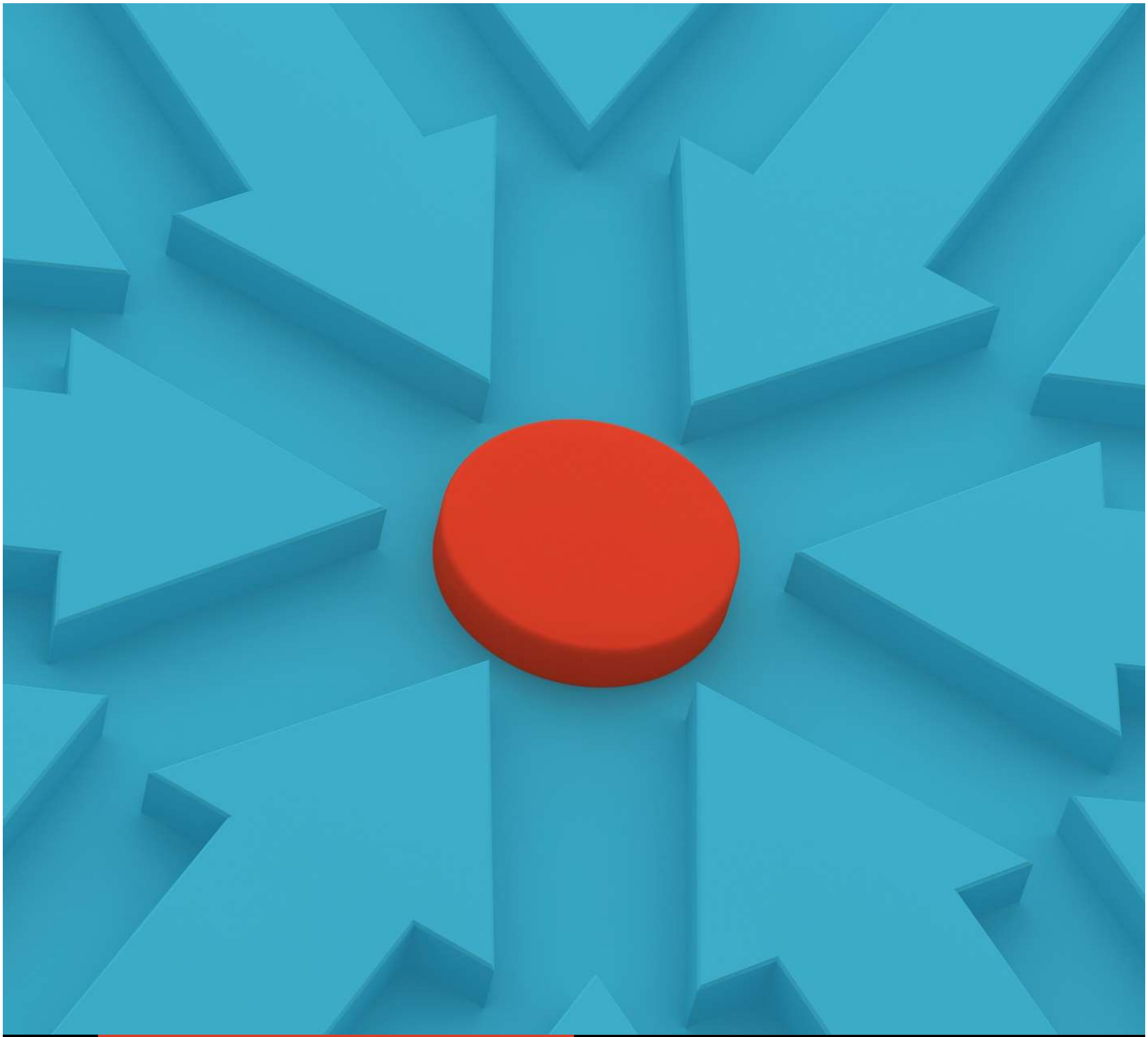
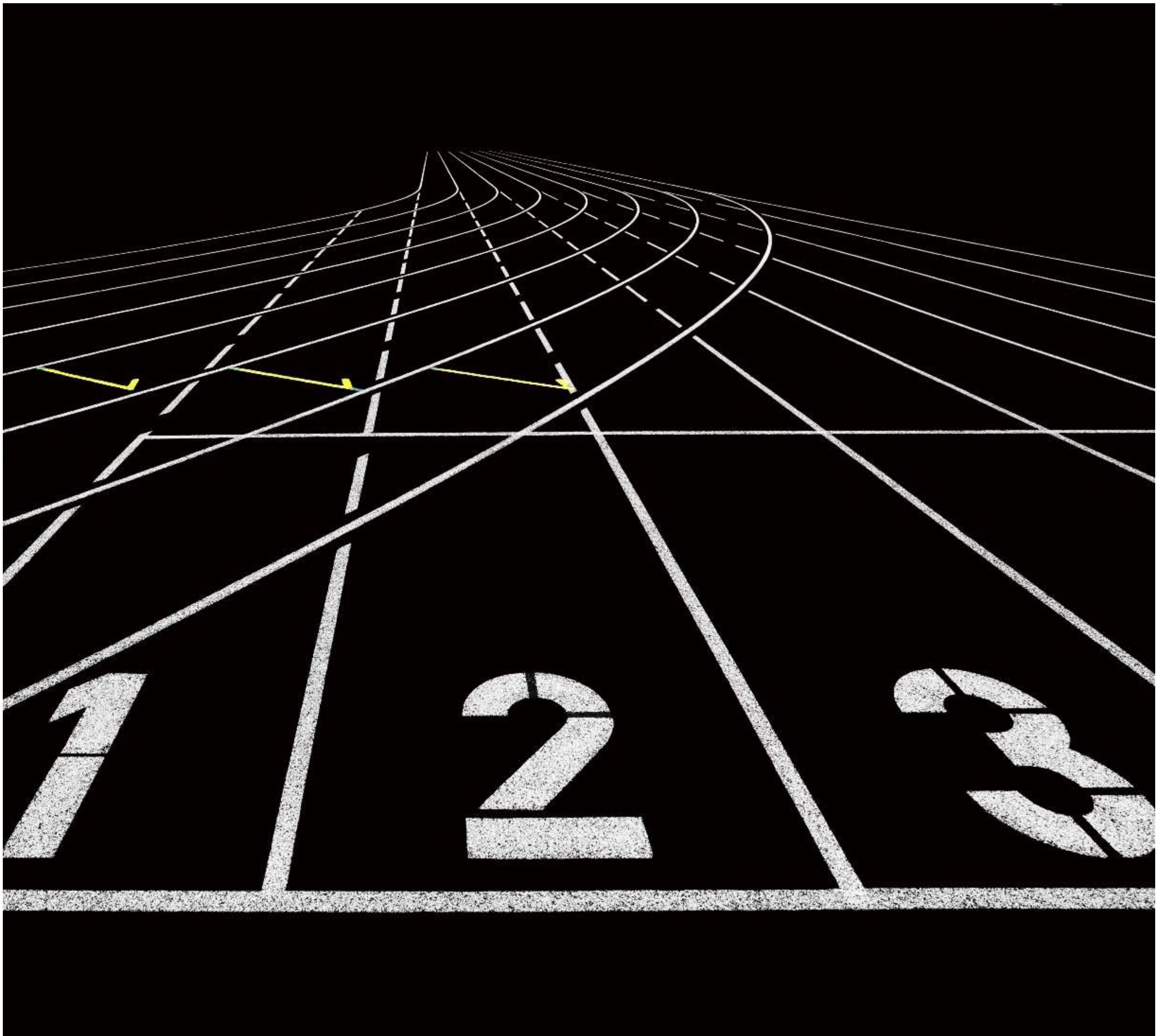


# BUSINESS REPORT

of NBFC Loan foreclosure

Neethu. Sidhardhan




## Introduction

A Non-Banking Financial Company (NBFC) is a company registered under the Companies Act,1956 engaged in the business of loans and advances etc.Foreclosure is a legal process in which a lender attempts to recover the balance of a loan from a borrower who has stopped making payments to the lender by forcing the sale of the asset usedas the collateral for the loan. Foreclosures costs are high and lenders want to find a suitable solution to avoid foreclosures.

# Business Problem

## Defining Problem Statement:

A Non-Banking Financial Company (NBFC) is a company registered under the Companies Act, 1956 engaged in the business of loans and advances, acquisition of shares/stocks/bonds/debentures/securities issued by Government or local authority or other marketable securities of a like nature, leasing, hire-purchase, insurance business, chit business but does not include any institution whose principal business is that of agriculture activity, industrial activity, purchase or sale of any goods (other than securities) or providing any services and sale/purchase/construction of immovable property. A non-banking institution which is a company and has principal business of receiving deposits under any scheme or arrangement in one lump sum or in installments by way of contributions or in any other manner, is also a non-banking financial company (Residuary non-banking company).

## Need of the study/project

If the loan needs to be closed before the tenure, the lender may levy a prepayment penalty. The penalty is to cover the lost interest revenue from early closing of the loans. We need to analyze those customers who will be having the high probability in closing the loans taken from NBFC earlier than the tenor allotted.

