

**“Predicting Delinquency and Loan Default with Advanced Analytics, Machine Learning and analyzing the performance of Banks in MSME Sector”**

***Banking and Finance Domain***

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

Post Graduate Program in Business Analytics and Business Intelligence

Capstone Project Report

Submitted to



Submitted by

Jai Kushwaha(9693OHEPBG), Anu Rathi(6Y55BS1Z79),

Sanjeev Sidharth (VHBQSBXRC4),Amol Pol(1AEF0X1YL6)

(Group 4)

Under the guidance of

Richa Agarwal (Mentor)

Batch – (PGPBABI **July**’1**9**)

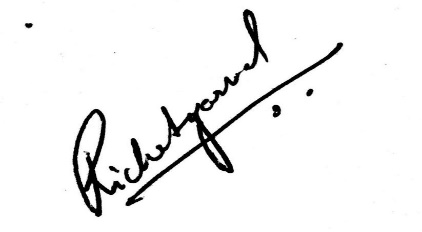
Year of Completion (20**20**)

**CERTIFICATE**

This is to certify that the participants Jai Kushwaha, Anu Rathi, Sanjeev Sidharth, Amol Pol who are the students of Great Lakes Institute of Management, has successfully completed their project on “Predicting Delinquency and Loan Default with Advanced Analytics, Machine Learning and analyzing the performance of Banks in MSME Sector”

This project is the record of authentic work carried out by them during the academic year 2019- 2020.

Mentor’s Name and Sign Program Office



Richa Agarwal

Date: 3rd January 2021

Place: Pune

Contents

[List of Figures 5](#_Toc60594411)

[List of Tables 6](#_Toc60594412)

[Abbreviations List 7](#_Toc60594413)

[1. Executive Summary 8](#_Toc60594414)

[1.1. Project Approach 8](#_Toc60594415)

[1.2. Achievements 9](#_Toc60594416)

[2. Introduction 10](#_Toc60594417)

[2.1. Need of the study 10](#_Toc60594418)

[2.2. Objectives 10](#_Toc60594419)

[2.2.1. The present study has been taken up with the following scopes 10](#_Toc60594420)

[2.2.2. Business Problem and suggestion for Business Implementation 10](#_Toc60594421)

[2.3. Data Sources 11](#_Toc60594422)

[2.4. Challenges faced by the bank 11](#_Toc60594423)

[2.5. Tools and techniques used 12](#_Toc60594424)

[3. Literature Review 13](#_Toc60594425)

[3.1. Banking and NPA History 13](#_Toc60594426)

[3.2. What is the extent and effect of the NPA problem in India? 13](#_Toc60594427)

[3.3. Banking and Analytics 15](#_Toc60594428)

[3.4. Background about MSME sector 15](#_Toc60594429)

[4. Detailed Methodology 17](#_Toc60594430)

[4.1. Data Extraction, Gathering, Cleaning and Pre-processing 17](#_Toc60594431)

[4.1.1. Data Extraction and Gathering 17](#_Toc60594432)

[4.1.2. Raw data Exploratory data analysis. 17](#_Toc60594433)

[4.1.3. Missing Values 19](#_Toc60594434)

[4.1.4. Data cleaning and variable datatype modification of the variables 19](#_Toc60594435)

[4.1.5. Missing value imputation 19](#_Toc60594436)

[4.2. Exploratory Data Analysis - Explore and gain insights from a dataset. 20](#_Toc60594437)

[4.2.1. Data Understanding 20](#_Toc60594438)

[4.2.2. Understanding the target variable 24](#_Toc60594439)

[4.2.3. Five-point Summary: 25](#_Toc60594440)

[4.3. Feature Importance Understanding 29](#_Toc60594441)

[4.3.1. Feature importance using decision tree and its score 29](#_Toc60594442)

[4.3.2. Feature importance using random forest technique 29](#_Toc60594443)

[4.4. Feature Engineering 30](#_Toc60594444)

[4.4.1. Converting relevant numeric fields to factor 30](#_Toc60594445)

[4.4.2. One hot Encoding 30](#_Toc60594446)

[4.5. Data splitting 31](#_Toc60594447)

[4.6. Modeling Technique 32](#_Toc60594448)

[4.6.1. Modeling for delinquency 32](#_Toc60594449)

[4.6.2. Modeling for Loan Default 33](#_Toc60594450)

[4.7. Component Reduction 37](#_Toc60594451)

[4.7.1. Principal Component Analysis 37](#_Toc60594452)

[4.7.2. Component Selection using VIF 38](#_Toc60594453)

[4.8. Modeling comparison 39](#_Toc60594454)

[4.9. Model interpretation through LIME: 40](#_Toc60594455)

[4.10. Deployment 41](#_Toc60594456)

[4.10.1. User Interfaces for banking enterprise 41](#_Toc60594457)

[5. Achievements 42](#_Toc60594458)

[5.1. Achieved in the current 42](#_Toc60594459)

[5.2. Future Enhancements 43](#_Toc60594460)

[6. Recommendations and Conclusion 44](#_Toc60594461)

[7. Bibliography 45](#_Toc60594462)

[8. Annexure 46](#_Toc60594463)

[8.1. Data Description 46](#_Toc60594464)

[8.2. Detailed EDA 49](#_Toc60594465)

[8.3. Feature Importance for all variables 49](#_Toc60594466)

[8.4. Detailed VIF values 52](#_Toc60594467)

[8.5. Code Files 53](#_Toc60594468)

# ****List of Figures****

[Figure 1: Delinquency and Default understanding 8](#_Toc60594469)

[Figure 2: Approach for Data collection and understanding 8](file:///C:\Users\sidht\Downloads\Capstone%20Project_Final%20report%20v1_Final.docx#_Toc60594470)

[Figure 3: Project Lifecycle 12](#_Toc60594471)

[Figure 4: Identified reasons from NPA from literature review 14](file:///C:\Users\sidht\Downloads\Capstone%20Project_Final%20report%20v1_Final.docx#_Toc60594472)

[Figure 5: Distribution of non- performing assets 14](#_Toc60594473)

[Figure 6: Understanding of MSME Sector 15](#_Toc60594474)

[Figure 7: Role of Analytics in Banking 16](#_Toc60594475)

[Figure 8: VoterID details of users 20](file:///C:\Users\sidht\Downloads\Capstone%20Project_Final%20report%20v1_Final.docx#_Toc60594476)

[Figure 9: Mobile Details of User 21](#_Toc60594477)

[Figure 10: CNS Score Details 21](#_Toc60594478)

[Figure 11: Age Detail of customers 22](#_Toc60594479)

[Figure 12: Perform CNS score 22](#_Toc60594480)

[Figure 13: Box Plot for "lvt" variable 23](#_Toc60594481)

[Figure 14: Numerical and Categorical association for "ltv" 24](file:///C:\Users\sidht\Downloads\Capstone%20Project_Final%20report%20v1_Final.docx#_Toc60594482)

[Figure 15: Target Variable Distribution 24](#_Toc60594483)

[Figure 16: Feature Importance score based on random forest 30](#_Toc60594484)

[Figure 18: Model comparison based on performance parameters 39](#_Toc60594485)

[Figure 17: Model Interpretation with LIME 40](#_Toc60594486)

[Figure 19: Sample Screen shot for banking enterprise user interface 41](#_Toc60594487)

# ****List of Tables****

[Table 1: Abbreviation List 7](#_Toc60594488)

[Table 2: Data Understanding 17](#_Toc60594489)

[Table 3: Detailed Column wise overview of Data master file 18](#_Toc60594490)

[Table 4: Data Understanding - Missing value columns 19](#_Toc60594491)

[Table 5: Five-point summary details 25](#_Toc60594492)

[Table 6: Top Identified Features based on decision tree 29](#_Toc60594493)

[Table 7: Delinquency Model Performance 32](#_Toc60594494)

[Table 8: Delinquency model performance for scaled data 33](#_Toc60594495)

[Table 9: Model Performance Interpretation 33](#_Toc60594496)

[Table 10: Logistic Regression Model Performance 34](#_Toc60594497)

[Table 11: Decision Tree Model Performance 34](#_Toc60594498)

[Table 12: Adaboost Model Performance 34](#_Toc60594499)

[Table 13: Random Forest Model Performance 35](#_Toc60594500)

[Table 14: SVC Model Performance 35](#_Toc60594501)

[Table 15: Naive Bayes Model Performance 35](#_Toc60594502)

[Table 16: KNN Model Performance 36](#_Toc60594503)

[Table 17: CatBoost Model Performance 36](#_Toc60594504)

[Table 18: LGBM Model Performance 36](#_Toc60594505)

[Table 19: EGBM Model Performance 37](#_Toc60594506)

[Table 20: Collective Thresholds for models 37](#_Toc60594507)

[Table 21: VIF sample values 39](#_Toc60594508)

[Table 22: Model performance after applying SMOTE technique 40](#_Toc60594509)

[Table 23: Data Description 46](#_Toc60594510)

[Table 24: Feature Importance 49](#_Toc60594511)

[Table 25: Detailed VIF values 52](#_Toc60594512)

[Table 26: Code Files (Attachments) 53](#_Toc60594513)

# ****Abbreviations List****

Table 1: Abbreviation List

|  |  |  |
| --- | --- | --- |
| # | Abbreviation | Definition |
| 1 | NPA | Non-Performing assets |
| 2 | EDA | Exploratory data analysis |
| 3 | Sec. | Sector |
| 4 | Accts. | Accounts |
| 5 | SME | Small and medium-sized enterprises |
| 6 | MSME | Micro, Small & Medium Enterprises |
| 7 | Gov. | Government |
| 8 | ID | Identification |
| 9 | PAN | Permanent Account Number |
| 10 | IDE | Integrated Development Environment |
| 11 | LGBM | Light Gradient Boosting Modeling |
| 12 | EGBM | Extreme Gradient Boosting Modeling |
| 13 | SVC | Support Vector Classification |
| 14 | AUC | Area Under Curve |

# Executive Summary

Academics and practitioners have studied over the years models for predicting firm’s bankruptcy, using statistical and machine-learning approaches. An earlier sign that a company has financial difficulties and may eventually bankrupt, i.e. it is going in default, can be diagnosed has been having difficulties in repaying its loans towards the banking system. A Firm’s default status is not technically a failure, but it is very relevant for bank lending policies and often anticipates the failure of the company.

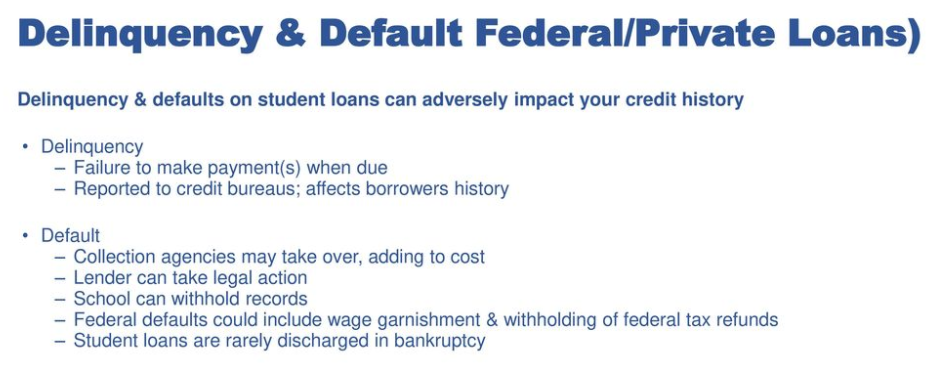


Figure 1: Delinquency and Default understanding

## Project Approach

The work that we have completed:

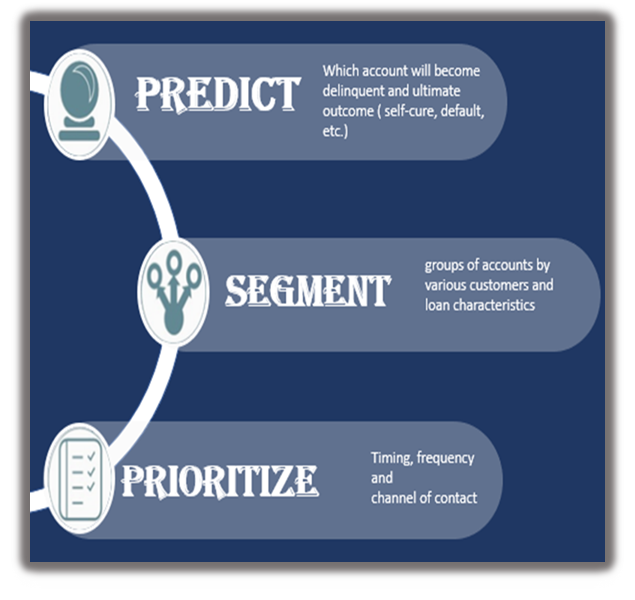
* Collected more data that was available with the bank for the individuals
* Merged data from other sources like, demographic information, account details etc. for further deep analysis
* Collected the CNS score from the data available for the users
* All the collected data was further segregated the data into Customer Master Information, Loan Profile Information and Customer Loan Transactional Data.
* Feature engineered new and derived attributes for the data model
* Data Quality and preparation activities were performed like missing value treatment, imputation, data type conversions for homogeneity in the data set.

Figure 2: Approach for Data collection and understanding

* Performed EDA on the data to understand the data and to determine any outlier and treatment for the same
* Narrowed our choices to Random Forest and XG boost for modeling based our analysis criteria mentioned in detail in modeling section
* Also used PCA and SMOTE techniques to understand if the model performance can be improved.

