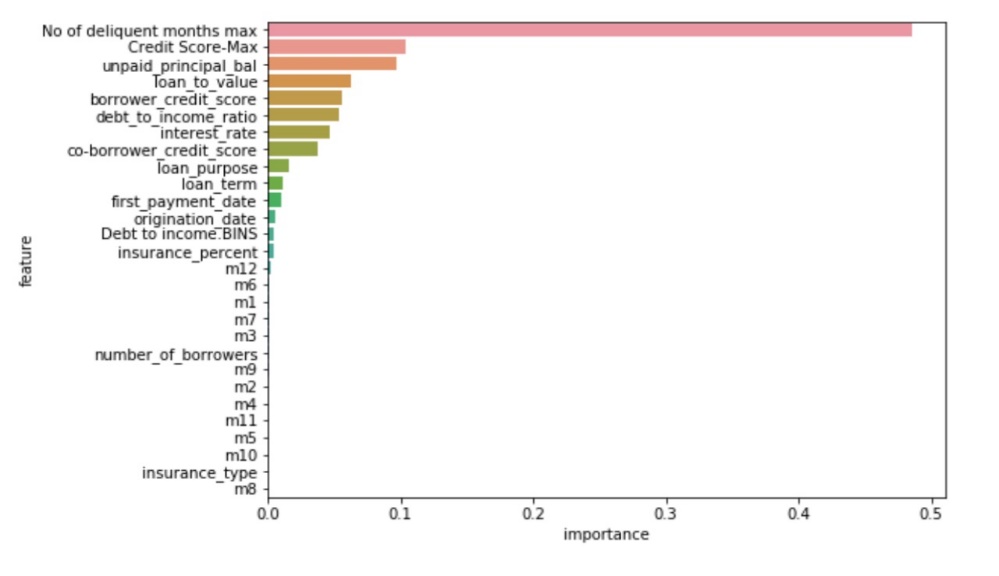
#### Random Forest



#### Ada Boost



#### Gradient Boost

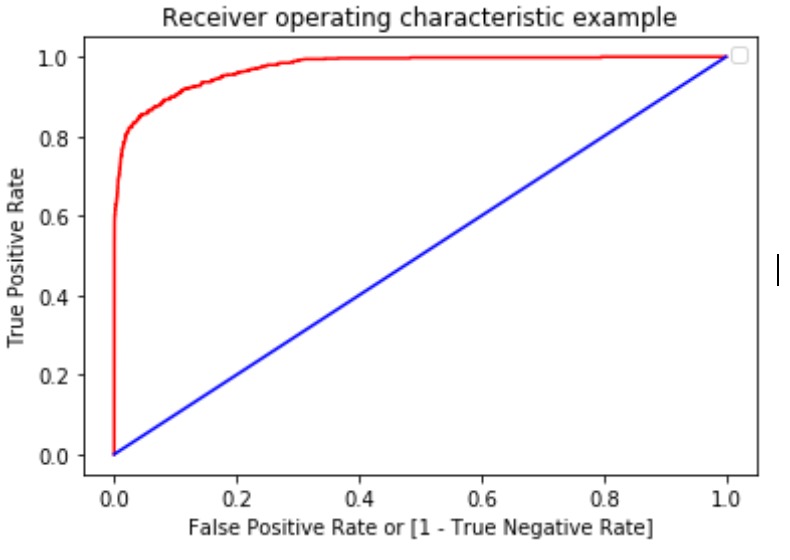


## Final Model:

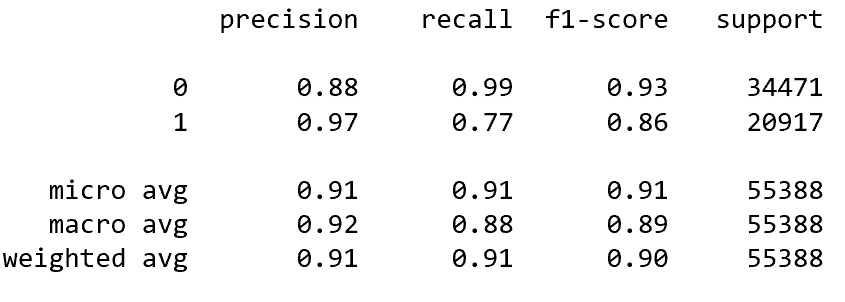
### Random Forest:

From the above algorithms we conclude that **“Random Forest”** is the best model after comparing the F1-Score and ROC-AUC Scores.