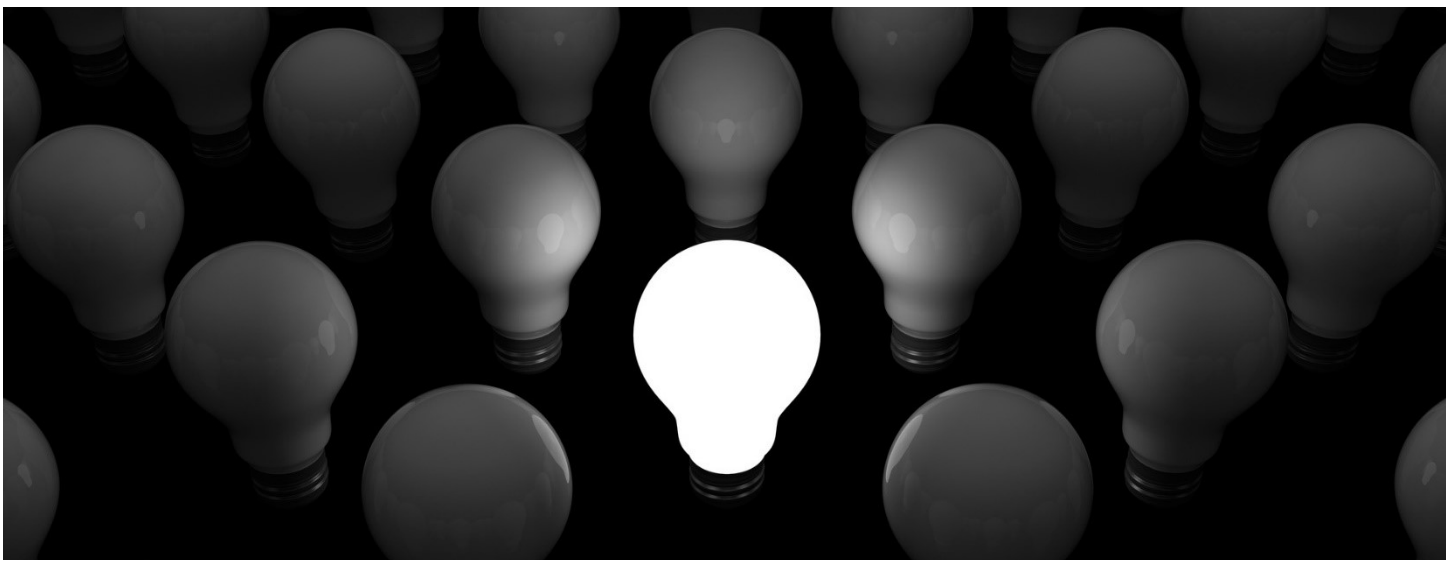


**Figure 1**

## Understanding business/social opportunity

Financially foreclosure weaken the business of NBFC and they are prone to run mismatches in the short-term maturity. Our aim here is to predict those customers that can prone to be early payer of loans. We should present different services and products to the customers that will not encourage them to close the loans earlier than allotted tenor
Foreclosure is a legal process in which a lender attempts to recover the balance of a loan from a borrower who has stopped making payments to the lender by forcing the sale of the asset used as the collateral for the loan.

## 2. EDA and Business Implication

The data shared from NBFC is collected from 2010 to 2019. The data consist of 20012 rows and 53 variables.

| Features | Description | Features | Description |
|---|---|---|---|
| AGREEMENTID | Agreement ID of the loan account (a customer can have multiple loans) | EMI_OS_AMOUNT | EMI outstanding amount |
| AUTHORIZATIONDATE | Authorization date of the loan | EMI_RECEIVED_AMT | EMI received amount |
| BALANCE_EXCESS | Balance of excess amount | EXCESS_ADJUSTED_AMT | Excess adjusted amount |
| BALANCE_TENURE | Remaining tenure | EXCESS_AVAILABLE | Excess received |
| CITY | City of origination | FOIR | Fixed obligation to income ratio (Value should range from 0-1 – Derived variable) |
| COMPLETED_TENURE | Completed tenure | INTEREST_START_DATE | Interest start date on the loan |
| CURRENT_INTEREST_RATE | Current rate of interest on the loan. Renamed field (Old Name: CURRENT_ROI ) | LAST_RECEIPT_AMOUNT | Last receipt amount |
| CURRENT_INTEREST_RATE_MAX | Maximum value of the CURRENT ROI across transactions | LAST_RECEIPT_DATE | Last receipt date |
| CURRENT_INTEREST_RATE_MIN | Minimum value of the CURRENT ROI across transactions | LATEST_TRANSACTION_MONTH | Month of last receipt date. In case account is Foreclosed, it will be month of Foreclosure |
| CURRENT_INTEREST_RATE_CHANGES | Number of times the CURRENT ROI has changed | LOAN_AMT | Loan amount which was sanctioned |
| CURRENT_TENOR | Current tenor of the loan | MAX_EMI_AMOUNT | Maximum receipt amount |
| CUSTOMERID | Unique Customer ID given to each customer | MIN_EMI_AMOUNT | Minimum receipt amount |
| DIFF_AUTH_INT_DATE | Difference between authorization and interest start date | MONTHOPENING | Month of opening |
| DIFF_CURRENT_INTEREST_RATE_MAX_MIN | Difference between the maximum and minimum interest rate per agreement | NET_DISBURSED_AMT | Amount that was disbursed |
| DIFF_EMI_AMOUNT_MAX_MIN | Difference in original and current tenor (ORIGNAL_TENOR | NET_LTV | Net Loan to Value ratio (Value ranges from 0-100 (in %) – Derived variable) |

| | | | |
|---|---|---|---|
| | - CURRENT_TENOR) | | |
| DIFF_ORIGINAL_CURRENT_INTEREST_RATE | Difference in original ROI and current ROI (ORIGNAL_ROI - CURRENT_ROI) | NET_RECEIVABLE | Net receivable (EMI_DUEAMT - EMI_RECEIVED_AMT = EMI_OS_AMOUNT) + (EXCESS_AVAILABLE - EXCESS_ADJUSTED_AMT = BALANCE_EXCESS) = NET_RECEIVABLE) |
| DIFF_ORIGINAL_CURRENT_TENOR | Difference in original and current tenor (ORIGNAL_TENOR - CURRENT_TENOR) | NUM_EMI_CHANGES | Number of different values in the receipts amount |
| DPD | Days past due | NUM_LOW_FREQ_TRANSACTIONS | Number of transactions done in less than 28 days |
| DUEDAY | Next due date of the loan | ORIGNAL_INTEREST_RATE | Original rate of interest on the loan (when the loan was sanctioned). Renamed field (Old Name: ORIGNAL_ROI) |
| EMI_AMOUNT | Mode of the receipt amount | ORIGNAL_TENOR | Original tenor of the loan (when the loan was sanctioned) |
| EMI_DUEAMT | EMI due amount | OUTSTANDING_PRINCIPAL | Outstanding principal |
| PAID_INTEREST | Paid interest | PAID_PRINCIPAL | PAID_PRINCIPAL |
| PRE_EMI_DUEAMT | Pre EMI due amount for the loan | PRE_EMI_OS_AMOUNT | Pre EMI-Outstanding amount |
| PRE_EMI_RECEIVED_AMT | Pre EMI that was received | PRODUCT | Loan product |
| SCHEMEID | Scheme ID under which loan was given | NPA_IN_LAST_MONTH | Whether NPA in last month |
| NPA_IN_CURRENT_MONTH | Whether NPA in current month | MOB | Internal code |
| FORECLOSURE | Labelled Field | | |

## **Type of Data**:

The data consist of Date formats, integers, floats and objects.

The data consist of both categorical variable and numeric variables.

Categorical Variables in the data seta are cities, NPA last month and NPA current month and Products

# Descriptive Analysis:

| | Mean | Standard Error | Median | Mode | Standard Deviation | Sample Variance | Kurtosis | Skewness | Range | Min | Max | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Balance Excess | 78995.98 | 9533.44 | 0.00 | 0.00 | 1348636.32 | 1818819934619.75 | 1499.50 | 34.31 | 75555999.48 | 0.00 | 75555999.48 | 1580867545.24 |
| BALANCE TENURE | 172.82 | 0.45 | 174.00 | 180.00 | 64.00 | 4096.57 | 2.31 | 0.31 | 674.00 | 0.00 | 674.00 | 3458566.00 |
| Completed Tenure | 17.27 | 0.12 | 12.00 | 0.00 | 16.49 | 271.80 | 2.37 | 1.53 | 98.00 | 0.00 | 98.00 | 345589.00 |
| Current interest | 14.78 | 0.02 | 14.55 | 17.48 | 2.49 | 6.18 | -0.54 | 0.29 | 15.19 | 9.90 | 25.10 | 295816.00 |
| Current Int Max | 14.90 | 0.02 | 14.67 | 17.48 | 2.48 | 6.15 | 0.09 | 0.29 | 27.03 | 10.43 | 37.46 | 298183.76 |
| Current int Min | 14.30 | 0.02 | 13.73 | 17.48 | 2.68 | 7.17 | -0.37 | 0.39 | 29.09 | -5.06 | 24.03 | 286209.08 |
| Current Int Change | 0.76 | 0.01 | 0.00 | 0.00 | 1.13 | 1.29 | 2.62 | 1.56 | 9.00 | 0.00 | 9.00 | 15171.00 |
| Current Tenure | 190.09 | 0.41 | 180.00 | 180.00 | 58.56 | 3429.27 | 3.21 | 0.49 | 707.00 | 6.00 | 713.00 | 3804155.00 |
| Diff Auth date | 0.01 | 0 | 0 | 0 | 0.57 | 0.32 | 11499.96 | 94.53 | 87.00 | -17.00 | 70.00 | 126.00 |
| Diff in int max and min | 0.60 | 0.01 | 0 | 0 | 0.97 | 0.93 | 61.65 | 4.18 | 24.35 | 0 | 24.35 | 11974.67 |
| Diff EMI min and max | 115209.42 | 6851.51 | 19885.00 | 0 | 967082.44 | 935248446862.33 | 3317.06 | 46.40 | 84968249.90 | 0 | 84968249.90 | 2295317193.70 |
| Diff Original and current int | -0.38 | 0.01 | 0 | 0 | 0.88 | 0.78 | 5.56 | 0.28 | 17.50 | -7.18 | 10.32 | -7614.64 |
| Diff original and current tenor | -6.80 | 0.24 | 0 | 0 | 33.53 | 1123.98 | 16.30 | -0.15 | 695.00 | -461.00 | 234.00 | -136009 |
| DPD | 7.57 | 0.47 | 0 | 0 | 66.10 | 4369.06 | 293.75 | 15.50 | 2054 | 0 | 2054 | 151572 |
| Due day | 5.78 | 0.02 | 5.00 | 5.00 | 2.72 | 7.39 | 7.55 | 3.05 | 14 | 1 | 15 | 115602 |
| EMI Amount | 43609.50 | 799.72 | 18937.50 | 118.00 | 113131.82 | 12798809358.61 | 317.76 | 13.34 | 4879479.00 | 0.00 | 4879479.00 | 872713217.44 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EMI O/S | 33297.35 | 4638.16 | 0.00 | 0.00 | 656131.13 | 430508065929.47 | 4118.48 | 55.96 | 58995308.80 | 0.00 | 58995308.80 | 666346538.06 |
| EMI Received | 1958255.83 | 47807.18 | 537657.63 | 0.00 | 6762984.20 | 45737955299673.90 | 505.19 | 16.18 | 354610409.64 | 0.00 | 354610409.64 | 39188615638.77 |
| Excess Adj | 359900.21 | 27733.92 | 0.00 | 0.00 | 3923345.60 | 15392640672853.30 | 2442.28 | 41.57 | 284164206.62 | 0.00 | 284164206.62 | 7202323067.92 |
| Excess Avail | 438896.19 | 29475.81 | 260.61 | 0.00 | 4169759.35 | 17386893062735.80 | 1946.38 | 36.12 | 284164207.07 | 0.00 | 284164207.07 | 8783190613.12 |
| FOIR | 27.96 | 27.36 | 0.52 | 0.49 | 3871.06 | 14985142.35 | 20011.99 | 141.46 | 547786.33 | -170.33 | 547616.00 | 559536.21 |
| Last Receipt Amt | 80674.46 | 5750.15 | 19642.00 | 118.00 | 808402.70 | 653514917507.88 | 6427.43 | 67.90 | 84968810.90 | 1.00 | 84968811.90 | 1594530661.70 |
| Loan Amt | 5897355.27 | 91794.95 | 2684572.11 | 1501295.80 | 12985660.51 | 168627378954917 | 199.87 | 10.86 | 424528919.85 | 37532.40 | 424566452.24 | 118017873678.61 |
| Max EMI | 122254.44 | 6875.38 | 23600.00 | 267280.00 | 970451.59 | 941776289557.743 | 3271.08 | 45.97 | 84968798.56 | 13.34 | 84968811.90 | 2435675237.02 |
| Min EMI | 7045.03 | 307.66 | 133.18 | 118.00 | 43425.49 | 1885773032.28 | 1835.75 | 33.19 | 3156964.99 | 0.01 | 3156965.00 | 140358043.32 |
| Month opening | 5447511.16 | 83685.83 | 2503693.74 | 3753239.50 | 11838513.01 | 140150390379834 | 214.67 | 11.20 | 381836715.30 | 0.00 | 381836715.30 | 109015593296.32 |
| Net Disbursed | 5847665.54 | 91273.77 | 2640779.31 | 1501295.80 | 12911931.76 | 166717981822539 | 203.77 | 10.97 | 424528919.85 | 37532.40 | 424566452.24 | 117023482743.20 |
| Net LTV | 51.19 | 0.15 | 53.30 | 80.00 | 21.11 | 445.50 | -0.79 | -0.21 | 99.62 | 0.38 | 100.00 | 1024399.08 |
| Original Int | 14.40 | 0.02 | 13.73 | 13.42 | 2.60 | 6.78 | -0.57 | 0.41 | 18.13 | 9.65 | 27.78 | 288201.36 |
| Original Tenor | 183.30 | 0.32 | 180.00 | 180.00 | 44.60 | 1989.18 | 0.16 | -0.21 | 286.00 | 14.00 | 300.00 | 3668146 |
| Original Principal | 5212982.40 | 81443.84 | 2394655.38 | 3753239.50 | 11521352.56 | 132741564914715.00 | 233.70 | 11.67 | 381836716 | -0.75 | 381836715.3 | 104322203839 |
| Paid Int | 989054.69 | 21391.01 | 309724.83 | 0.00 | 3026052.53 | 9156993905212.06 | 290.36 | 13.21 | 9156993905212.06 | 0.00 | 381836715.304797 | 104322203839.547 |
| Paid principal | 866763.73 | 245275.38 | 78786.50 | 0.00 | 34697580.79 | 1203922112837700.00 | 19625.77 | 139.43 | 1203922112837700 | 0.00 | 123036220.646358 | 19792962427.2723 |
| Pre EMI | 57804.47 | 2669.69 | 10696.02 | 0.00 | 377664.74 | 142630656963.55 | 2929.58 | 43.42 | 31775396.135568 | 0.00 | 31775396.135568 | 1156783045.07047 |