## Achievements

* Deep dived into various aspects of loans given by the banks to MSME sectors and the important measures that are required while giving the loan.
* Early detection of a default in case of these kind of loans helps bank reduce their NPA and overall reducing the impact on the Indian economy.
* Early detection by detecting if the customer will go delinquent by identifying the respective customers and effectively reducing the NPA
* Factors while giving the loan and information to be taken from the for effective performance of a loan.
* Better scrutiny process for the predicted default.
* Constantly monitoring the already given loans and its health for preventing it into going to NPA.

# Introduction

## Need of the study

Small borrowers, whose default rates have traditionally been among the lowest in India, are increasingly missing loan repayments as rising unemployment and stagnant-to-declining wages put pressure on finances of small companies as well as households.

At the SIDBI National Microfinance Congress 2019, Reserve Bank of India deputy governor M.K. Jain said banks need to focus on repayment capacity of borrowers at the appraisal stage and monitor the loans through their lifecycle much more closely.

Indian banks, which hold the dubious distinction of having the world’s worst bad loan ratio, have so far seen most of its asset-quality troubles originate in the corporate sector. With the economy slowing down sharply, smaller borrowers are also finding it difficult to repay loans.

The reasons for the rising defaults are multifold. Experts said the slowdown in economic growth is hampering the loan-servicing capability of small businesses. For small retail borrowers, slackening growth in rural income has been a pain point.

As working capital cycles of large and mid-corporates stretch, the SMEs bear the brunt of the same. With banks hesitating to lend to them incrementally if there are even marginal cash flow issues, the defaults in SME could materially increase even from 10-11% seen currently for public sector banks (PSBs).

## Objectives

As India eyes towards 5$ Trillion economy by 2024, 2$ trillion are proposed by MSME. However, the MSME sector of India faces challenges to stay afloat and keep running. This in turn impacts banks which are the financial backbone of this industry. And an unstable MSME result in financial loss for banks which transpires to slow growth of Indian economy.

As a part of our project, we are proposing the use of Advanced Analytics and Machine learning to help financial institutes like banks understands their MSME borrowers and predict possible delinquency. It is then that adequate steps could be taken to understand the reasons behind payment defaults and thereby prevent NPAs. Through this project we propose to address the interdependence, reduce financial risk and improve financial stability.

### **The present study has been taken up with the following scopes**

* To understand the performance of the bank in various regions and for individual branches.
* To make a study to understand magnitude of impact of different schemes and specific industry sectors.
* To examine the causes for incidence and trends of NPAs in MSME sector for the bank.
* To create a machine learning model which will be able to predict delinquency and default based on account parameters.

### **Business Problem and suggestion for Business Implementation**

* Objective to make suitable suggestions for any Financial Institution based on the performance of the bank as per the predictive model.
* As we know Government is pushing banks to disburse as many MSME loans for reviving the economy. And Banks are constantly under pressure to deliver the targets with limited number of manpower available to PSU banks as many retirements are taking place and creating a resource bubble. However, this leads to increase number of advances and which in turn increases the of loan going delinquent or defunct and banks’ ability to reduce overall NPA.
* So as to help the business our model will help detect customers which will be having high risk of defaulting at an early stage and based on simple data provided of customer and help business target these customers from going defaulting and also, with efficient usage of man power and come up an early resolution plans for these kinds of advances.

## Data Sources

We had received earlier is the live data with some masked fields from a bank for which we are looking to create a predictive model to predict delinquency of the customers. The data is the primary data that bank collects for its various customers for bank’s internal records. The data set has been received from various departments and we have collated it as per our analytical problem statement. We also derived a few new columns from the current data for better learning of the model.

At the time of writing this report, we have been successful in getting more data around the details of the customer, their details, payment schedules etc. which in term will help us in accurately creating a model for delinquency in loan default. The key data sets we received till now are:

1. Loan Account Status
2. KYC data (masked) like mobile number, type of documents submitted as a part of KYC, employment condition etc.
3. Customer’s financial data (liabilities and credit scores)
4. Loan Sanction Details like asset cost, disbursement amount.

Based on the earlier data available and more data that we received, we have created a consolidated data set which will be taken as the basis of model creation for prediction the loan default.

***For details of data dictionary please refer annexure section 8.1.***

## Challenges faced by the bank

* Data is distributed and not available at one location
* Departments working in silos
* Irregularities in document verification
* Payments defaults by the customer
* MSME sector loan defaults increasing
* Economic slow down
* Changes due to global banking developments, policy environment, operations and performance of commercial banks, developments in cooperative banking and non-banking financial institutions.

## Tools and techniques used

We have used the statistical modelling technique to come up with a solution for predicting defaults which can serve as a useful tool for the banks for their futures proceedings.

We used the following process where various techniques like data mining, data exploration, cleaning, features engineering, predictive modeling, deep learning and visualization techniques were used.



Figure : Project Lifecycle

The major tools that we used during our project are:

IDE – Notebook, R studio

Language – Python, R, Flask

Vitalization Tool – Tableau

Other tools – MS Excel, Word, PowerPoint

Online IDE – Google Colab

Model Deployment – Heroku which is a platform as a service (PaaS)

# Literature Review

## Banking and NPA History

Authentic history of banking tells that it deals with lending and collection of money. The concept of Non-Performing Assets (NPAs) was introduced for the first time in the Narasimhan Committee on “Financial System Reforms” that was tabled in Parliament on December 17th 1991. The Committee studied the prevailing financial system, identified its short comings and weaknesses and made with ranging suggestions and recommendations in line with internationally accepted norms. Based on the recommendations of the Committee on “Financial System Reforms”, the RBI evolved prudential norms on Income recognition, Asset classification and Provisioning and issued revised instructions to banks in April 1992. While conveying non-performing category and their anxiety to present rosy picture of their affairs the above instructions to banks also 4 advised them that as per practice followed internationally, income on NPAs is not to be recognized on accrual basis but is to be looked only when it is actually realized because an asset becomes non-performing when it ceases to generate income. The above instructions of RBI have since been implemented by banks from the financial year ended March 1998. The problem of NPAs is linked to the function of lending money. The lending of money collected from the public, for interest, instead of one’s own money, was the beginning of banking. Though the present-day banking does not restrict itself to traditional deposit collection and money lending, encompassing a wide sphere of financial activity, lending still remains the prime activity connected with banking. Most credit needs of the society, for carrying, commercial activities are fulfilled by the banks. The conventional credit from the banking system to the Commercial sector comprises bank loans and advances in the form of term loans, demand loans, cash credit, overdrafts, inland and foreign bills purchased and discounted as well as investments in instruments issued by non-government sector.

## What is the extent and effect of the NPA problem in India?

Banks give loans and advances to borrowers. Based on the performance of the loan, it may be categorized as:

1. a standard asset (a loan where the borrower is making regular repayments), or
2. a non-performing asset. NPAs are loans and advances where the borrower has stopped making interest or principal repayments for over 90 days.

**What led to the rise in NPAs?**

Some of the factors leading to the increased occurrence of NPAs are external, such as decreases in global commodity prices leading to slower exports. Some are more intrinsic to the Indian banking sector. One of the prominent one is the MSME sector.

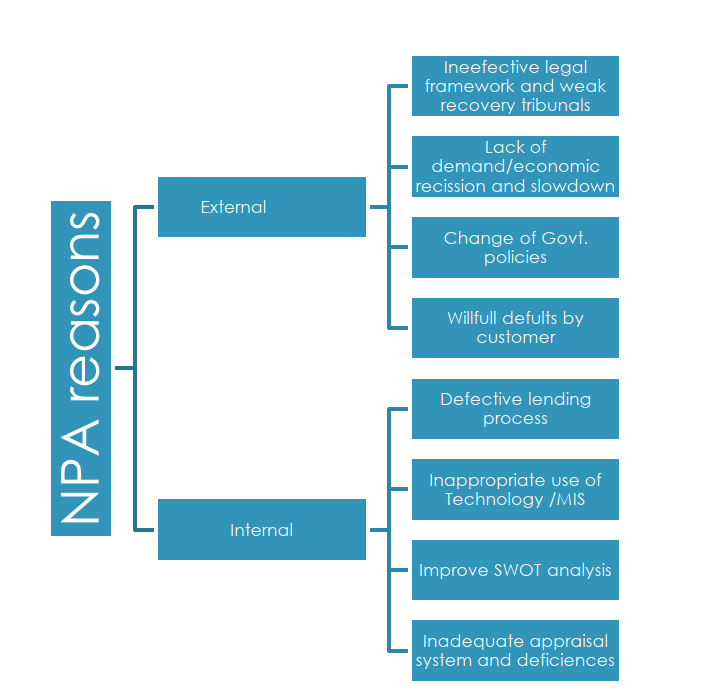


Figure 4: Identified reasons from NPA from literature review

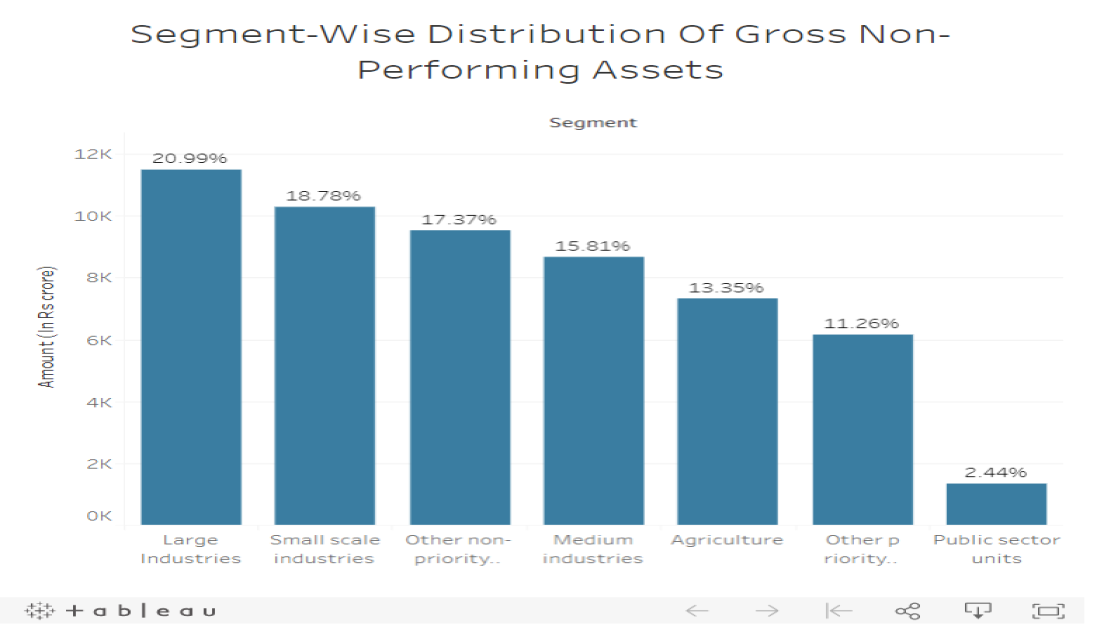


Figure 5: Distribution of non- performing assets

So, from the above it is evident the MSME sector plays 33% part in the overall in the Indian economy

## Banking and Analytics

The Banking industry generates a huge volume of data on a day-to-day basis. To differentiate itself from the competition, banks are increasingly adopting big data analytics as part of their core strategy. Analytics will be the critical game changer for the banks. Adopting it has become necessary in almost all sectors that banks deal with. One such sector that we are going into depth is the MSME sector.

## Background about MSME sector

The Micro, Small and Medium Enterprises (MSME) sector has emerged as a highly vibrant and dynamic sector of the Indian economy over the last five decades. It contributes significantly to the economic and social development of the country by fostering entrepreneurship and generating largest employment opportunities at comparatively lower capital cost, next only to agriculture. MSMEs are complementary to large industries as ancillary units and this sector contributes significantly in the inclusive industrial development of the country. The MSMEs are widening their domain across sectors of the economy, producing diverse range of products and services to meet the demand of domestic as well as global markets.

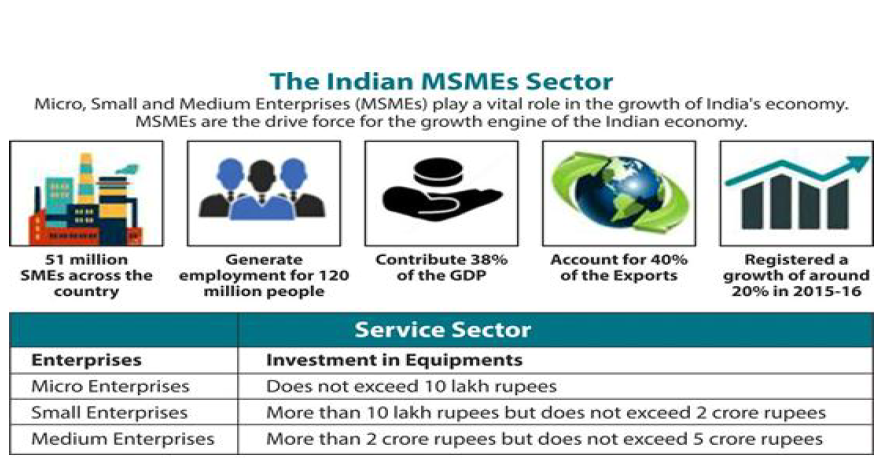


Figure 6: Understanding of MSME Sector

**Source:** International Chamber of Service Industry (ICSI)

The Role of Analytics in Banking, or applications of data mining in banking, enhances the performance of the banks by improving how banks segment, target, acquire, and retain customers.