#### ROC-AUC Curve:



#### Classification Report**:**



* From Classification report of Random Forest, we can infer that, after implementing resampling technique to match majority class overfitting avoided in the model with f1-score of non-delinquent case is 0.93 and delinquent case is 0.83
* The AUC-ROC score of Random Forest is 0.97318, we can infer that True Positive rate has been increased and False Positive Rate has been decreased.

# Conclusions:

In this project we aim to construct a real time delinquency prediction in the services for different financial institutions. We use analytics data to perform the experiments. In order to predict the delinquent status which is the target variable here, various univariate and bivariate analysis is done initially to check the relationship between the variables. Some great insights were found such as financial institutions which have higher interest rates have higher delinquency rates, customers who have insured their loans tend to repay the loan without missing the payment date. The graphs used were visualised with the help of Tableau.

For the imbalanced data, oversampling technique has been used here to treat the underrepresented minority class. Feature engineering has been done to get better features and different feature selection methods like RFE and Backward elimination were also used. This was a classification model so the accuracy metrics taken into consideration were the f1 score and ROC-AUC score. As the model was tending to overfit after using the feature selection, we considered all the features after doing more of literature survey for the features.

Finally, after fitting the model into the oversampled data we found that amongst the models like Logistic Regression, Naïve Bayes, Ada Boost, Gradient Boosting, Random Forest; the ROC-AUC score and F1 score of Random forest was about 97% and 86% respectively which inferred that Random Forest is the best model for the dataset.

# Recommendations:

From the above dataset, following methods are recommended: -

1. The loan duration should be increased, so as to give enough time to borrowers to repay the debt.
2. The interest rate can be lowered. It has been observed from the trend that with lower interest rate the delinquency rate is minimal.
3. The number of dependents should have been mentioned in the dataset since; more inference could have been drawn. With increase in the number of dependents, and with lesser income it can be said that the customer is more delinquent.
4. Also, the features like salary and salary pay date are missing which should have been available for inferring further. The salary and the pay date would have provided more accurate results for the delinquency status.
5. The origination and the first payment date of the loan should have been available as a part of the data since it relates with the salary pay date. If the first payment date is within a week after receiving the salary the customers tend to pay the debt often with less frequency of being delinquent.
6. More number of insurance on loan should be encouraged, as customers who had their loan insured tend to pay the loans dutifully in comparison to the customers who did not insure their loans.
7. Depending on the income of a particular customer, proper allocation of loans should be done that is amount of loan that will be sanctioned should be in accordance to the income.
8. Proper research techniques on past behaviour of the borrower should be taken into consideration. With various Artificial intelligence techniques and text mining, customers having ongoing debts at the time applying the loan again can be detected and such applications could be kept on hold till a proper conclusion and need is made.
9. Better norms and regulations should be made for the customers who are unable to pay loans for more than 12 months. This can be observed from the delinquency status (Target Variable) in the dataset. With a constant delinquent trend of a customer, an attempt can be made to motivate the borrower for repaying the loan by reducing the interest rate by some percentage so that the financial institution does not bear a huge loss and also gets back the original amount that had been sourced out.
10. Proper reminders should be made out to the defaulters from the financial institutions and after giving ample amount of time and lowering the interest rate, if the defaulters still do not pay back the loan additional penalties like fine can be imposed so that the borrowers repay the money back and the institutes do not fall under huge loss.

# References and Bibliography:

* <https://cafral.org.in/sfControl/content/Speech/621201790033PMLoan_Delinquency_in_Banking_Systems.pdf> - Loan Delinquency in Banking Systems
* <http://www.pbr.co.in/2017/2017_month/August/14.pdf> - An Empirical Analysis of Factors affecting the Loan Repayment
* <https://pdfs.semanticscholar.org/1130/aa51c0d4ce6ba47c5d0d2df6ebd230d02990.pdf> - Causes and Control of Loan Default/Delinquency in Microfinance Institutions in Ghana
* <https://www.lendfoundry.com/regression-analysis-as-a-fintech-tool-for-predicting-successful-loan-applications/> - Regression Analysis as a successful fin-tech tool