In descriptive statistics we calculate the mean, median, mode, standard deviation, range, minimum, and maximum of the data.

## Multivariate Analysis:

Multivariate analysis is based on the principles of multivariate statistics, which involves observation and analysis of more than one variable. The relation of each variables with one can be done by using a correlation matrix and regression model by using stats model.

## Correlation Matrix:

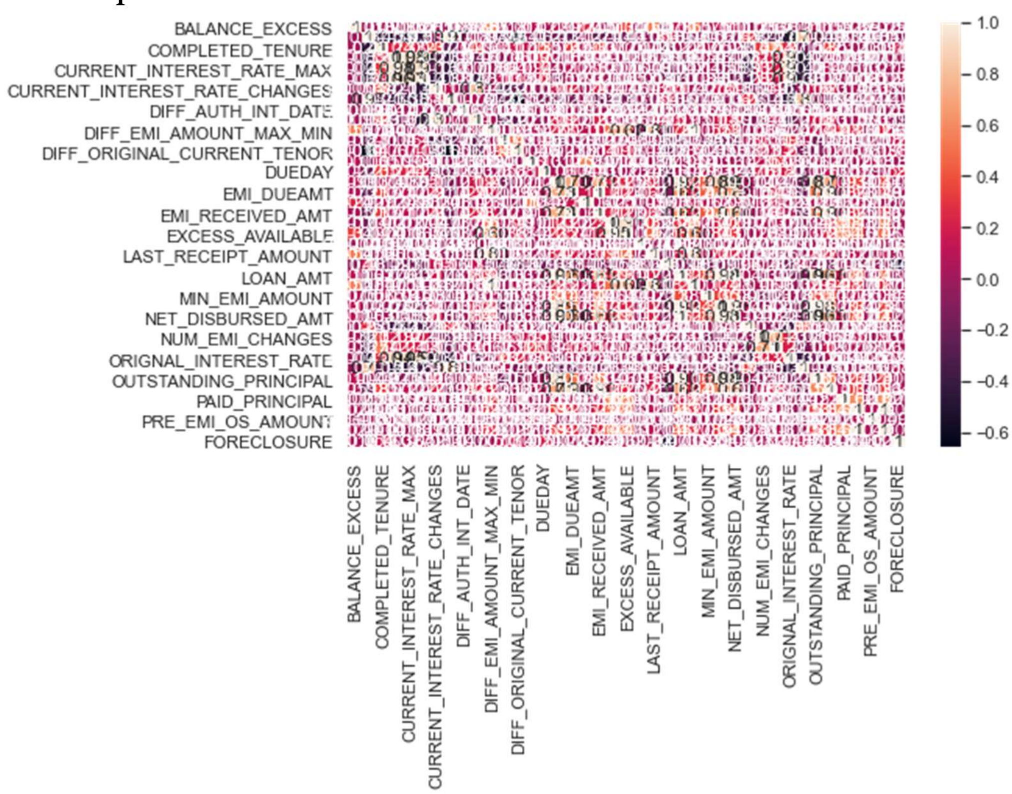
| | BALANCE_EXCE | ALANCE_TENU | MPLETED_TEN | NT_INTEREST | INTEREST_RA | INTEREST_RA | NTEREST_RATE | JRRENT_TENC | AUTH_INT_D | INTEREST_RA | AMOUNT_M | CURRENT_IN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BALANCE_EXCESS | 1.00 | | | | | | | | | | | |
| BALANCE_TENURE | -0.01 | 1.00 | | | | | | | | | | |
| COMPLETED_TENURE | 0.02 | -0.44 | 1.00 | | | | | | | | | |
| JRRENT_INTEREST_RA | -0.01 | -0.39 | 0.24 | 1.00 | | | | | | | | |
| RENT_INTEREST_RATE_ | 0.00 | -0.43 | 0.33 | 0.98 | 1.00 | | | | | | | |
| RENT_INTEREST_RATE_ | -0.01 | -0.39 | 0.20 | 0.96 | 0.93 | 1.00 | | | | | | |
| NT_INTEREST_RATE_CH | 0.01 | 0.01 | 0.37 | -0.21 | -0.13 | -0.41 | 1.00 | | | | | |
| CURRENT_TENOR | -0.01 | 0.97 | -0.20 | -0.35 | -0.38 | -0.37 | 0.11 | 1.00 | | | | |
| DIFF_AUTH_INT_DATE | 0.00 | -0.02 | 0.04 | 0.01 | 0.02 | 0.01 | 0.01 | -0.01 | 1.00 | | | |
| RENT_INTEREST_RATE_ | 0.01 | -0.01 | 0.29 | -0.15 | -0.02 | -0.38 | 0.81 | 0.07 | 0.00 | 1.00 | | |
| _EMI_AMOUNT_MAX_ | 0.55 | -0.08 | 0.04 | -0.01 | 0.01 | -0.01 | 0.04 | -0.07 | 0.00 | 0.05 | 1.00 | |
| GINAL_CURRENT_INTE | 0.02 | -0.23 | 0.11 | -0.04 | 0.06 | 0.13 | -0.30 | -0.22 | -0.06 | -0.21 | 0.01 | 1.00 |
| ORIGINAL_CURRENT_T | 0.00 | -0.64 | 0.16 | 0.07 | 0.10 | 0.18 | -0.21 | -0.65 | 0.01 | -0.22 | 0.09 | 0.49 |
| DPD | 0.10 | -0.06 | 0.15 | 0.09 | 0.09 | 0.09 | 0.00 | -0.02 | 0.00 | -0.01 | 0.05 | 0.00 |
| DUEDAY | 0.02 | -0.20 | 0.35 | 0.22 | 0.24 | 0.23 | -0.05 | -0.12 | 0.04 | -0.03 | 0.03 | 0.06 |
| EMI_AMOUNT | 0.10 | -0.13 | 0.08 | 0.00 | 0.02 | -0.05 | 0.17 | -0.12 | 0.01 | 0.19 | 0.23 | -0.11 |
| EMI_DUEAMT | 0.07 | -0.21 | 0.33 | 0.07 | 0.12 | 0.03 | 0.22 | -0.14 | 0.03 | 0.23 | 0.20 | -0.03 |
| EMI_OS_AMOUNT | 0.24 | -0.03 | 0.09 | 0.03 | 0.04 | 0.03 | 0.02 | 0.00 | 0.00 | 0.01 | 0.11 | -0.03 |
| EMI_RECEIVED_AMT | 0.05 | -0.21 | 0.33 | 0.07 | 0.12 | 0.03 | 0.22 | -0.14 | 0.03 | 0.23 | 0.20 | -0.03 |
| XCESS_ADJUSTED_AM | 0.02 | -0.14 | 0.07 | 0.00 | 0.02 | 0.00 | 0.05 | -0.13 | 0.00 | 0.06 | 0.51 | 0.03 |
| EXCESS_AVAILABLE | 0.34 | -0.14 | 0.07 | 0.00 | 0.02 | -0.01 | 0.05 | -0.13 | 0.00 | 0.06 | 0.66 | 0.04 |
| FOIR | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.02 | 0.00 | 0.00 | 0.00 | 0.00 |
| AST_RECEIPT_AMOUN | 0.54 | -0.06 | 0.03 | -0.01 | 0.01 | -0.01 | 0.03 | -0.06 | 0.00 | 0.05 | 0.84 | 0.01 |
| LOAN_AMT | 0.11 | -0.06 | 0.02 | -0.06 | -0.05 | -0.11 | 0.17 | -0.06 | 0.01 | 0.19 | 0.26 | -0.12 |
| MAX_EMI_AMOUNT | 0.55 | -0.08 | 0.05 | -0.01 | 0.01 | -0.01 | 0.04 | -0.08 | 0.00 | 0.05 | 1.00 | 0.00 |
| MIN_EMI_AMOUNT | 0.07 | -0.06 | 0.07 | 0.04 | 0.05 | 0.04 | 0.01 | -0.04 | 0.03 | 0.01 | 0.06 | -0.06 |
| MONTHOPENING | 0.11 | -0.03 | -0.01 | -0.08 | -0.06 | -0.12 | 0.16 | -0.03 | 0.01 | 0.18 | 0.23 | -0.14 |
| NET_DISBURSED_AMT | 0.11 | -0.06 | 0.03 | -0.06 | -0.04 | -0.11 | 0.17 | -0.06 | 0.01 | 0.19 | 0.27 | -0.13 |
| NET_LTV | 0.00 | 0.28 | -0.14 | -0.32 | -0.33 | -0.28 | -0.06 | 0.26 | 0.00 | -0.08 | 0.01 | 0.03 |
| NET_RECEIVABLE | -0.88 | 0.00 | 0.02 | 0.03 | 0.02 | 0.02 | 0.01 | 0.01 | 0.00 | 0.00 | -0.50 | -0.03 |
| RIGNAL_INTEREST_RA | 0.00 | -0.45 | 0.27 | 0.94 | 0.95 | 0.96 | -0.30 | -0.41 | -0.01 | -0.22 | 0.00 | 0.30 |
| ORIGNAL_TENOR | -0.01 | 0.79 | -0.15 | -0.41 | -0.42 | -0.36 | -0.02 | 0.82 | -0.01 | -0.07 | -0.03 | 0.07 |
| UTSTANDING_PRINCIP | 0.11 | 0.00 | -0.03 | -0.08 | -0.07 | -0.12 | 0.16 | 0.00 | 0.00 | 0.17 | 0.14 | -0.15 |
| PAID_INTEREST | 0.06 | -0.18 | 0.35 | 0.07 | 0.12 | 0.03 | 0.23 | -0.09 | 0.03 | 0.23 | 0.22 | -0.04 |
| PAID_PRINCIPAL | 0.00 | -0.04 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | -0.04 | 0.00 | 0.02 | 0.28 | 0.01 |
| PRE_EMI_DUEAMT | 0.03 | -0.05 | 0.01 | 0.00 | 0.01 | -0.01 | 0.05 | -0.05 | 0.00 | 0.06 | 0.26 | 0.00 |
| PRE_EMI_OS_AMOUN | 0.00 | 0.01 | -0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| RE_EMI_RECEIVED_AM | 0.03 | -0.05 | 0.01 | 0.00 | 0.01 | -0.01 | 0.05 | -0.05 | 0.00 | 0.06 | 0.26 | 0.00 |
| MOB | 0.02 | -0.45 | 0.99 | 0.25 | 0.33 | 0.20 | 0.36 | -0.21 | 0.04 | 0.28 | 0.05 | 0.11 |