Furthermore, improvements in risk management, customer understanding, and fraud empower banks to maintain and grow a profitable customer base. The application of data mining and predictive analytics to extract actionable insights and quantifiable predictions can help the banks to gain insights that comprise of all types of customer behavior, including channel transactions, account opening and closing, default, fraud, and customer departure.

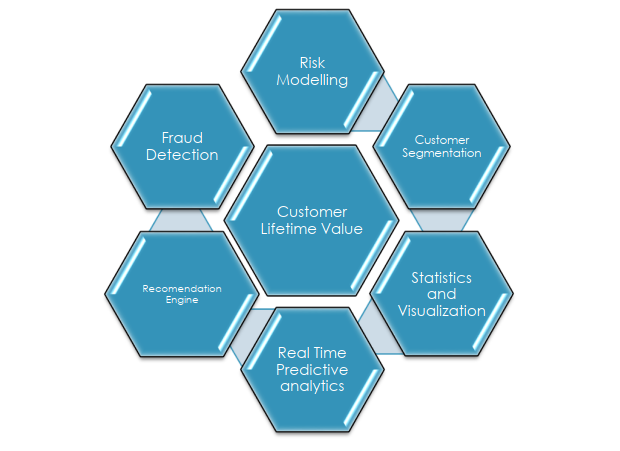


Figure 7: Role of Analytics in Banking

Banking industry has a very rich data application. The current state as per our study suggests that the banks have vast amount of data stored in various systems by different department in their own way for keeping customer records. While working on some of the data for a bank we realized there is large amount of data which is fragmented, and we even got to know there is some data which is never used. With proper channels to save data in customized data format, it can be of immense value to the bank and also to its customers. We tried to collect data from various departments in the bank, starting from loan service department to various transactional data records for different branches to understand the functioning of the bank and how this data can be used to predict defaults, assess performance of banks in MSME sector.

# Detailed Methodology

## Data Extraction, Gathering, Cleaning and Pre-processing

### Data Extraction and Gathering

The dataset we received was in excel format from ***PSU Bank regarding the MSME sector delinquency and their default status***. The data was disparate covered the features in different excels. Our first task was to merge the data so that we could derive meaning out of it.

Data is in the form of:

* **MSME - Customer Financial Data:** Data consist of record of customer loan accounts specifically primary as well as secondary. It also delinquent status and Credit score of the customer.
* **MSME - Customer KYC data:** Customer details as per the customer ID, pan Aadhaar, mobile, passport, employment and date of birth
* **MSME - Loan Account Details:** Loan account details of the customer as per customer ID, account no., open date, sector, sector name category, industry ty, branch state.
* **MSME - Loan Account Status:** Target variables are present in this section
  + Delinquency – nonpayment of EMI in las six months as 0,1
  + Loan Default -whether the customer defaulted the account and let it become NPA by not paying installment for 3 months continuously.
* **MSME - Loan Sanction Details:** Consist of information regarding account no., disbursed amount and the security value purchased through the loan amount.

We merged the data and created two master files containing all the variables from all the above files on the basis of customer ID as unique key as:

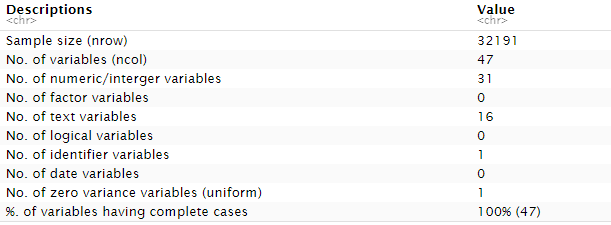
* **Data\_MSME\_merged\_Loan\_default:** Target variable present is whether the customer does loan default or not.
* **Data\_MSME\_merged\_Delinquent:** Target variable present is Delinquent – whether the customer has paid EMI on the respective dates or missed EMI payment date.

We also created a master data with all the variables as:

* **MSME Data\_Master**

### Raw data Exploratory data analysis.

Table 2: Data Understanding



We read the **MSME data master file** and tried to do some basic analysis so as we can get a basic overview of the data. Some basic analysis of raw data are as follows:

Table 3: Detailed Column wise overview of Data master file



***As from the above we can see there are no missing value shown in the data as the blanks are considered as strings by default. So, we drilled down further.***

### Missing Values

On drilling down further, we found that NA is considered as factor and some blank space are present in the data so initially data was showing no missing, so we modified the code to induce NA in the data. On further converting the missing string in the data as NA. **The total missing values in master file is 6707.**

We further also searched for NA along the variables for getting better understanding of the data. For more details on the code please refer to ***annexure section 8.5 code files (EDA)***.

### Data cleaning and variable datatype modification of the variables

As most of the fields are read as integer while importing so we converted them to factor as per business understanding.

* We started with the CNS score which is **credit score/rating for an individual customer or a firm.**
* As per business understanding we categorized CNS score into 6 different categories based on severity of risk with score ranging from 0 to 900 (900 is the lowest risk).
* Variable **perform\_CNS.score.Description** was having multiple categories which can be converted to simple levels so we feature engineered the variable.
* The final frequency counts that we got in the column perform\_CNS.score.Description are as below:



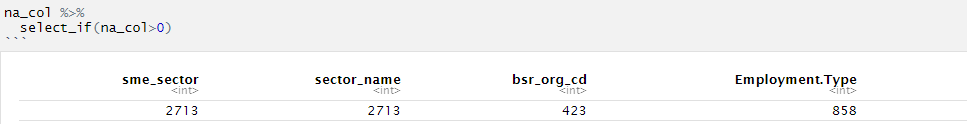
* Cleaning up SME category field as in some places category ‘Medium’ was showing as “Mediu’
* Converted the rest of the variables as factor as factor.
* Converted the date field [open\_dt, Date.of.Birth] in date format (‘DD-MM-YYYY’) as initially it was read as character then converted to factor then to a date format.
* We further omitted the NA’s and created the data set having proper date column and total rows now reduced to 28696.

### Missing value imputation

We imputed the missing value keeping our target variable in mind by two methods:

* Median Imputation as it will not be affected by the outliers in the four columns:

Table 4: Data Understanding - Missing value columns



* Imputed using K nearest neighbor (kNN) with k value as 5.

In KNN as result of imputation some extra variables are created on the basis of importance from which we removed the irrelevant columns and extracted the file to csv.

***Please refer to annexure section 8.5 code files for codes.***

## Exploratory Data Analysis - Explore and gain insights from a dataset.

The EDA process was meant to uncover patterns, trends and initial insights that could highlight relevant features for further investigation in the modeling process. Our initial hypothesis was that the features such as Annual Income, Interest Rate, Credit Grade and Loan Amount would have a bearing on whether the borrower defaulted.

So, after cleaning the data and imputation we received basic EDA insights about the data how is it structured, types of variables, missing data and if any variable is impacting target variable. Then we tried to understand our target variables where we got insight about delinquency and defaulter customers.

Post this we studied understanding of the each and every variable and the impact of it on target variables.

### Data Understanding

#### **Understanding the missing data, maximum and minimum values and frequency of the variable**

Please find the few examples mentioned below and for more details of the other variables please refer annexure ***section 8.2 – Detailed EDA***.

* ***VoterId Details***: As seen from the bar plot, we have observed that a smaller number of customers have shared the voderId details and yet there is no significant difference on the default rate.

Left bar shows the customers who have not submitted the voter ID details and right bar shows the customers who have shared.

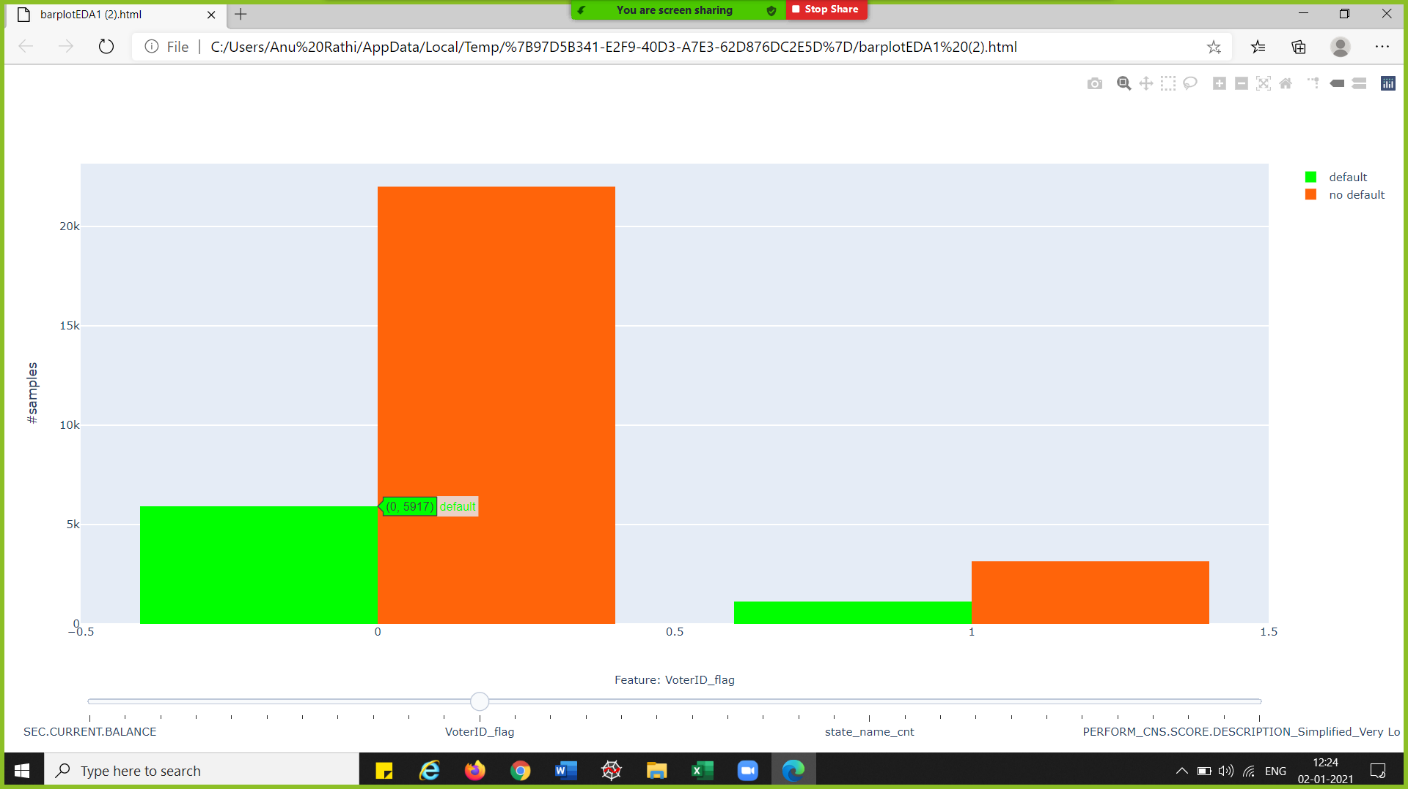


Figure 8: VoterID details of users

* ***Mobile Details of Users*:** As seen from the bar plot we can infer that most of the customers have provided the mobile details.

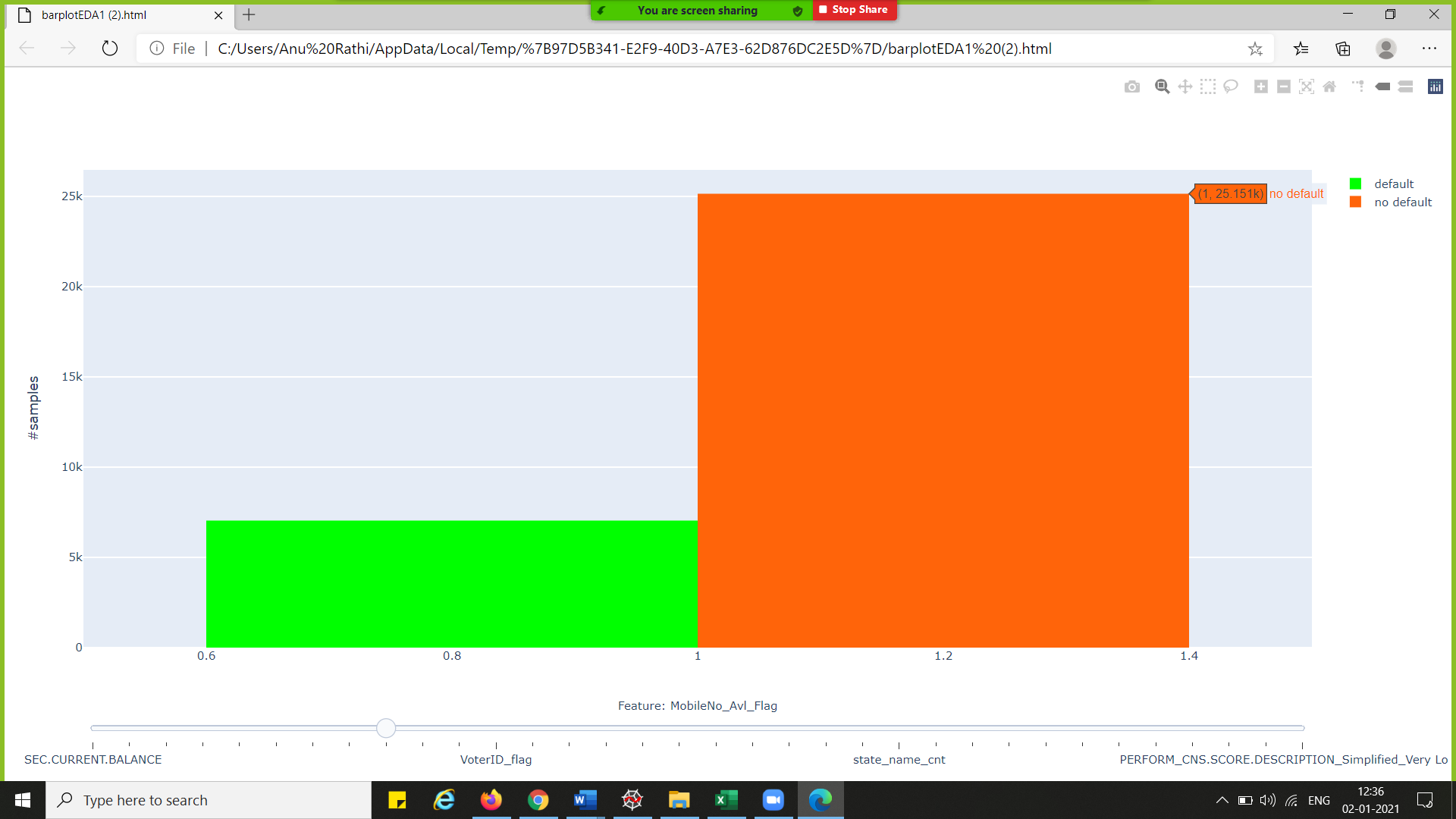


Figure 9: Mobile Details of User

Further exploration of categorical variable we found:

Most customer have Aadhar linked to account but 1n case of Pan, Voter ID, Driving License and Passport is not available.