| | INAL_CURREN | DPD | DUEDAY | EMI_AMOUNT | EMI_DUEAMT | I_OS_AMOUN | RECEIVED_A | S_ADJUSTED | CESS_AVAILAB | FOIR | RECEIPT_AMC | LOAN_AMT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BALANCE_EXCESS | | | | | | | | | | | | |
| BALANCE_TENURE | | | | | | | | | | | | |
| COMPLETED_TENURE | | | | | | | | | | | | |
| JRRENT_INTEREST_RA | | | | | | | | | | | | |
| RENT_INTEREST_RATE_ | | | | | | | | | | | | |
| RENT_INTEREST_RATE_ | | | | | | | | | | | | |
| NT_INTEREST_RATE_CH | | | | | | | | | | | | |
| CURRENT_TENOR | | | | | | | | | | | | |
| DIFF_AUTH_INT_DATE | | | | | | | | | | | | |
| RENT_INTEREST_RATE_ | | | | | | | | | | | | |
| _EMI_AMOUNT_MAX_ | | | | | | | | | | | | |
| GINAL_CURRENT_INTE | | | | | | | | | | | | |
| ORIGINAL_CURRENT_T | 1.00 | | | | | | | | | | | |
| DPD | 0.00 | 1.00 | | | | | | | | | | |
| DUEDAY | 0.08 | 0.13 | 1.00 | | | | | | | | | |
| EMI_AMOUNT | -0.02 | 0.01 | 0.06 | 1.00 | | | | | | | | |
| EMI_DUEAMT | 0.04 | 0.07 | 0.13 | 0.73 | 1.00 | | | | | | | |
| EMI_OS_AMOUNT | -0.01 | 0.58 | 0.04 | 0.07 | 0.16 | 1.00 | | | | | | |
| EMI_RECEIVED_AMT | 0.04 | 0.02 | 0.13 | 0.73 | 1.00 | 0.07 | 1.00 | | | | | |
| XCESS_ADJUSTED_AM | 0.19 | 0.01 | 0.02 | 0.23 | 0.22 | 0.03 | 0.22 | 1.00 | | | | |
| EXCESS_AVAILABLE | 0.18 | 0.04 | 0.02 | 0.25 | 0.24 | 0.10 | 0.23 | 0.95 | 1.00 | | | |
| FOIR | 0.00 | 0.00 | 0.02 | 0.03 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 1.00 | | |
| AST_RECEIPT_AMOUN | 0.05 | 0.04 | 0.03 | 0.22 | 0.20 | 0.10 | 0.19 | 0.36 | 0.53 | 0.00 | 1.00 | |
| LOAN_AMT | -0.04 | 0.01 | 0.04 | 0.91 | 0.65 | 0.09 | 0.65 | 0.29 | 0.31 | 0.02 | 0.22 | 1.00 |
| MAX_EMI_AMOUNT | 0.09 | 0.05 | 0.03 | 0.25 | 0.22 | 0.11 | 0.21 | 0.52 | 0.66 | 0.00 | 0.84 | 0.28 |
| MIN_EMI_AMOUNT | 0.00 | 0.01 | 0.08 | 0.41 | 0.44 | 0.06 | 0.44 | 0.06 | 0.08 | 0.00 | 0.07 | 0.38 |
| MONTHOPENING | -0.06 | 0.02 | 0.03 | 0.89 | 0.57 | 0.09 | 0.56 | 0.21 | 0.23 | 0.02 | 0.22 | 0.98 |
| NET_DISBURSED_AMT | -0.04 | 0.01 | 0.05 | 0.92 | 0.65 | 0.09 | 0.65 | 0.29 | 0.31 | 0.02 | 0.22 | 1.00 |
| NET_LTV | 0.00 | 0.02 | -0.08 | 0.01 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | -0.01 | 0.01 | 0.04 |
| NET_RECEIVABLE | -0.01 | 0.18 | 0.00 | -0.06 | 0.01 | 0.24 | -0.02 | 0.00 | -0.29 | 0.00 | -0.50 | -0.07 |
| RIGNAL_INTEREST_RA | 0.23 | 0.09 | 0.23 | -0.04 | 0.06 | 0.02 | 0.06 | 0.01 | 0.01 | 0.00 | 0.00 | -0.10 |
| ORIGNAL_TENOR | -0.11 | -0.03 | -0.10 | -0.18 | -0.16 | -0.01 | -0.16 | -0.03 | -0.03 | -0.02 | -0.03 | -0.11 |
| UTSTANDING_PRINCIP | -0.10 | 0.01 | 0.03 | 0.86 | 0.53 | 0.08 | 0.53 | 0.05 | 0.08 | 0.02 | 0.15 | 0.96 |
| PAID_INTEREST | 0.02 | 0.03 | 0.13 | 0.71 | 0.96 | 0.09 | 0.96 | 0.25 | 0.25 | 0.00 | 0.21 | 0.68 |
| PAID_PRINCIPAL | 0.05 | 0.00 | 0.01 | 0.09 | 0.07 | 0.00 | 0.07 | 0.58 | 0.54 | 0.00 | 0.03 | 0.18 |
| PRE_EMI_DUEAMT | 0.04 | 0.02 | 0.06 | 0.34 | 0.26 | 0.06 | 0.25 | 0.48 | 0.46 | 0.01 | 0.12 | 0.42 |
| PRE_EMI_OS_AMOUN | 0.00 | 0.10 | 0.02 | -0.01 | -0.01 | 0.01 | -0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 |
| RE_EMI_RECEIVED_AM | 0.04 | 0.02 | 0.06 | 0.34 | 0.26 | 0.06 | 0.26 | 0.48 | 0.46 | 0.01 | 0.12 | 0.42 |
| MOB | 0.17 | 0.16 | 0.35 | 0.07 | 0.33 | 0.09 | 0.32 | 0.08 | 0.08 | 0.00 | 0.03 | 0.02 |

| | MAX_EMI_AMOU | N_EMI_AMOU | ONTHOPENIN | DISBURSED_ | NET_LTV | ET_RECEIVABL | IAL_INTEREST | RIGNAL_TENO | ANDING_PRIN | AID_INTERES | AID_PRINCIPA | E_EMI_DUEAN | EMI_OS_AMC | MI_RECEIVED | MOB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MAX_EMI_AMOUNT | 1.00 | | | | | | | | | | | | | | |
| MIN_EMI_AMOUNT | 0.10 | 1.00 | | | | | | | | | | | | | |
| MONTHOPENING | 0.24 | 0.35 | 1.00 | | | | | | | | | | | | |
| NET_DISBURSED_AMT | 0.28 | 0.38 | 0.98 | 1.00 | | | | | | | | | | | |
| NET_LTV | 0.01 | 0.01 | 0.04 | 0.04 | 1.00 | | | | | | | | | | |
| NET_RECEIVABLE | -0.50 | -0.04 | -0.07 | -0.07 | 0.01 | 1.00 | | | | | | | | | |
| RIGNAL_INTEREST_RA | 0.00 | 0.02 | -0.12 | -0.10 | -0.30 | 0.01 | 1.00 | | | | | | | | |
| ORIGNAL_TENOR | -0.03 | -0.05 | -0.09 | -0.11 | 0.35 | 0.00 | -0.37 | 1.00 | | | | | | | |
| UTSTANDING_PRINCIP | 0.16 | 0.35 | 0.98 | 0.96 | 0.04 | -0.07 | -0.12 | -0.08 | 1.00 | | | | | | |
| PAID_INTEREST | 0.24 | 0.41 | 0.61 | 0.68 | 0.01 | -0.02 | 0.06 | -0.11 | 0.58 | 1.00 | | | | | |
| PAID_PRINCIPAL | 0.28 | 0.03 | 0.04 | 0.18 | 0.00 | 0.00 | 0.00 | -0.01 | 0.02 | 0.11 | 1.00 | | | | |
| PRE_EMI_DUEAMT | 0.27 | 0.08 | 0.33 | 0.42 | 0.01 | 0.00 | 0.00 | -0.03 | 0.31 | 0.34 | 0.62 | 1.00 | | | |
| PRE_EMI_OS_AMOUN | 0.00 | 0.00 | 0.03 | 0.02 | 0.01 | 0.01 | 0.00 | 0.01 | 0.03 | 0.00 | 0.00 | 0.08 | 1.00 | | |
| RE_EMI_RECEIVED_AM | 0.27 | 0.08 | 0.33 | 0.42 | 0.01 | 0.00 | 0.00 | -0.03 | 0.31 | 0.34 | 0.62 | 1.00 | 0.05 | 1.00 | |
| MOB | 0.05 | 0.07 | -0.01 | 0.03 | -0.14 | 0.03 | 0.27 | -0.14 | -0.03 | 0.35 | 0.02 | 0.06 | 0.00 | 0.06 | 1.00 |

We have taken 39 numeric features in the data set to check the correlation of each variables. The darker the color shows they are highly correlated and lighter the color shows they are negative correlated. The correlation matrix shows that there are certain variables like "Current interest", interest rate max and interest rate min, interest rate change, original interest rate is highly correlated among each other.

## Heat Map:



**Logistic Regression Stats Model**:

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | FORECLOSURE | R-squared: | 0.263 |
| Model: | OLS | Adj. R-squared: | 0.262 |
| Method: | Least Squares | F-statistic: | 214.9 |
| Date: | Thu, 04 Feb 2021 | Prob (F-statistic): | 0.00 |
| Time: | 18:18:20 | Log-Likelihood: | -315.83 |
| No. Observations: | 19922 | AIC: | 699.7 |
| Df Residuals: | 19888 | BIC: | 968.2 |
| Df Model: | 33 | | |
| Covariance Type: | nonrobust | | |

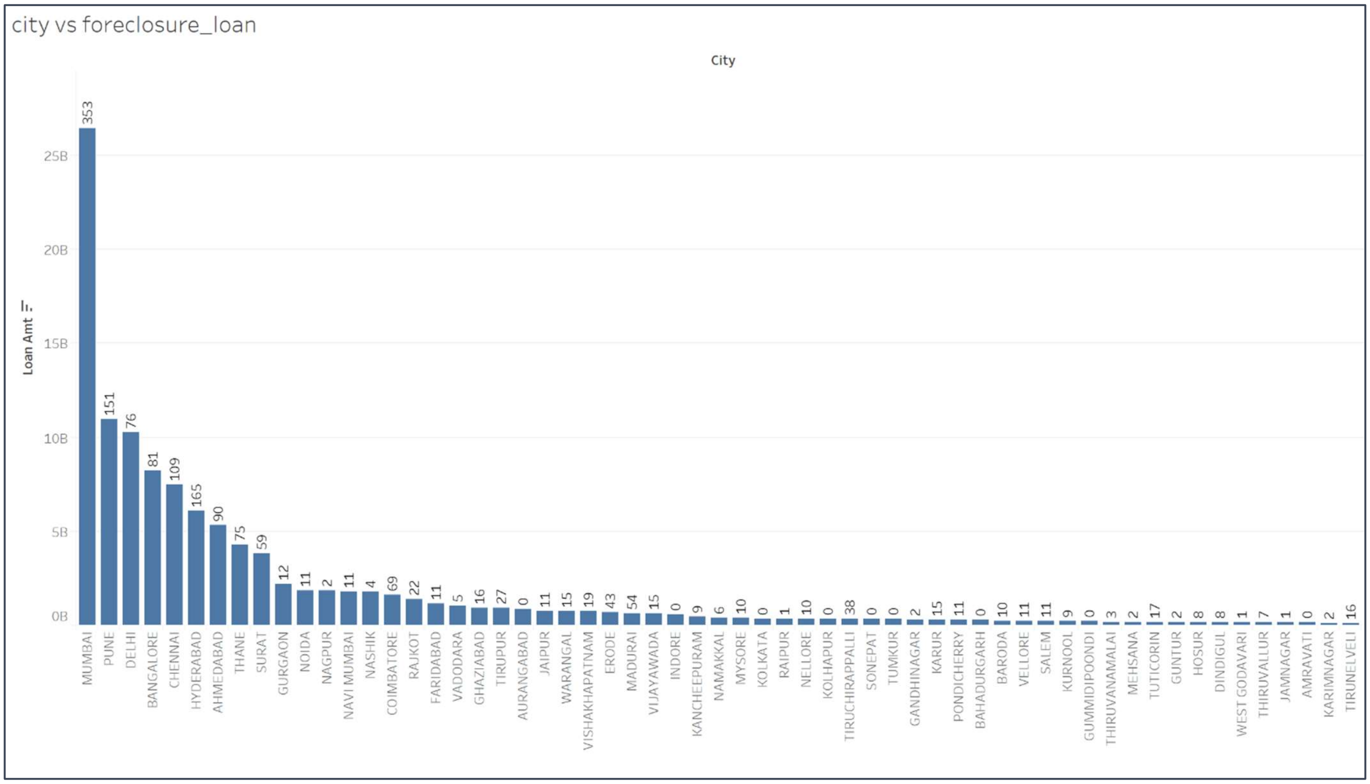
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 0.3944 | 0.020 | 19.784 | 0.000 | 0.355 | 0.433 |
| NET_DISBURSED_AMT | 6.833e-09 | 7.17e-09 | 0.952 | 0.341 | -7.23e-09 | 2.09e-08 |
| BALANCE_EXCESS | -5.9819 | 5.624 | -1.064 | 0.288 | -17.006 | 5.043 |
| BALANCE_TENURE | -0.0008 | 6.35e-05 | -11.911 | 0.000 | -0.001 | -0.001 |
| COMPLETED_TENURE | 0.0014 | 0.000 | 13.784 | 0.000 | 0.001 | 0.002 |
| CURRENT_INTEREST_RATE | -0.0377 | 0.004 | -9.634 | 0.000 | -0.045 | -0.030 |
| CURRENT_INTEREST_RATE_MAX | 0.0262 | 0.004 | 7.157 | 0.000 | 0.019 | 0.033 |
| CURRENT_INTEREST_RATE_MIN | 0.0393 | 0.004 | 10.831 | 0.000 | 0.032 | 0.046 |
| CURRENT_INTEREST_RATE_CHANGES | 0.0269 | 0.003 | 8.707 | 0.000 | 0.021 | 0.033 |
| CURRENT_TENOR | 0.0006 | 5.08e-05 | 12.324 | 0.000 | 0.001 | 0.001 |
| DIFF_AUTH_INT_DATE | -0.0030 | 0.003 | -0.971 | 0.332 | -0.009 | 0.003 |
| DIFF_CURRENT_INTEREST_RATE_MAX_MIN | -0.0131 | 0.002 | -6.129 | 0.000 | -0.017 | -0.009 |
| DIFF_EMI_AMOUNT_MAX_MIN | -2.079e-06 | 2.49e-06 | -0.837 | 0.403 | -6.95e-06 | 2.79e-06 |
| DIFF_ORIGINAL_CURRENT_INTEREST_RATE | 0.0096 | 0.002 | 5.386 | 0.000 | 0.006 | 0.013 |
| DIFF_ORIGINAL_CURRENT_TENOR | 0.0003 | 5.18e-05 | 6.206 | 0.000 | 0.000 | 0.000 |
| DPD | -8.42e-05 | 3.42e-05 | -2.463 | 0.014 | -0.000 | -1.72e-05 |
| EMI_AMOUNT | 2.055e-07 | 5.23e-08 | 3.926 | 0.000 | 1.03e-07 | 3.08e-07 |
| EMI_DUEAMT | -9.8572 | 11.192 | -0.881 | 0.378 | -31.794 | 12.080 |
| EMI_OS_AMOUNT | 9.8572 | 11.192 | 0.881 | 0.378 | -12.080 | 31.794 |
| EXCESS_ADJUSTED_AMT | -5.9819 | 5.624 | -1.064 | 0.288 | -17.006 | 5.043 |
| EXCESS_AVAILABLE | 5.9819 | 5.624 | 1.064 | 0.288 | -5.043 | 17.006 |
| FOIR | -2.936e-07 | 4.5e-07 | -0.652 | 0.514 | -1.18e-06 | 5.89e-07 |
| LAST_RECEIPT_AMOUNT | 1.114e-08 | 4.38e-09 | 2.541 | 0.011 | 2.55e-09 | 1.97e-08 |
| LATEST_TRANSACTION_MONTH | -0.0458 | 0.001 | -68.349 | 0.000 | -0.047 | -0.044 |
| LOAN_AMT | -4.296e-09 | 2.21e-09 | -1.941 | 0.052 | -8.63e-09 | 4.25e-11 |
| MAX_EMI_AMOUNT | 2.063e-06 | 2.49e-06 | 0.830 | 0.406 | -2.81e-06 | 6.93e-06 |
| MIN_EMI_AMOUNT | -2.001e-06 | 2.49e-06 | -0.805 | 0.421 | -6.88e-06 | 2.87e-06 |
| MONTHOPENING | -2.505e-10 | 5.08e-09 | -0.049 | 0.961 | -1.02e-08 | 9.71e-09 |
| NET_LTV | 0.0001 | 9.18e-05 | 1.126 | 0.260 | -7.66e-05 | 0.000 |
| NUM_EMI_CHANGES | 0.0087 | 0.001 | 8.679 | 0.000 | 0.007 | 0.011 |
| NUM_LOW_FREQ_TRANSACTIONS | -0.0104 | 0.001 | -9.069 | 0.000 | -0.013 | -0.008 |
| ORIGNAL_INTEREST_RATE | -0.0281 | 0.003 | -8.683 | 0.000 | -0.034 | -0.022 |
| ORIGNAL_TENOR | 0.0010 | 4.43e-05 | 21.842 | 0.000 | 0.001 | 0.001 |
| OUTSTANDING_PRINCIPAL | -3.546e-09 | 4.17e-09 | -0.850 | 0.395 | -1.17e-08 | 4.63e-09 |
| PAID_INTEREST | 6.597e-09 | 7.38e-09 | 0.894 | 0.371 | -7.87e-09 | 2.11e-08 |
| PAID_PRINCIPAL | 1.434e-10 | 2.66e-10 | 0.539 | 0.590 | -3.78e-10 | 6.65e-10 |
| PRE_EMI_DUEAMT | 5.557e-07 | 7.76e-07 | 0.716 | 0.474 | -9.66e-07 | 2.08e-06 |
| PRE_EMI_OS_AMOUNT | -6.669e-07 | 7.82e-07 | -0.853 | 0.394 | -2.2e-06 | 8.65e-07 |
| PRE_EMI_RECEIVED_AMT | -5.683e-07 | 7.77e-07 | -0.732 | 0.464 | -2.09e-06 | 9.54e-07 |