* ***CNS Score Details:*** CNS Score is simple to CNS score simplified. and from the last bar blot we can imply that most of the customers have no credit history.

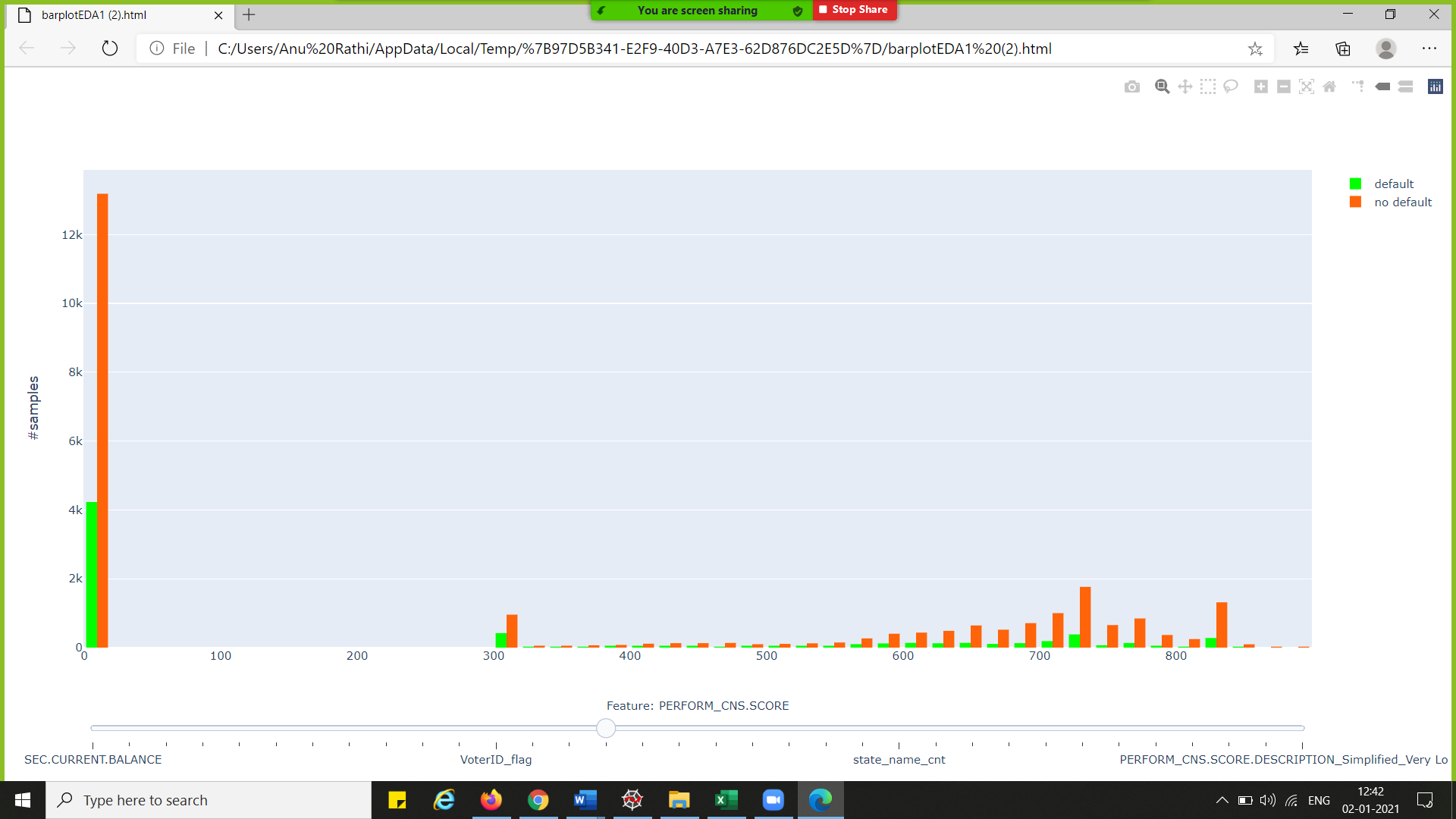


Figure 10: CNS Score Details

* ***Age Details of customers*:** The age of the customers taking loans lies mostly between 25 years to 35 years.

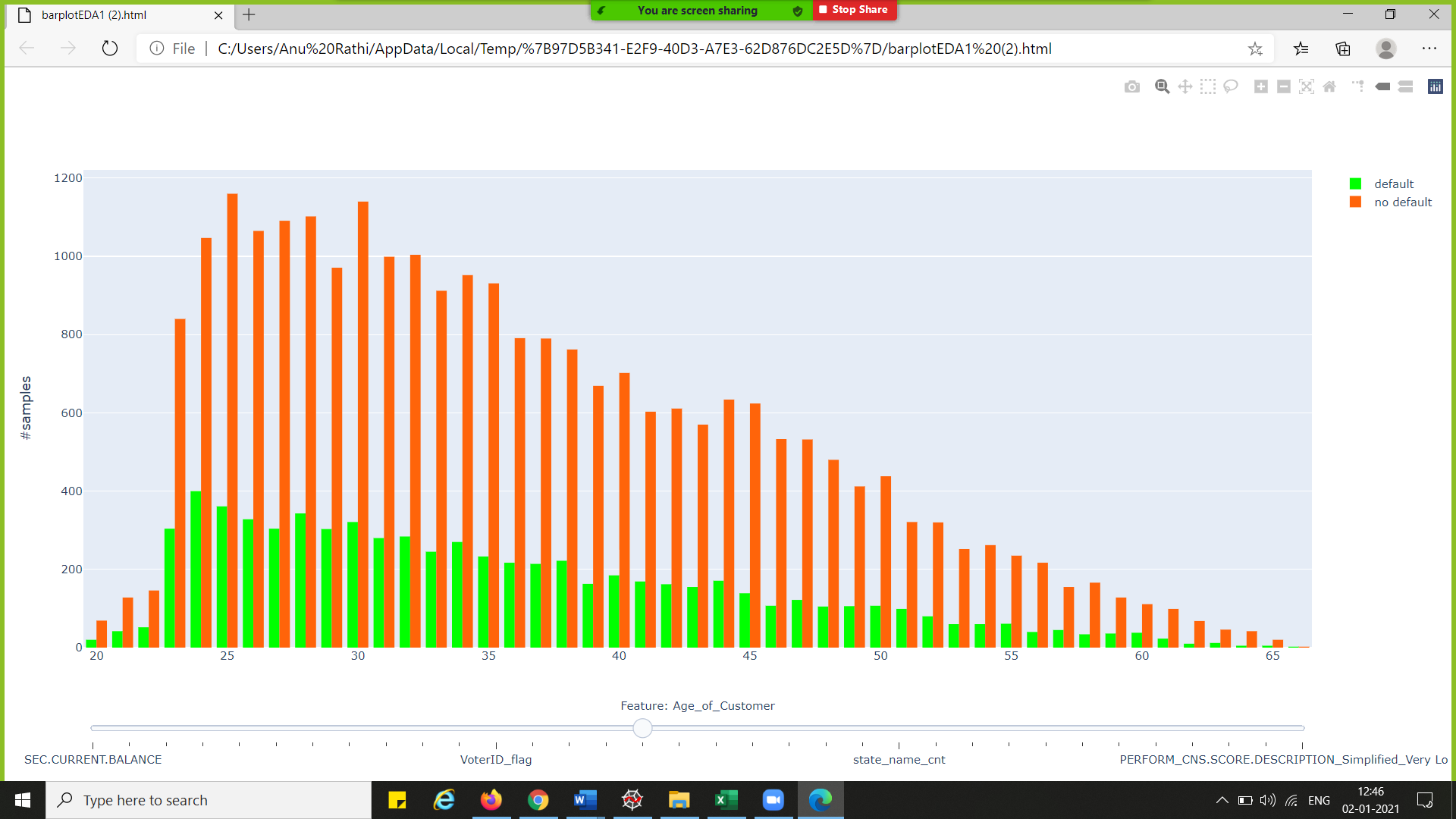


Figure 11: Age Detail of customers

* ***Perform CNS Score:*** As from the above boxplot we can see the 75% of the customers who are defaulters have low scores compared to non-defaulters. Which justifies better the score less are the chances the customer will default.



Figure 12: Perform CNS score

#### **Outliers understanding and treatment:**

We found a few outliers as shown in the figure below for LTV. Specifically, the customers having low LTV are the outliers. Similarly, for other important variables we found a few outliers but as this is a live data from the bank to maintain the integrity of the data, we kept the data as it is.

We further performed the (Min-Max) scaling. The cost of having this bounded range in contrast to standardization is that we end up with smaller standard deviations, which can suppress the effect of outliers.

For more details please refer annexure section 8.2 – Detailed EDA.

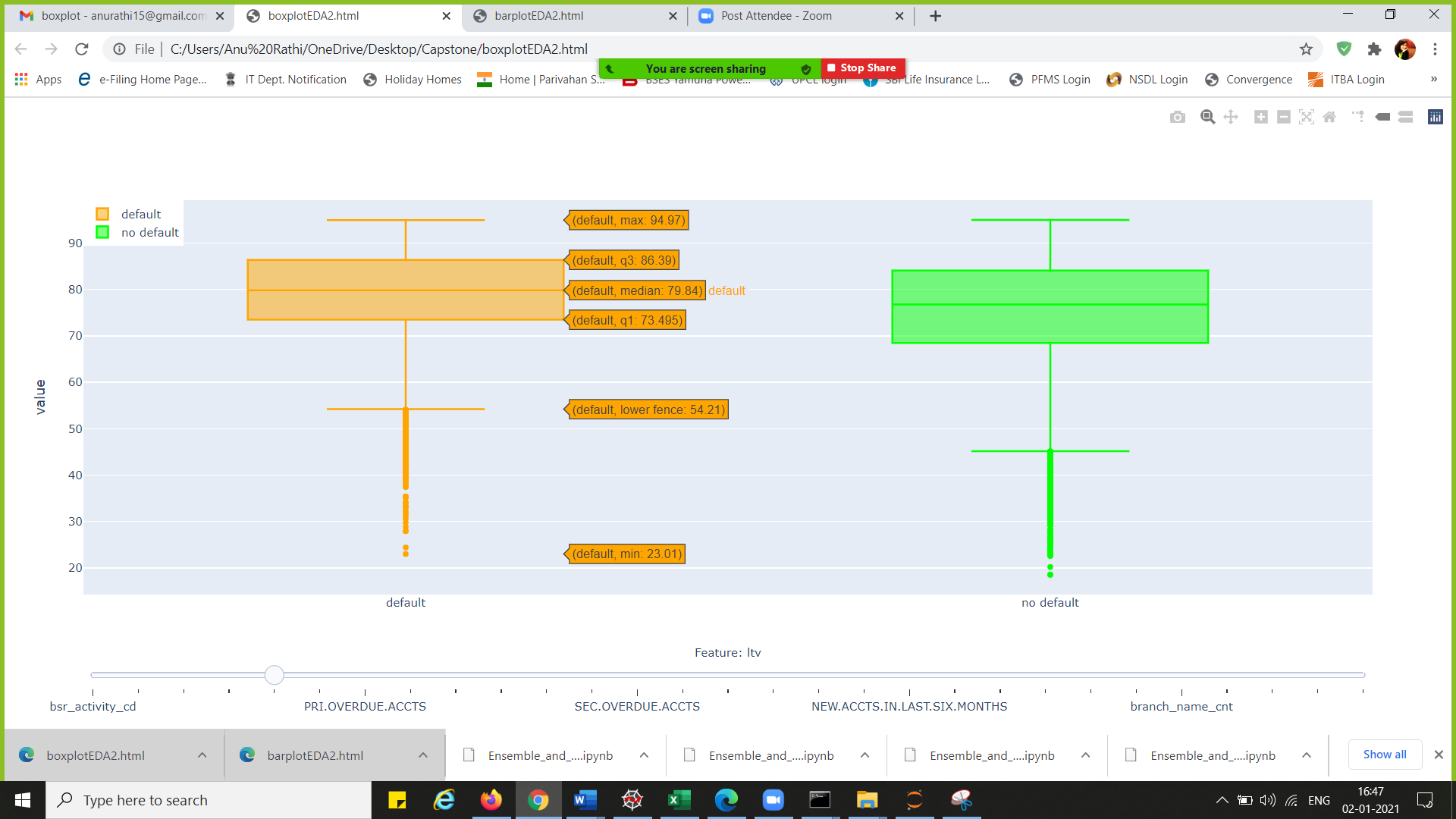


Figure : Box Plot for "ltv" variable

#### **Numerical and Categorical Association of the variables.**

We worked to find out the numerical and categorical association of different key variables as explained for LTV. LTV is one of the key variables in predicting the results.

LTV ratio is a number lenders use to determine how much risk they're taking on with a [secured loan](https://www.experian.com/blogs/ask-experian/secured-vs-unsecured-loans-what-you-should-know/). It measures the relationship between the loan amount and the market value of the asset securing the loan, such as a house or car.

For more details, please refer annexure section 8.2 – Detailed EDA.

**Numerical Association for the disbursed amount is highest which implies if LTV is high disbursed amount will be high as well and relatively Asset cost is the minimum.**

Figure : Numerical and Categorical association for "ltv"

**LTV categorical association with loan default is highest and minimum for delinquency and with 2 others.**

### Understanding the target variable

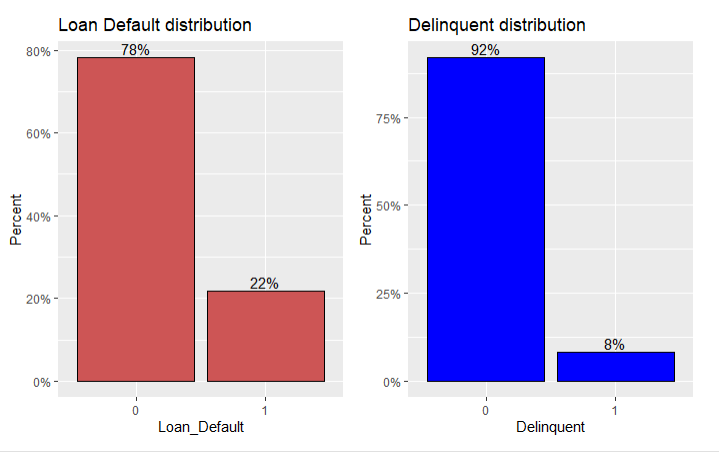


Figure 15: Target Variable Distribution

**Target variable are biased. As we can see 8% are delinquent customer while loan defaulters are 22%. So, most of the non-delinquent customer committed loan default.**

### Five-point Summary:

When executed on raw data the variables were considered as numeric and hence the below values are obtained. For example: BSR code was UNKNOWN and to identify the unknown values we passed the specific values such as 99999.00. Based on the O/P we received we engineered the data.

Table 5: Five-point summary details

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | Mean | std | min | 25% | 50% | 75% | max | Understanding |
| sme\_sector | 32191 | 1.78 | 0.42 | 1 | 2 | 2 | 2 | 2.00 | Most of the data belongs to service sector as the mean is 1.78. |
| bsr\_activity\_cd | 32191 | 53753.60 | 22637.27 | 1101 | 51201 | 52319 | 60203 | 99999.00 | BSR code are given by banks based on the activities for the loan purpose. 99999.00 is given to the companies whose activities are not mentioned. |
| bsr\_org\_cd | 32191 | 11.68 | 1.28 | 1 | 11 | 11 | 12 | 25.00 | Most of the data lies b/w bsr\_org\_cd 11 and 12 hence most of the loans are given for government organisation and individual loans. |
| disbursed\_amount | 32191 | 515182.10 | 129127.80 | 139900 | 448490 | 513030 | 570130 | 9905720.00 | Average disbursed amount by the bank is 5 lakhs and more. |
| asset\_cost | 32191 | 713149.80 | *191733.50* | 370000 | 627000 | 667150 | 730555 | 16289920.00 | Asset purchased through loan has average cost of above 7 lakhs and median is above 6 lakhs. |
| ltv | 32191 | 75.61 | 11.50 | 18.51 | 69.64 | 77.68 | 84.61 | 94.99 | Most of the loans are secured at 75% risk. LTV -Loan to value: Relationship b/w loan amount and market value of the asset. |
| PRI.ACTIVE.ACCTS | 32191 | 1.12 | 2.04 | 0 | 0 | 0 | 1 | 52.00 | Most of the customers are new customers as the Median value is zero. |
| PRI.OVERDUE.ACCTS | 32191 | 0.17 | 0.58 | 0 | 0 | 0 | 0 | 23.00 | Since there are new customers hence the overdue amount is 0. |
| PRI.SANCTIONED.AMOUNT | 32191 | 2247365.00 | 10543730.00 | 0 | 0 | 0 | 711645 | 585530000.00 | Avg. sectioned amount is more than 2 lakhs. |
| PRI.DISBURSED.AMOUNT | 32191 | 2229950.00 | 10524480.00 | 0 | 0 | 0 | 691335 | 585530000.00 | As the sanctioned and disbursed amounts are almost same so we can see that mean amount is also 2 lakhs. |
| SEC.NO.OF.ACCTS | 32191 | 0.09 | 0.83 | 0 | 0 | 0 | 0 | 46.00 | On an avg. 9% of the customers have secondary account. |
| SEC.ACTIVE.ACCTS | 32191 | 0.04 | 0.42 | 0 | 0 | 0 | 0 | 26.00 | Only 4% of the sec accounts are active. 44% of the secondary accounts are active (0.04/0.09). |
| SEC.OVERDUE.ACCTS | 32191 | 0.01 | 0.13 | 0 | 0 | 0 | 0 | 6.00 | Only 1% of the secondary accounts are overdue. Also there are customers having 6 overdue accounts which is high risk. |
| SEC.SANCTIONED.AMOUNT | 32191 | 101805.90 | 1733274.00 | 0 | 0 | 0 | 0 | 110000000.00 | Avg. amount for the secondary sanctioned amount is more than 1 lakh. |
| SEC.DISBURSED.AMOUNT | 32191 | 99753.19 | 1718551.00 | 0 | 0 | 0 | 0 | 110000000.00 | Avg. amount for the secondary sanctioned amount is almost equal to 1 lakh. |
| PRIMARY.INSTAL.AMT | 32191 | 149220.90 | 1725651.00 | 0 | 0 | 0 | 21760 | 154204100.00 | Avg. instalment is app 1.5 lakhs. |
| SEC.INSTAL.AMT | 32191 | 8590.58 | 347202.70 | 0 | 0 | 0 | 0 | 41709010.00 | Avg. instalment is app 9K. |
| NEW.ACCTS.IN.LAST.SIX.MONTHS | 32191 | 0.39 | 0.96 | 0 | 0 | 0 | 0 | 20.00 | As the values are 0 hence these are the new customers. |
| DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS | 32191 | 0.10 | 0.40 | 0 | 0 | 0 | 0 | 11.00 | Most of the accounts are regulated for last 6 months only 10% has gone delinquent. |
| MobileNo\_Avl\_Flag | 32191 | 1.00 | 0.00 | 1 | 1 | 1 | 1 | 1.00 | 1 indicates bank is having mobile contact details of all customers. |
| Aadhar\_flag | 32191 | 0.85 | 0.36 | 0 | 1 | 1 | 1 | 1.00 | 85% of the customers has given their aadhar details. |
| PAN\_flag | 32191 | 0.08 | 0.27 | 0 | 0 | 0 | 0 | 1.00 | 8% of the customers has given their PAN details. |
| VoterID\_flag | 32191 | 0.13 | 0.34 | 0 | 0 | 0 | 0 | 1.00 | 13% of the customers has given their VoterID details. |
| Driving\_flag | 32191 | 0.03 | 0.17 | 0 | 0 | 0 | 0 | 1.00 | 3% of the customers has given their Driving\_flag details. |
| Passport\_flag | 32191 | 0.00 | 0.05 | 0 | 0 | 0 | 0 | 1.00 | Most of the customers do not have passport details. |
| PERFORM\_CNS.SCORE | 32191 | 299.25 | 339.70 | 0 | 0 | 15 | 681 | 890.00 | We have average score of 300 which is not good. |
| Age\_of\_Customer | 32191 | 36.00 | 9.73 | 20 | 28 | 34 | 43 | 66.00 | Average customer age is 36 and median is 34 years. |
| Age\_of\_Loan | 32191 | 0.72 | 0.48 | 0 | 0 | 1 | 1 | 7.00 | Most of the loans are new loans. Therefore, age is less than 1 year. |
| scheme\_cd\_cnt | 32191 | 13984.17 | 9594.95 | 1 | 1384 | 20988 | 20988 | 20988.00 | Encoded feature based on count frequency. |
| industry\_name\_cnt | 32191 | 3737.08 | 4601.01 | 1 | 187 | 1018 | 10369 | 10369.00 | Encoded feature based on count frequency. |
| branch\_name\_cnt | 32191 | 21.40 | 19.22 | 1 | 9 | 16 | 28 | 136.00 | Encoded feature based on count frequency. |
| district\_name\_cnt | 32191 | 299.69 | 290.83 | 1 | 85 | 214 | 428 | 1262.00 | Encoded feature based on count frequency. |
| region\_name\_cnt | 32191 | 760.70 | 262.02 | 323 | 535 | 781 | 910 | 1443.00 | Encoded feature based on count frequency. |
| state\_name\_cnt | 32191 | 4930.87 | 4892.47 | 1 | 866 | 1783 | 11469 | 11469.00 | Encoded feature based on count frequency. |
| open\_dt\_day\_cnt | 32191 | 5659.92 | 1150.67 | 3163 | 5481 | 5626 | 6353 | 7055.00 | Day of disbursement was calculated and encoded feature based on count frequency. |
| sector\_name\_Services | 32191 | 0.78 | 0.42 | 0 | 1 | 1 | 1 | 1.00 | 78% of the loans belong to service sector. |
| sme\_category\_MICRO | 32191 | 0.92 | 0.27 | 0 | 1 | 1 | 1 | 1.00 | 92% are the loans for MICRO segment of MSME. |
| sme\_category\_SMALL | 32191 | 0.07 | 0.26 | 0 | 0 | 0 | 0 | 1.00 | 7% are the loans for SMALL segment of MSME. |
| module\_id\_cdcc | 32191 | 0.15 | 0.36 | 0 | 0 | 0 | 0 | 1.00 | 15% of the loans are for the current account which is for the day to day operations. |
| Employment.Type\_Self employed | 32191 | 0.53 | 0.50 | 0 | 0 | 1 | 1 | 1.00 | 53% of the customers are self-employed. |
| PERFORM\_CNS.SCORE.DESCRIPTION\_Simplified\_Low Risk | 32191 | 0.08 | 0.27 | 0 | 0 | 0 | 0 | 1.00 | Encoded feature based on CNS score with 8% having a low risk. |
| PERFORM\_CNS.SCORE.DESCRIPTION\_Simplified\_Medium Risk | 32191 | 0.05 | 0.22 | 0 | 0 | 0 | 0 | 1.00 | Encoded feature based on CNS score with 5% having a medium risk. |
| PERFORM\_CNS.SCORE.DESCRIPTION\_Simplified\_Not Scored | 32191 | 0.54 | 0.50 | 0 | 0 | 1 | 1 | 1.00 | Encoded feature based on CNS score with 54% having no score representing that they new customers. |
| PERFORM\_CNS.SCORE.DESCRIPTION\_Simplified\_Very High Risk | 32191 | 0.05 | 0.21 | 0 | 0 | 0 | 0 | 1.00 | Encoded feature based on CNS score with 5% having a high risk. |
| PERFORM\_CNS.SCORE.DESCRIPTION\_Simplified\_Very Low Risk | 32191 | 0.23 | 0.42 | 0 | 0 | 0 | 0 | 1.00 | Encoded feature based on CNS score with 23% having a very low risk. |

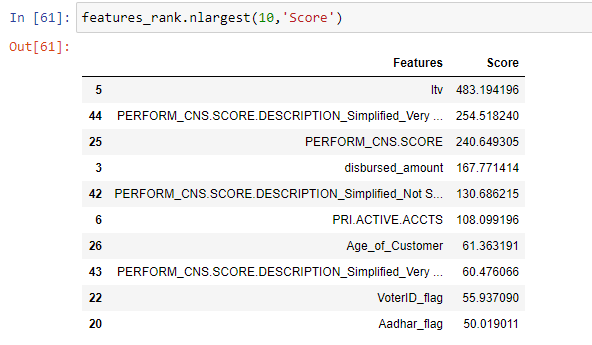
## Feature Importance Understanding

We further deep dive into the variable with impact our target variable. Using decision and random forest selection, we found out the importance of variable.

***Feature Importance of all the variables is mentioned in the annexure section 8.2***

### Feature importance using decision tree and its score

Table 6: Top Identified Features based on decision tree



**Our Observation:**

* From above it is evident the credit score of the customer reflects most of the aspect of type of customer.
* LTV is loan to value of asset ratio describes the intention of customer whether he is in a practical need loan and an honest customer
* Number of active accounts tells about the credits portfolio and future indication can be sorted out while giving the in practical in banks.
* Age, VoterID and Aadhaar gives relevant emphasis of KYC plays an important role in advance. Better to you know the customer the better chances that you can predict the delinquency and loan default of a loan account.

### Feature importance using random forest technique

Further using Random Forest technique, we tried to understand the importance of the features This technique gives you a score for each feature of your data, the higher the score more relevant it is.