| | | | |
|---|---|---|---|
| Omnibus: | 7445.957 | Durbin-Watson: | 1.901 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 31249.794 |
| Skew: | 1.824 | Prob(JB): | 0.00 |
| Kurtosis: | 7.933 | Cond. No. | 1.04e+16 |

Both the results are compared with each other and those features whose P-value is greater than α are dropped as they are not significant for predicting the foreclosure. With the help of multivariate analysis, we will get significant features that would result a better result for model building.

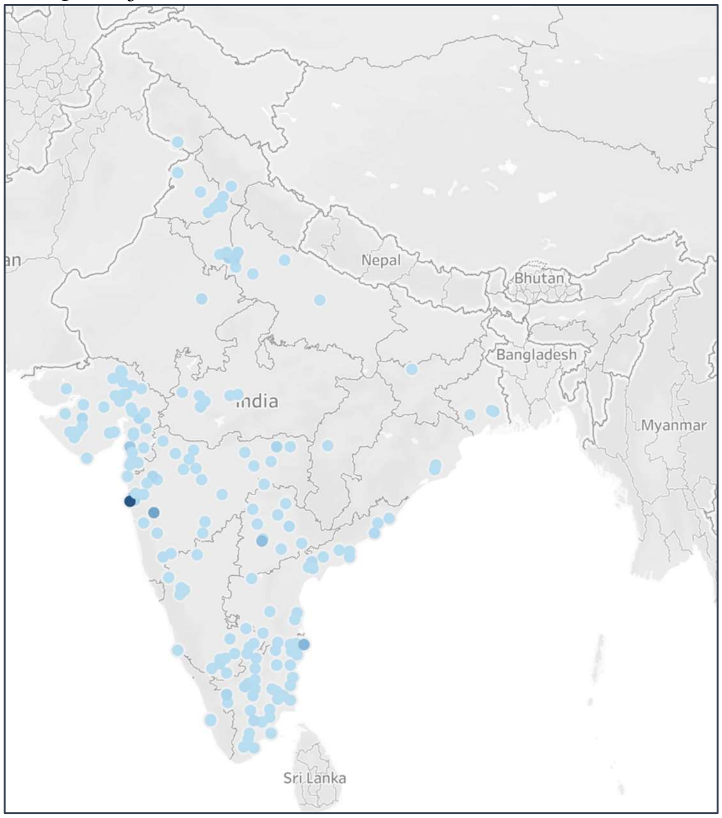
**Bivariate Analysis**:

*<u>Analysis of Foreclosure trends in Cities</u>*:



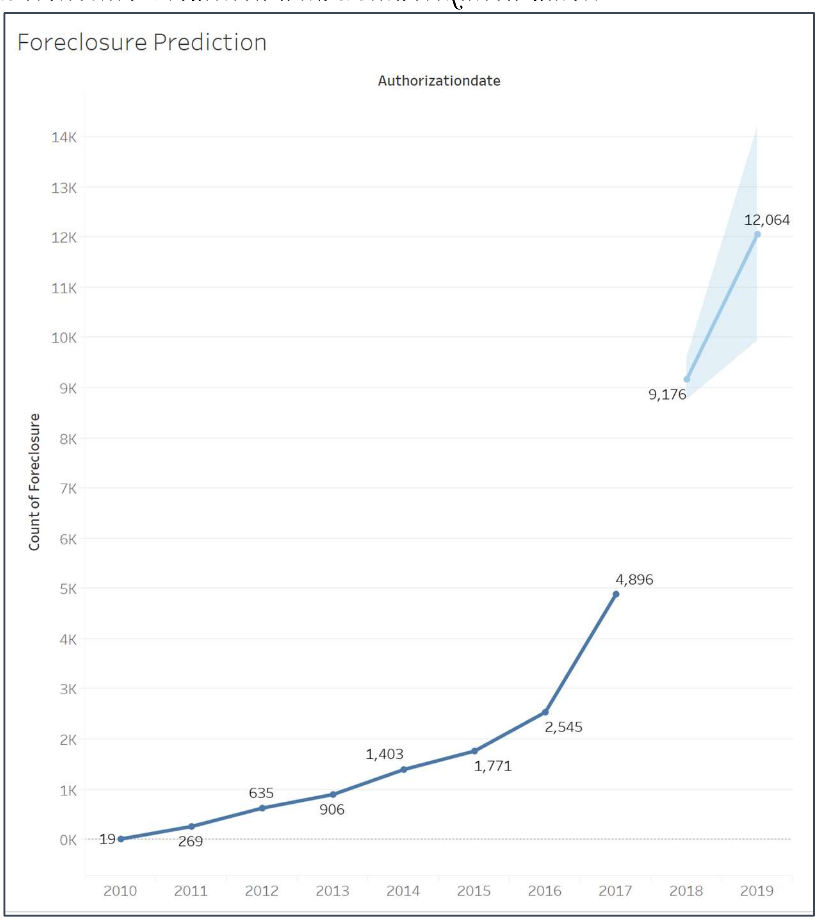

In terms of loan and foreclosure Mumbai is on the top followed by Pune, Bangalore and Hyderabad. Metro cities of India is having a high movement of loans as well as foreclosure where in this data we could find there are places like Pondicherry, Coimbatore, Madurai, Nellore, Mysore, Tirunelveli, etc. which is not a falling under the title of cities their foreclosure rates are very high. Erode is a town where the count of foreclosure is 43. NBFC has to implement different approaches to the borrowers of these under developed placed where the foreclosure rates are impacting high on the business.

## *Analysis of Cities and Loans*




As per the data shared, the maximum loans are taken from borrowers in Mumbai and Hyderabad. The geological graphs between city and loan proportion shows that the majority of the borrowers are from southern part of India and western ghats of India. There are no borrowers from Eastern part of India. Northern and Central India the borrowers are very minimal. The geological analysis recommends the expansion of NBFC to eastern part and the central part of India.

## *Foreclosure Prediction with Authorization dates:*

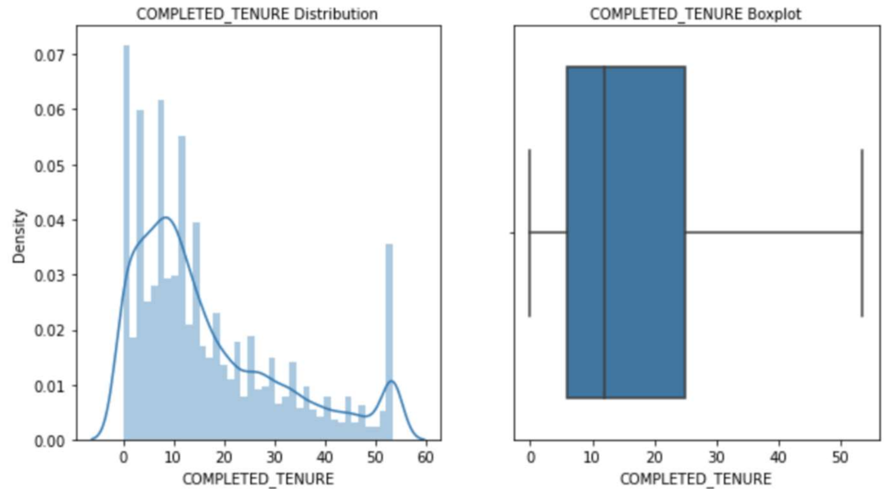



The data shared by NBFC is having authorization of loan from 2010 till 2018 where the data is compared with the volume of foreclosures happening for different year. Every year the foreclosure is showing an increase in volume of closures happening. For the year 2018 to 2019 the expected foreclosures that can happen falls between 9176 to 12064. This foreclosure proceedings are a trigger for the business as currently RBI has banned NBFC in taking penalty from the borrowers. Considering this, NBFC need to revalidate the borrowers CBIL Score and their annual earning as well as the Net LTV ratio of the borrowers. NBFC should calculate the frequency of missed EMI from the borrower and highly recommended to take the next further initiatives that could bring down the non-defaulters in becoming defaulters.

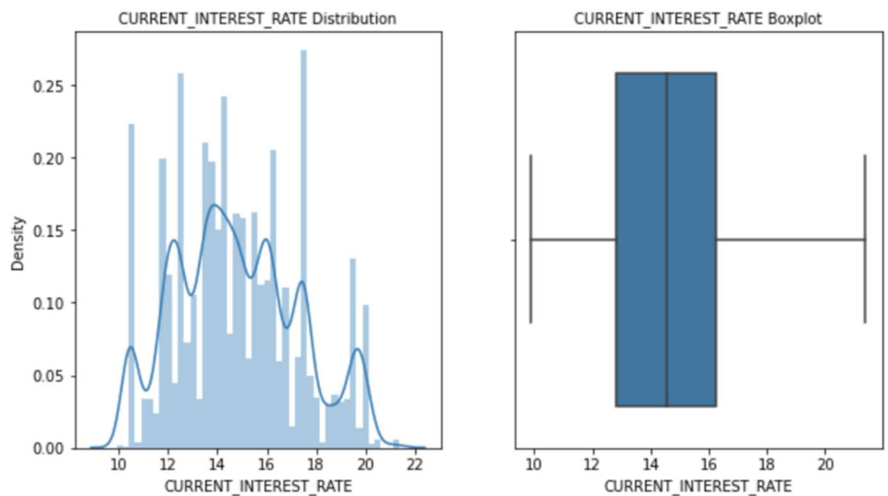
# Univariate Analysis:

Box plot and the distribution plots are used to identify the skewness of the data and outliers of the account.

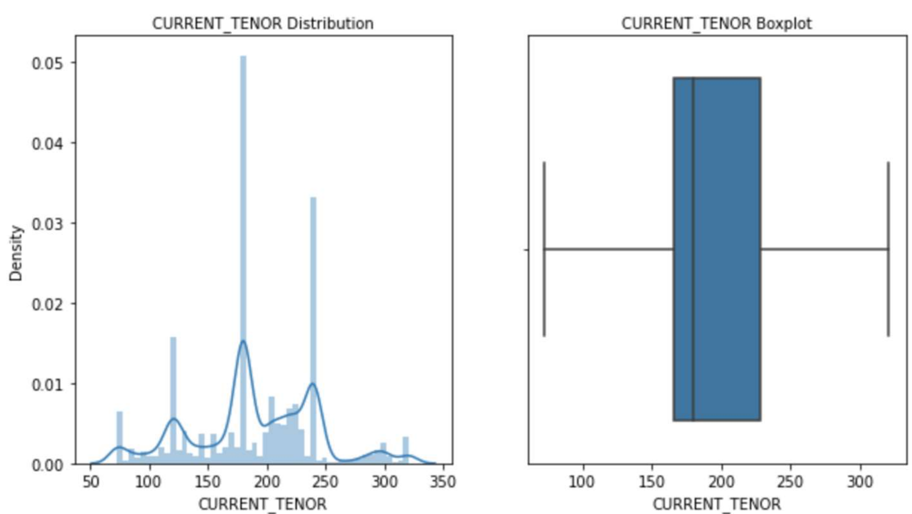
***Completed Tenure***: This is one of the significant features that will help in getting a better prediction on foreclosure. Out of original tenure the tenure that the borrower has paid the EMI.



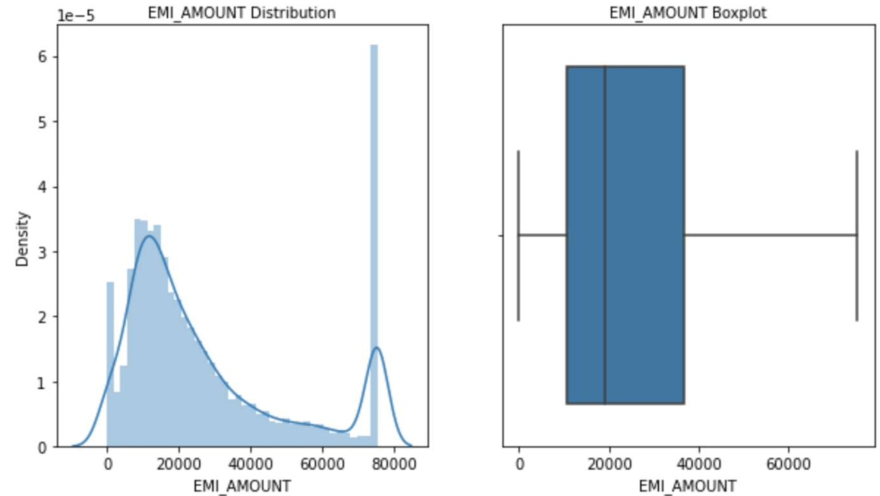

***Current Interest Rate:*** The interest rate is a fluctuating change made with depend with repo rates. This one of the important aspects for Indian economy.



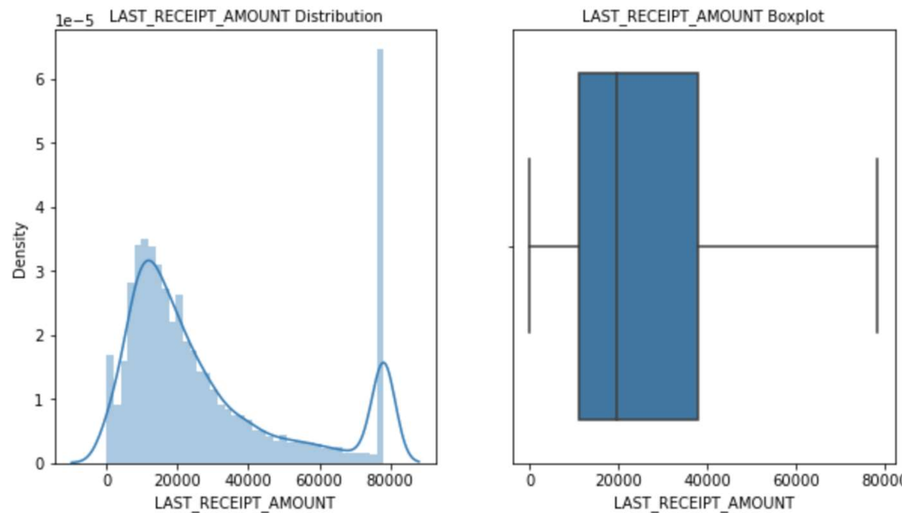

**Current Tenor**: Available tenor with the borrowers. This would help in calculating the foreclosure prediction whether the borrower can fall from defaulters to non-defaulters.



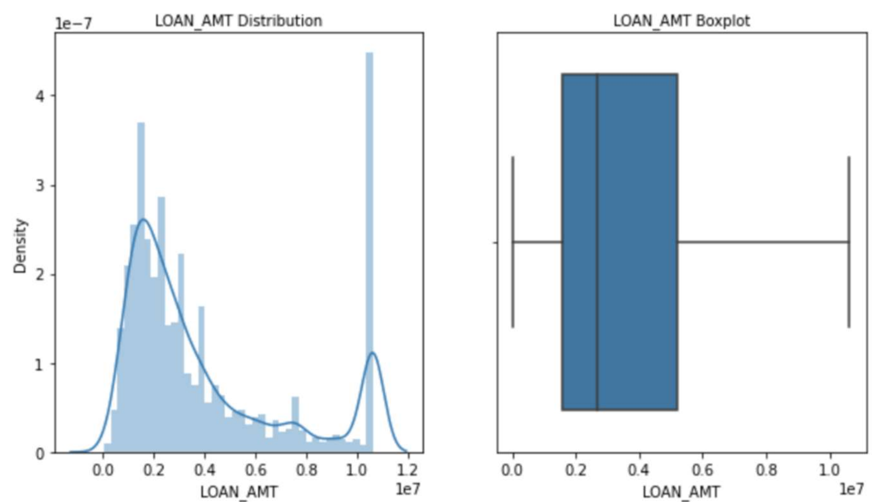
**EMI Amount:** The EMI that the borrower is required to be paid to NBFC



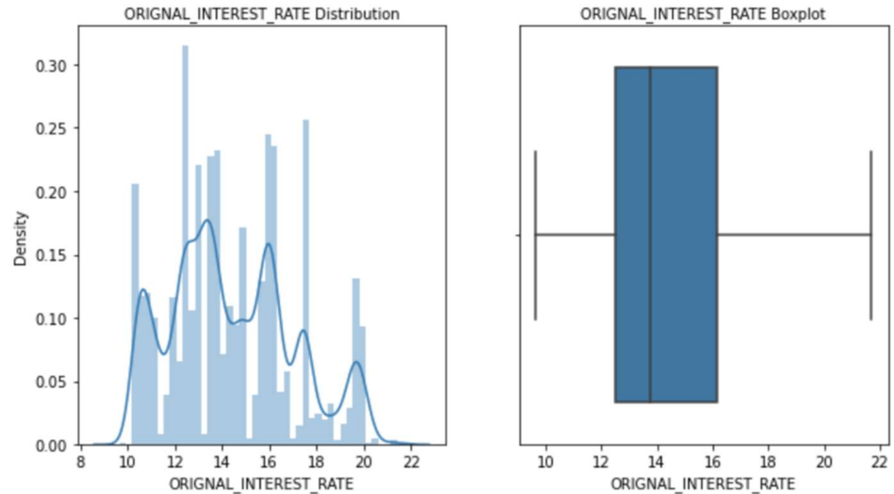

**Last Receipt Amount**: The amount that was paid by the borrower to NBFC.



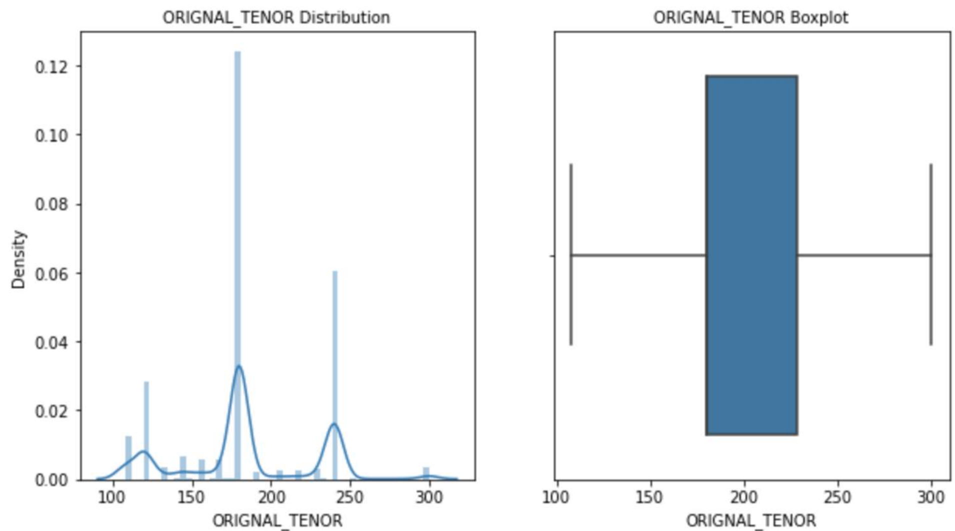

**Loan Amount**: The amount that the borrower owes to NBFC. This is an important feature that would help in predicting the foreclosure of the borrower from default to non-defaulters.