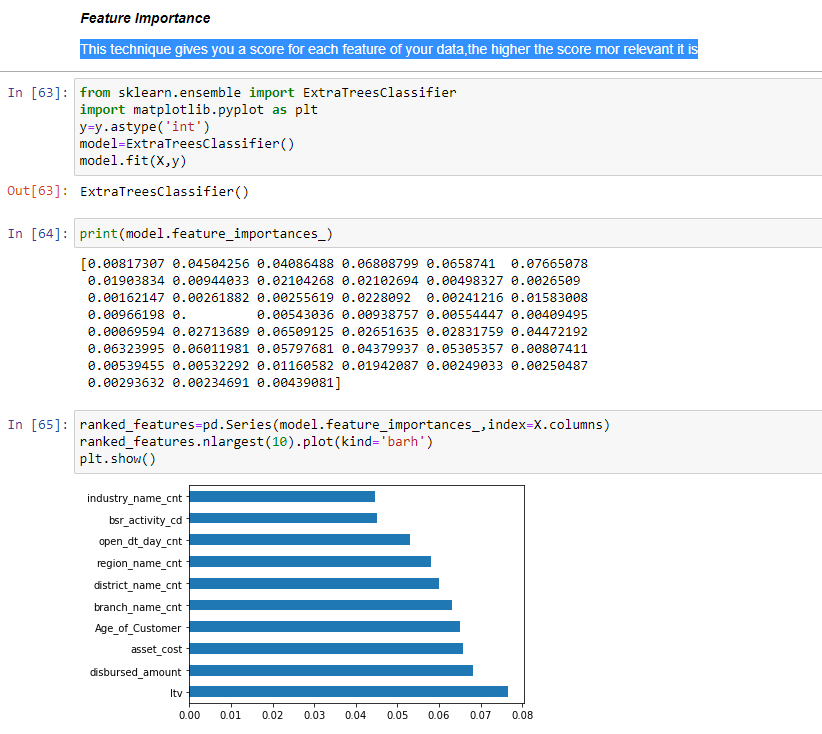


Figure 16: Feature Importance score based on random forest

**Observation:**

Top 2 features playing an important role in determining the delinquency and default in MSME sector are LTV and disbursed amount.

As business understand disbursed amount and sanctioned amount will be corelated and data correlation matrix proves it. The same goes for sec sanction and disbursal as it was evident from the data.

## Feature Engineering

### Converting relevant numeric fields to factor

Earlier in the code have converted the numeric field to factor as per data understanding.

Specifically, variables like:

* data\_msme
* Aadhar\_flag
* PAN\_flag
* VoterID\_flag
* Driving\_flag
* Passport\_flag
* Delinquent (target)
* sme\_category

### One hot Encoding

Since we have some categorical variables for the analysis and the machine learning algorithms doesn't take categorical and string variables directly, we have to create dummy variables for them. We can either encode them using label encoder available for R and python both, but it would be wrong in our analysis since a lot of these variables have multiple categories. Just using weights can cause discrepancies in the algorithm. Instead, we will one hot encode these so that we have a 1 wherever that category turns up and 0 otherwise. This will also create separate columns for each level of category.

We tried one hot encoding on the imputed data, and it resulted in giving 4019 columns based on factors level present. Also we have dropped some columns from the data which are not required in the final modelling such as ConsumerID, account no., account name) and redundant data PERFORM\_CNS.SCORE.DESCRIPTION as we have earlier engineered the new column PERFORM\_CNS.SCORE.DESCRIPTION\_Simplified

**Further for efficient encoding we feature engineered our date variables open\_dt and Date.of.Birth** As our data was reading as date but while converting in one hot encoding it’s showing 325 and 9986 labels in open\_dt and Date.of.Birth respectively.

Based on this we were also able to add 2 new columns in the main data Age\_of\_Customer and Age\_of\_Loan.

**Count frequency encoding for variable with factor levels greater than 10.**

Firstly, extracted categorical features:

**Categorical feature with factor levels greater than 10:**

1. 'scheme\_cd',
2. 'industry\_name',
3. 'branch\_name',
4. 'district\_name',
5. 'region\_name',
6. 'state\_name'

So, extracted the level count and mapping through dictionary in the master data.

Advantage of this technique helps reduce dimensionality while one hot encoding. So, it gives ease of not increasing the feature space.

On the downturn, only disadvantage will be providing same weight to the categories is there.

Checked for remaining categorical vector levels are less than 10.

Finally, we can do the one hot encoding.

So, in all we have 49 total variables.

***Code files are available in annexure section 8.5 Code Files.***

**Independent Variables: 47**

**Dependent variables: 2**

## Data splitting

We divided the data into three parts basically:

* Training: Consisting of 70% of the data
* Testing: Containing 25% of the data
* Validation: Consisting of 5% of the data.

## Modeling Technique

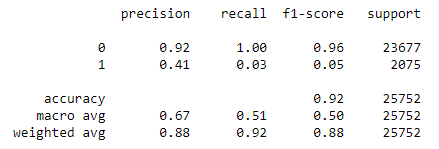
The modeling process also takes major amount of time particularly as we are looking to attempt different features, models, parameters etc. As this is a classification problem, we decided to narrow down the focus on Logistic Regression for delinquency and following mentioned techniques for loan default.

* Logistic Regression
* Decision Tree
* Random Forest
* SVC
* KNN
* CatBoost
* Adaboost
* Light gradient boosting
* Extreme gradient boosting
* Naïve Bayes
* ANN

### Modeling for delinquency

#### For modeling we started working with Logistic regression where we received the following scores on the data were obtained:

Table : Delinquency Model Performance

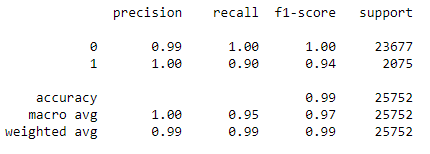


Observation: As we can see the model is performing well for predicting whether the customer will go delinquent or not pretty well.

When we performed logistic regression on the unscaled data accuracy for delinquent customers was coming very low.

#### On further enhancing the model and scaling the data we were able to get good results.

Table : Delinquency model performance for scaled data



Using tuned parameter and scaled data (min max scaling used to normalize the data) in logistic regression prediction power for delinquent customers increased to 0.94.

***Code files are available in annexure section 8.5 Code Files.***

### Modeling for Loan Default

We tried different models and analyzed the accuracy to obtain the best prediction model and compare the results.

In the below tables we have calculated various the performance measures to understand how the model is working to predict the loan defaulters.

The performance measures we have used are as follows:

**AUC**: AUC signifies the chance that model will be able to distinguish between defaulters and non-defaulters.

**Precision**: The precision is the proportion of relevant results in the list of all returned search results.

**F-Score**: The F-score, also called the F1-score, is a measure of a model's accuracy on a dataset. The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

**Recall**: The recall is the ratio of the relevant results returned by the search engine to the total number of the relevant results that could have been returned.

Table : Model Performance Interpretation

|  |  |  |
| --- | --- | --- |
| S. No. | Evaluation Metric | Interpretation |
| 1 | Sensitivity | What percentage of all 1's i.e. default is correctly predicted? |
| 2 | Specificity | What percentage of all 0's i.e. customer is not default were correctly predicted? |
| 3 | Precision | What percentage of predicted 1's i.e. default is correct? |
| 4 | Recall | What percentage of all 1's i.e. default is correctly predicted? |
| 5 | F1 Score | A combination of Precision and Recall |
| 6 | AUROC | Model's true performance considering all possible probability cutoffs. |
| 7 | Gini Coefficient | How the model exceeded random predictions in terms of ROC |
| 8 | KS Statistic | Used to decide how many customers to target |

#### Logistic Regression

Logistic train roc-auc: 0.58

Logistic test roc-auc: 0.57

Table : Logistic Regression Model Performance

|  |  |
| --- | --- |
| Train | Test |
|  |  |

#### Decision Tree

Decision Tree train roc-auc: 1.0

Decision Tree roc-auc: 0.51

Table : Decision Tree Model Performance

|  |  |
| --- | --- |
| Train | Test |
|  |  |

#### Adaboost test

Adaboost train roc-auc: 0.65

Adaboost test roc-auc: 0.64

Table : Adaboost Model Performance

|  |  |
| --- | --- |
| Train | Test |
|  |  |

#### Random Forest

RF train roc-auc 1.0

RF test roc-auc: 0.63

Table : Random Forest Model Performance

|  |  |
| --- | --- |
| Train | Test |
|  |  |

#### SVC

SVC train roc-auc: 0.56

SVC test roc-auc: 0.51

Table : SVC Model Performance

|  |  |
| --- | --- |
| Train | Test |
|  |  |

#### Naïve Bayes

Naive Bayes train roc-auc: 0.59

Naive Bayes test roc-auc: 0.54

Table : Naive Bayes Model Performance

|  |  |
| --- | --- |
| Train | Test |
|  |  |

#### KNN

KNN train roc-auc: 0.80

KNN test roc-auc: 0.54

Table : KNN Model Performance

|  |  |
| --- | --- |
| Train | Test |
|  |  |

#### CatBoost Classifier

Table : CatBoost Model Performance

|  |  |
| --- | --- |
| Train | Test |
|  |  |

#### LGBM

Light GBM train roc-auc: 0.82

Light GBM test roc-auc: 0.64

Table : LGBM Model Performance

|  |  |
| --- | --- |
| Train | Test |
|  |  |

#### Extreme Gradient Boosting Modeling

Extreme GBM with Grid search train roc-auc: 0.91

Extreme GBM with Grid search test roc-auc: 0.61

Table : EGBM Model Performance

|  |  |
| --- | --- |
| Train | Test |
|  |  |

#### Ensemble of all the models and predicted the loan defaulters based on different techniques.

Calculated the prediction of different models and concatenated means of all the techniques and then extracting the best 5 predictions and threshold to predict the defaulters.

Ensemble test roc-auc: 0.63

Table : Collective Thresholds for models

|  |  |
| --- | --- |
| Accuracy | F1 Score |
|  |  |

## Component Reduction

### Principal Component Analysis

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

We can see from the graph that at least 30 components will be required to explain 95% variance and 22 componets for 80% variance hence considering the large number of components, we are not going ahead to use the PCA technique.

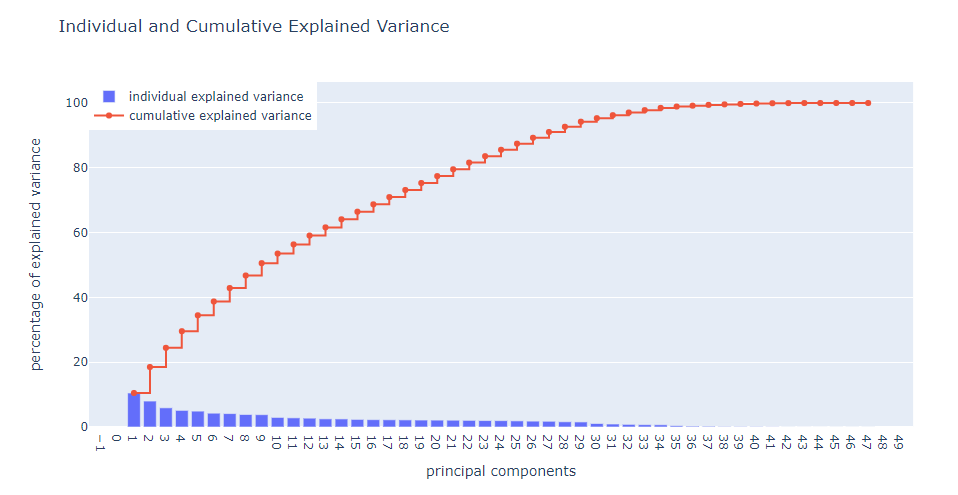


Figure 14: Scree plot capturing the variance by Principal Components

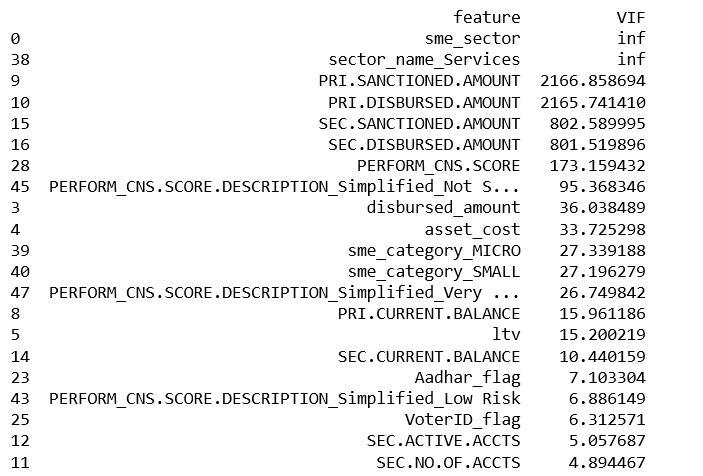
### Component Selection using VIF

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple variables. There are some guidelines we can use to determine whether our VIFs are in an acceptable range. A rule of thumb commonly used in practice is if a VIF is > 10, you have high multicollinearity. In our case, with values around 1, we are in good shape.

Please find a few sample values that we have got based on the VIF. In our analysis we have included a few variables even after the higher values in the calculation as they were important variable as per business perspective:

* disbursed\_amount
* asset\_cost
* sme\_category\_MICRO
* sme\_category\_SMALL
* PERFORM\_CNS.SCORE.DESCRIPTION\_Simplified\_Very Low Risk
* PRI.CURRENT.BALANCE
* ltv
* SEC.CURRENT.BALANCE

Table 21: VIF sample values



For details list of variables and their values please refer to ***annexure section 8.4- Detailed VIF values.***

## Modeling comparison

Now, in order to decide with which model(s) we will proceed further, we used the confusion matrix outputs of these models to evaluate their accuracy. A confusion matrix is a table that allows us to visualize the performance of a classification model. One can also use the information in it to calculate measures that can help us determine the usefulness of the model. As our primary focus here is to predict the ‘loan defaults’, the most important measure that we focused on was f1 score. As we can see in the image below, the Random Forest and XGB model are the top two performers.

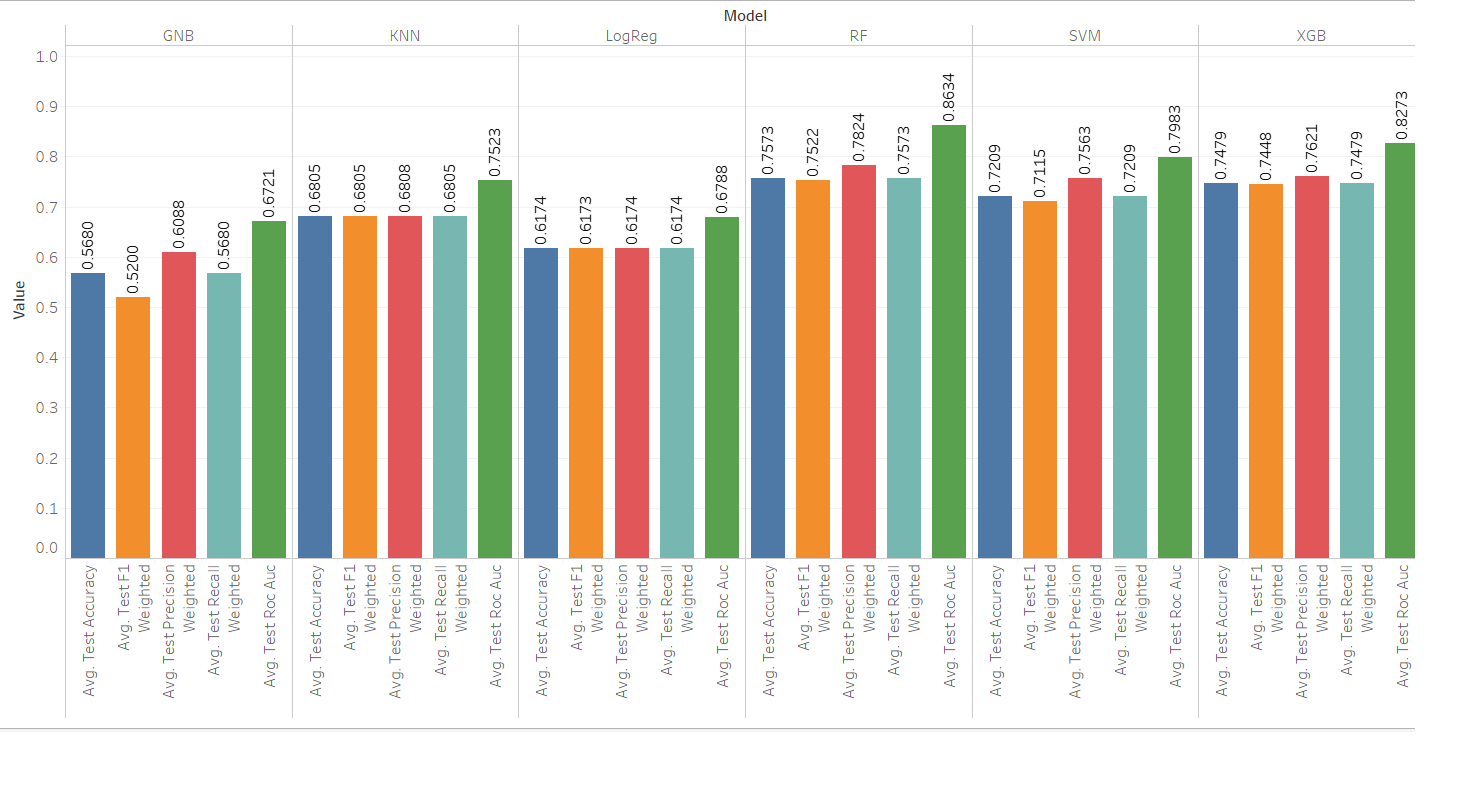
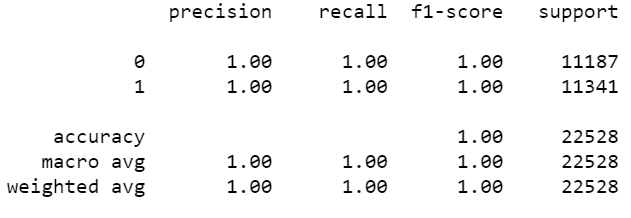


Figure : Model comparison based on performance parameters

We also tried to engineer the data with Smote to further improve our accuracy levels for prediction of defaulters. The output we got was biased and as we are using live data we wanted to decided not to use this technique.

Table : Model performance after applying SMOTE technique



***Finally, we have selected Random Forest and XG boost for deploying our application.***

## Model interpretation through LIME:

LIME is model-agnostic, meaning that it can be applied to any machine learning model. The technique attempts to understand the model by perturbing the input of data samples and understanding how the predictions change.

The output of LIME is a list of explanations, reflecting the contribution of each feature to the prediction of a data sample. This provides local interpretability, and it also allows to determine which feature changes will have most impact on the prediction.

From the below image we can see that out of 10 variables 7 are contributing towards determining the non-defaulters whereas 3 are contributing towards the defaulter identification. The mentioned variables are supported by feature importance as well. We used Random Forest technique to check the feature importance on the prediction.

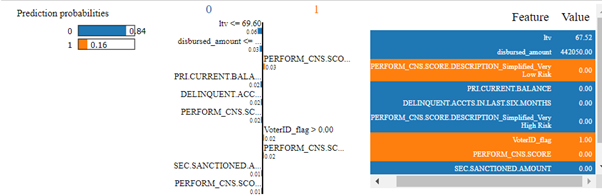


Figure : Model Interpretation with LIME

## Deployment

The most important part while deploying an analytical solution is to bridge the gap between the IT team and data science team. The main objective is to ensure that the final solution is ready to be used within the operational environment and that end users have all the required tools to act upon the analytical insights discovered during the development phases of the project.

### User Interfaces for banking enterprise

We have created an end-to-end user-friendly interface front end application which will be deployed as a web application.

Banks will have to provide top 4 mentioned details of the customer to predict if the customer will default or not. On the basis of which the bank can decide if further details are required of the customer and if the loan can be given.

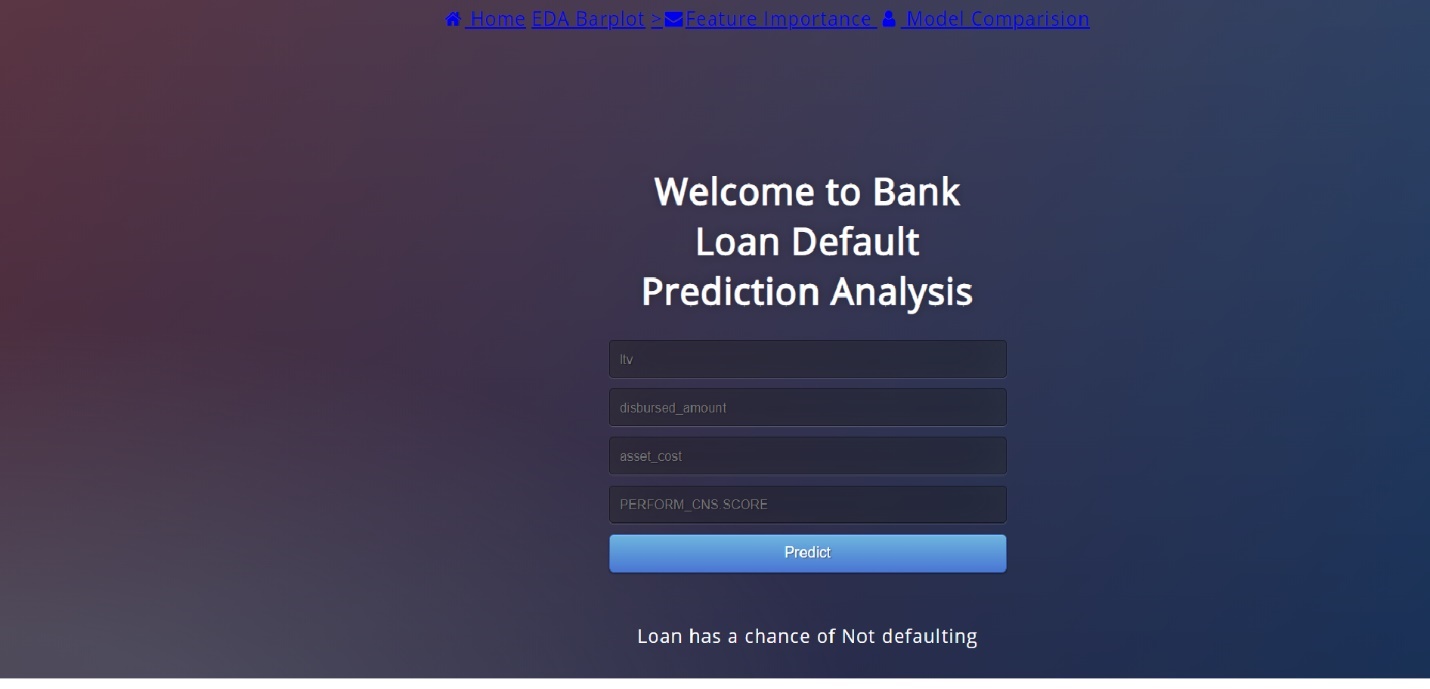


Figure : Sample Screen shot for banking enterprise user interface

We plan to improvise the application by including more features and data which will help the banks to understand the customers financial status and to better predict loan defaults to reduce NPAs.

In the future enhancements we also plan to create an application where the end users can themselves input their data in the required and can ensure their eligibility for the loan.

# Achievements

## Achieved in the current

We were able to achieve an accuracy of **94%** for predicting the delinquency of the customer and around **75%** for predicting the defaults much early in the lending cycle. There were also some analytics taken during the study which helped the bank to identify various other issues and how the bank can handle those.

The key achievements which can help the bank to cater to the current challenges is as below:

* **Consolidation of data spread across departments to achieve the analytical needs of the bank.** There was a lot of data available with the bank but as the departments are working in silos, they were not able to identify the key insights by the individualistic approach. We got all the data from different departments which we consolidated to bring out the hidden insights after processing and EDA analysis of the data. The insights were more around, who are people who tend to default, if there is any account which will be delinquency in coming few months, the insights on user profile where the instalments are getting differed, details on specific industries and regions.
* **Deep dived into various aspects of loans given by the banks to MSME sectors and the important measures that are required while giving the loan.** A common fallacy is to take the easy route and implement an out of box generic early warning application that will not consider the specific circumstances of the bank and thus not deliver the desired business impact in NPA reduction. A standalone warning system application can further complicate the IT application landscape of the bank and undermine its NPA management program.
* **Early detection of a default in case of these kind of loans helps bank reduce their NPA and overall reducing the impact on the Indian economy.** Non-performing assets are one of the biggest challenges facing the global banking system, and particularly Indian banks. The extent of the challenge for nationalized banks is that non-action is no longer an option. This issue is likely to get worse due to the overall economic slowdown impacting most customer segments across banks’ portfolio like MSMEs (micro, small and medium enterprises), large corporates, and agriculture to name a few.
* **Factors while giving the loan and information to be taken from the for effective performance of a loan.** While banks across regions have historically focused on traditional data sources, we have observed that recently more successful banks have differentiated themselves by leveraging non-traditional and powerful data that exists both within and outside their systems. For example; instances of bounced checks in customer’s deposit account, advance tax deposit receipts, stock market data, and more.
* **Better scrutiny process for the predicted default**. While studying the pre-default behavior of customers, it is important to identify key data attributes that show correlation with default behavior or stress scenario, for example a small company with limited liquidity facing decline in sales. This can help define scrutiny process further in the cycle and their relevance to the bank.
* **Constantly monitoring** the already given loans and its health for preventing it into going to NPA.
* **Sector wise planning** - Certain sectors need to be monitored more closely than other. The identified concerned sectors should be dealt with extreme caution and extensive due diligence should be carried out before sanctioning anything.

## Future Enhancements

A few points that we think can add to the benefits of our project as future enhancement are:

* Creation of a Central Data repository for all corporate borrowers of all banks would go a long way in establishing coordination and provide a way to track defaulters, thus restricting their entry into the system.
* A one stop solution to be for the credit team which would cater to specific industries and monitor new applications and their status quo, subsequently raising a red flag if any anomalies are found.
* Anomaly detection in regular transactions of the users
* Enablement of automatic alerts and notifications in case of predicted delinquency, default of anomaly in the transaction or records of the user.
* Generation of automatic periodic reports
* OSINT analysis of the applicants/ organization.
* Link analysis of the user to identify intestinal frauds early.

# Recommendations and Conclusion

Based on the analysis we could conclude that the delinquency and defaults can be predicted much early in the cycle and additional cost of maintaining these account and recovery cost can be avoided if the right measures are taken.

The created tool gives high accuracy for prediction of defaults based on the available information with the bank. The bank can start working of consolidation of data at single repository which can help in deploying the created tool in the real live environment faster. We have also shared our understanding of the future developments in the solution for which an early start can be really helpful to create and implement the solutions early.

Some other suggestion and recommendation could be:

The banks should put in place proper loan review mechanism for large value loans with following objectives:

* to identify promptly loans which develop credit weaknesses and initiate timely corrective action;
* to evaluate portfolio quality and isolate potential problem areas;
* to provide information for determining adequacy of loan loss provision;
* to assess the adequacy of and adherence to, loan policies and procedures, and to monitor compliance
* to provide top management with information on credit administration, including credit sanction process.
* Seizing the assets of willful defaulters with the assistance of seizing agencies.
* banks should perform the credit audit to review the sanction process and status of post sanction processes/procedures.
* Demanding borrowers to produce additional security, where tools are predicting the borrower as defaulter
* Enquiring borrowers to know if they have availed loans from other banks
* Demanding highly liquid security from borrowers,
* Maintenance of loan to value ratio at a satisfactory level.
* Monitoring the borrowers to ascertain whether the borrowed funds are utilized for productive purpose.

# Bibliography

* RBI Circular: RBI/2018-19/100 DBR.No.BP.BC.18/21.04.048/2018-19 Dated: 1st January 2020.
* Banerjee, R., Verma, D. and Jaiswal, B. (2018), “Non-performing assets: a comparative study of the
* Indian commercial banks”, International Journal of Social Relevance and Concern, Vol. 6 No. 2
* Dey, S. (2018), “Recovery mechanisms of non-performing assets in Indian commercial banks: an
* empirical study”, NSOU Open Journal, Vol. 1 No. 2.
* Kumar, S. (2018), “A study on Non-Performing assets of Indian banks: Trend and recovery”,
* International Journal of Electronics, Electrical and Computational System, Vol. 7 No. 3
* Kumar, R.B., Subba Rao, B.V.S.S. and Kusuma, G.D.V. (2018), “Genesis for increase of NPAs in Indian
* banks – an empirical analysis”, Journal of Banking and Finance Management, Vol. 1 No. 1  Mishra, M.K. and Pawaskar, J.R. (2017), “A study of non-performing assets and its impact on banking sector”, Journal for Research, Vol. 3 No. 1
* Sharma, G. (2018), “Unmasking the ordeal of growing NPAs in Indian banking sector”, International
* Journal of Business and Management Invention, Vol. 7 No. 7
* <https://www.thehindubusinessline.com/opinion/columns/s-murlidharan/npa-crisis-prevention-is-the-cure/article30551370.ece>
* <https://www.ndtv.com/business/rbi-npa-norms-rbi-issues-revised-circular-for-resolution-of-stressed-assets-2049736>
* <http://sudeep.co/data-science/Understanding-the-Data-Science-Lifecycle/>
* <http://14.139.58.147:8080/jspui/bitstream/123456789/339/9/chapter%206.pdf>
* <https://www.accenture.com/cr-en/~/media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Industries_6/Accenture-Effective-Early-Warning-for-Managing-Non-performing-Assets.pdf>
* <https://www.moneycontrol.com/news/business/npas-of-psbs-stands-at-rs-7-27-lakh-crore-4897741.html>
* <https://www.spssanalyticspartner.com/blog-post-e2e-sustain-deployment-in-advanced-analytics-projects/>
* <https://economictimes.indiatimes.com/small-biz/sme-sector/rbi-extend-one-time-restructuring-scheme-for-msme-loans-to-december-31-2020/articleshow/73977674.cms>
* <https://www.rbi.org.in/Scripts/BS_CircularIndexDisplay.aspx?Id=11445>
* <https://ijrar.org/papers/IJRAR2001597.pdf>
* <https://prsindia.org/content/banking-sector-india-issues-challenges-and-way-forward>
* <https://rbi.org.in/scripts/NotificationUser.aspx?Mode=0&Id=7736>
* <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>

# Annexure

## Data Description

Please find the details of the data below:

Table 23: Data Description

|  |  |  |
| --- | --- | --- |
| Variable name | Data Description | Additional Description Regarding the Variable |
| **module\_id** | MSME loan given in the form of  1. ADV General Advance  2. CDCC - Current account or Cash Credit |  |
| **cbr\_cd** | Branch Code of the Branch |  |
| **branch\_name** | Name of the Branch |  |
| **region\_name** | Region for which the MSME loan belongs |  |
| **ac\_no** | Account Number |  |
| **ac\_title** | Name of the account |  |
| **open\_dt** | Account open date |  |
| **sme\_sector** | SME sector Classification 1. SME Manufacturing 2. SME Services |  |
| **sector\_name** | Unit Classes under Sector 1. Manufacturing 2. Services | 1. **Manufacturing Enterprises**-the enterprises engaged in the manufacture or production of goods pertaining to any industry specified in the first schedule to the industries (Development and regulation) Act, 1951) or employing plant and machinery in the process of value addition to the final product having a distinct name or character or use. The Manufacturing Enterprise are **defined in terms of investment in Plant & Machinery.** 2. Service Enterprises: -The enterprises engaged in providing or rendering of services and are defined in terms of investment in equipment. |
| **sme\_category** | MSME category 1. Micro 2. Medium 3. Small | 1. Manufacturing Enterprises Investment in plant & machinery Micro Enterprises Does not exceed twenty-five lakh rupees Small Enterprises More than twenty-five lakh rupees but does not exceed five crore rupees Medium Enterprises More than five crore rupees but does not exceed ten crore rupees  2. Service Sector Enterprises Investment in equipment’s Micro Enterprises Does not exceed ten lakh rupees: Small Enterprises More than ten lakh rupees but does not exceed two crore rupees Medium Enterprises More than two crore rupees but does not exceed five crore rupees |
| **scheme\_cd** | Bank's scheme under which loan was given | Details in sheet 'SCHEMES' |
| **bsr\_activity\_cd** | Borrower Activity for which the MSME loan was given | In most cases, debt is classified as nonperforming when loan payments have not been made for a period of 90 days. While 90 days is the standard, the amount of elapsed time may be shorter or longer depending on the terms and conditions of each individual loan. A loan can be classified as a nonperforming asset at any point during the term of the loan or at its maturity. Details in sheet 'bsr\_activity\_code' |
| **industry\_name** | Industry to which the loan account belongs | Eg. Retail, Textile… others |
| **district\_name** | District to which loan belongs to |  |
| **state\_name** | State to which loan belongs to |  |
| **customer\_id** | Customer ID of the account |  |
| **bsr\_org\_cd** | Organisation Code ---> | Details in sheet 'bsr\_org\_cd' |
| **loan\_default** | Is NPA? 0 = No, 1 = Yes |  |
| **Delinquent** | Is Delinquent? 0 = No, 1 = Yes |  |
| **disbursed\_amount** | Amount of Loan disbursed |  |
| **asset\_cost** | Cost of the Asset |  |
| **ltv** | Loan to Value of the asset |  |
| **Date.of.Birth** | Date of birth of the customer |  |
| **Employment.Type** | Employment Type of the customer (Salaried/Self Employed) |  |
| **MobileNo\_Avl\_Flag** | if Mobile no. was shared by the customer then flagged as 1 |  |
| **Aadhar\_flag** | if aadhar was shared by the customer then flagged as 1 |  |
| **PAN\_flag** | if pan was shared by the customer then flagged as 1 |  |
| **VoterID\_flag** | if voter was shared by the customer then flagged as 1 |  |
| **Driving\_flag** | if DL was shared by the customer then flagged as 1 |  |
| **Passport\_flag** | if passport was shared by the customer then flagged as 1 |  |
| **PERFORM\_CNS.SCORE** | Bureau Score |  |
| **PERFORM\_CNS.SCORE.DESCRIPTION** | Bureau score description |  |
| **PRI.NO.OF.ACCTS** | count of total loans taken by the customer at the time of disbursement |  |
| **PRI.ACTIVE.ACCTS** | count of active loans taken by the customer at the time of disbursement |  |
| **PRI.OVERDUE.ACCTS** | count of default accounts at the time of disbursement |  |
| **PRI.CURRENT.BALANCE** | total Principal outstanding amount of the active loans at the time of disbursement |  |
| **PRI.SANCTIONED.AMOUNT** | total amount that was sanctioned for all the loans at the time of disbursement |  |
| **PRI.DISBURSED.AMOUNT** | total amount that was disbursed for all the loans at the time of disbursement |  |
| **SEC.NO.OF.ACCTS** | count of total loans taken by the customer at the time of disbursement |  |
| **SEC.ACTIVE.ACCTS** | count of active loans taken by the customer at the time of disbursement |  |
| **SEC.OVERDUE.ACCTS** | count of default accounts at the time of disbursement |  |
| **SEC.CURRENT.BALANCE** | total Principal outstanding amount of the active loans at the time of disbursement |  |
| **SEC.SANCTIONED.AMOUNT** | total amount that was sanctioned for all the loans at the time of disbursement |  |
| SEC.DISBURSED.AMOUNT | total amount that was disbursed for all the loans at the time of disbursement |  |
| NEW.ACCTS.IN.LAST.SIX.MONTHS | New loans taken by the customer in last 6 months before the disbursement |  |
| DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS | Loans defaulted in the last 6 months |  |

## Detailed EDA

***For more details on variable understanding please refer:***

* ***Bar plots to understand the missing data, maximum and minimum values for variables, frequency of the variable:***

****

* ***Boxplots to understand the spread of the data, outliers, comparison of the defaulters and non-defaulters:***

****

* ***EDA report is the detailed explanation of the variables with their numerical and categorical association.***

****

## Feature Importance for all variables

Table 24: Feature Importance

|  |  |  |
| --- | --- | --- |
| # | *Features* | *Score* |
| 0 | sme\_sector | 0.08 |
| 1 | bsr\_activity\_cd | 0.01 |
| 2 | bsr\_org\_cd | 0.08 |
| 3 | disbursed\_amount | 167.77 |
| 4 | asset\_cost | 0.24 |
| 5 | ltv | 483.19 |
| 6 | PRI.ACTIVE.ACCTS | 108.1 |
| 7 | PRI.OVERDUE.ACCTS | 48.58 |
| 8 | PRI.SANCTIONED.AMOUNT | 48.77 |
| 9 | PRI.DISBURSED.AMOUNT | 47.36 |
| 10 | SEC.NO.OF.ACCTS | 0.45 |
| 11 | SEC.ACTIVE.ACCTS | 0.09 |
| 12 | SEC.OVERDUE.ACCTS | 0.05 |
| 13 | SEC.SANCTIONED.AMOUNT | 0.01 |
| 14 | SEC.DISBURSED.AMOUNT | 0.01 |
| 15 | PRIMARY.INSTAL.AMT | 0.01 |
| 16 | SEC.INSTAL.AMT | 1.03 |
| 17 | NEW.ACCTS.IN.LAST.SIX.MONTHS | **45.15** |
| 18 | DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS | 28.99 |
| 19 | MobileNo\_Avl\_Flag | NaN |
| 20 | Aadhar\_flag | 50.02 |
| 21 | PAN\_flag | 0.00 |
| 22 | VoterID\_flag | 55.94 |
| 23 | Driving\_flag | 0.84 |
| 24 | Passport\_flag | 1.17 |
| 25 | PERFORM\_CNS.SCORE | 240.65 |
| 26 | Age\_of\_Customer | 61.36 |
| 27 | Age\_of\_Loan | 0.05 |
| 28 | scheme\_cd\_cnt | 2.08 |
| 29 | industry\_name\_cnt | 0.92 |
| 30 | branch\_name\_cnt | 0.66 |
| 31 | district\_name\_cnt | 3.92 |
| 32 | region\_name\_cnt | 1.15 |
| 33 | state\_name\_cnt | 4.55 |
| 34 | open\_dt\_day\_cnt | 0.70 |
| 35 | sector\_name\_Services | 0.08 |
| 36 | sme\_category\_MICRO | 2.61 |
| 37 | sme\_category\_SMALL | 2.97 |
| 38 | module\_id\_cdcc | 1.20 |
| 39 | Employment.Type\_Self employed | 9.67 |
| 40 | PERFORM\_CNS.SCORE.DESCRIPTION\_Simplified\_Low Risk | 39.01 |
| 41 | PERFORM\_CNS.SCORE.DESCRIPTION\_Simplified\_Mediu... | 1.96 |
| 42 | PERFORM\_CNS.SCORE.DESCRIPTION\_Simplified\_Not S... | 130.69 |
| 43 | PERFORM\_CNS.SCORE.DESCRIPTION\_Simplified\_Very ... | 60.48 |
| 44 | PERFORM\_CNS.SCORE.DESCRIPTION\_Simplified\_Very ... | 254.52 |

## Detailed VIF values

Table 25: Detailed VIF values



## Code Files

Table : Code Files (Attachments)

|  |  |  |
| --- | --- | --- |
| # | Coding Reference | Files |
| **1** | Data Preprocessing | <https://drive.google.com/drive/folders/1h4u6l__mTcf-A6lUe2cjxiAS-jIL11Yn?usp=sharing> |
| **2** | EDA | <https://drive.google.com/drive/folders/1advR2vMBAgfzxgP25H10-Zmgm8PJL6ok?usp=sharing> |
| **3** | Feature Engineering | <https://drive.google.com/drive/folders/1YueT3rHdJKsBru3arvHLcTY8tfjd3Otn?usp=sharing> |
| **4** | Feature Importance | <https://drive.google.com/drive/folders/1W6NbHZC-PsJwc4GL1McZd3t-74pwttSp?usp=sharing> |
| **5** | Modeling | <https://drive.google.com/drive/folders/1M8K8hiLrQFwkePNFRVyqW7yHgBN8o6cc?usp=sharing> |
| **6** | Deployment | <https://drive.google.com/drive/folders/1VuE6B0QWPmStueORIdhbUAdYED4VB_gU?usp=sharing> |