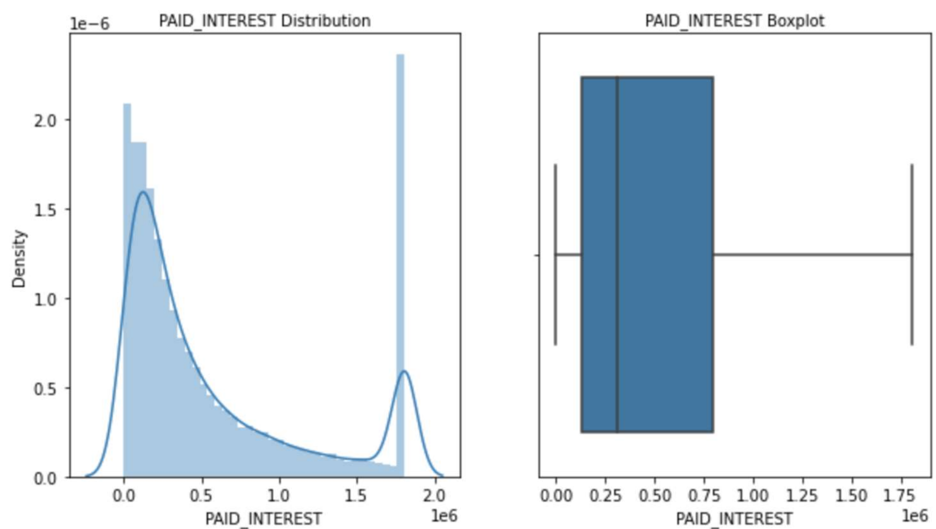
**Original Interest Rate**: The default interest is generally higher than the original interest as it reflects aggravation in the financial risk of the borrower



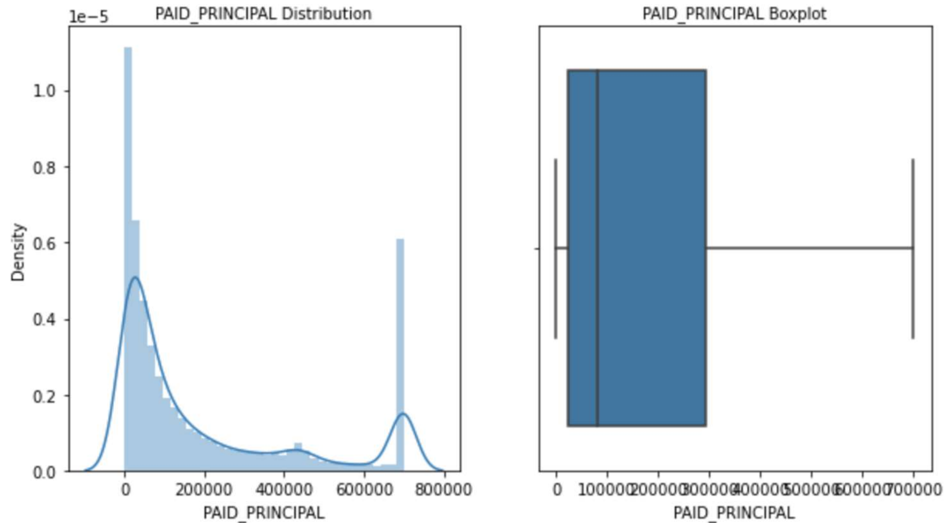

**Loan Amount**: The amount that the borrower owes to NBFC. This is an important feature that would help in predicting the foreclosure of the borrower from default to non-defaulters.



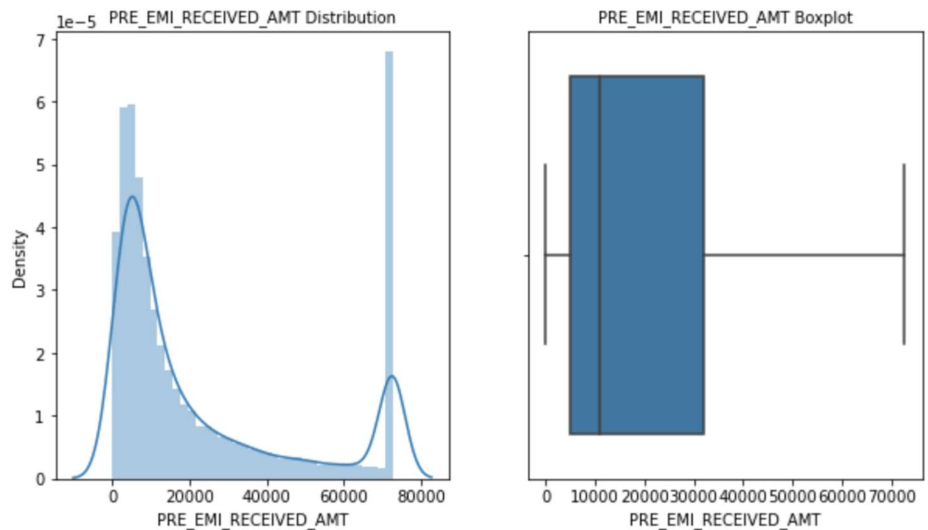

**Paid Interest**: The amount that the borrower has paid to NBFC as interest that excludes from the loan amount and principal amount.



**Paid Principal**: The amount that the borrower has paid towards the principal amount out of the loan amount




**Pre EMI received**: The amount that the borrower has paid to NBFC as interest on the disbursement amount.




# 3. Data Cleaning & Pre-Processing

## <u>Missing Value and Treatment:</u>

There were 10 constraints in the dataset where they need treatment for missing values. The features like “NPA in last month” and “NPA in current month”, 99% of the data are missing. Hence, we will remove these fields by considering the rule of missing value. If any data is missing for more than 70%, we can remove them as the impute of the variable is not possible. “Scheme ID” and “Customer ID” are interlinked where 281 customer details are missing. These been a unique variable representing the customer. Impute of the data in mean or mode will present an incorrect forecasting. Hence these two features are been removed from our study.

The features like “Min EMI Amount”, “Max EMI Amount” and “Difference of Min and Max” are linked with each other. 0.4% of the data are missing. We will remove the NAN values of the three features. The treatment for the rest of the features is done with the assistance of package named Simple Imputer. This would impute the missing values with the median as strategy.

## Removal of variables (if applicable)

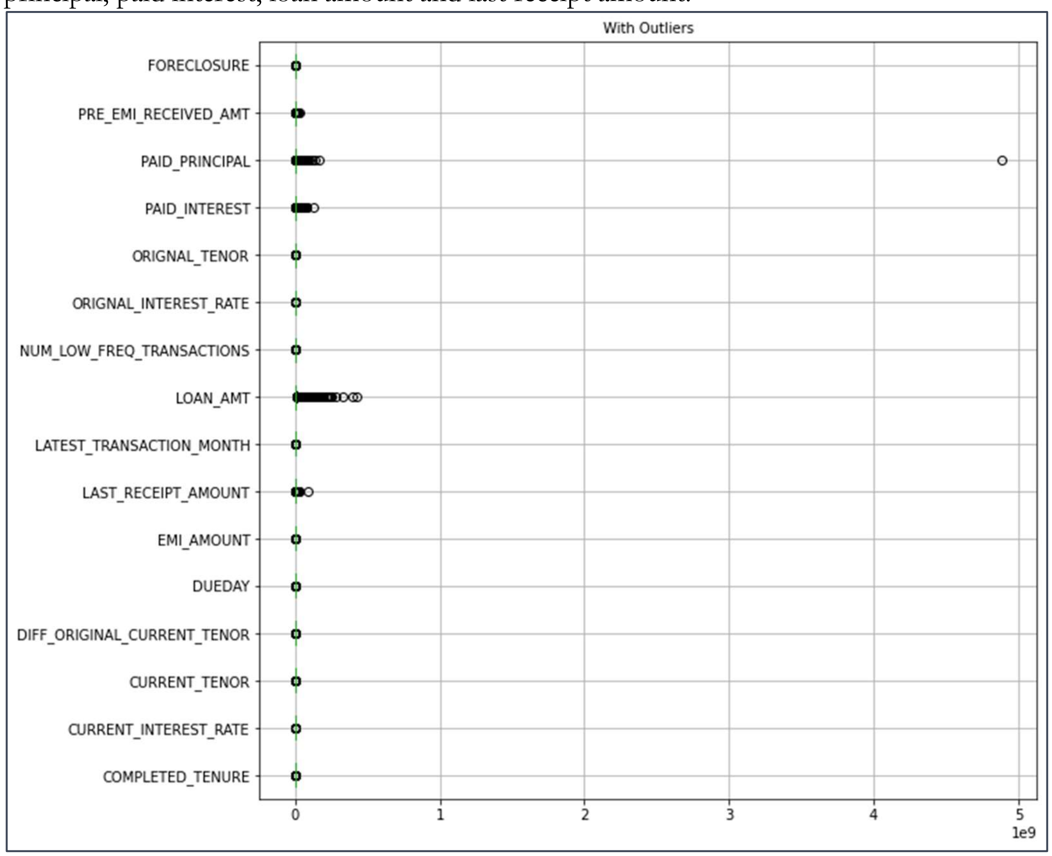
The correlation matrix and the logistic regression stats model shows that there is multicollinearity. Hence there are certain variables that need to be removed from the dataset as the highly correlated and need to be removed from the analysis. If the correlated variable near to 1 that need to be removed from the data. Those features whose P value is greater than the apha, those are not significant features as they will not give a better results. We will consider only those significant variables from the dataset that would lead to a better prediction of the model.

## Scaling of the Dataset: -

The dataset that we are handling are not normalized. We have month format whose values are numeric and we have digital variables. If we are processing the data without scaling, this would impact our results. Hence scaling the data is required before stepping into the modelling. We use standard scaler function from the package sklearn. Normalization is a process of structuring the relational database in accordance with series.

## Outlier and Treatment

After removing the insignificant features from the data, we need to check whether there are any outliers present in the data or not. The box plot helps in identifying the available outliers. From the significant variable we have outliers in features like paid principal, paid interest, loan amount and last receipt amount.



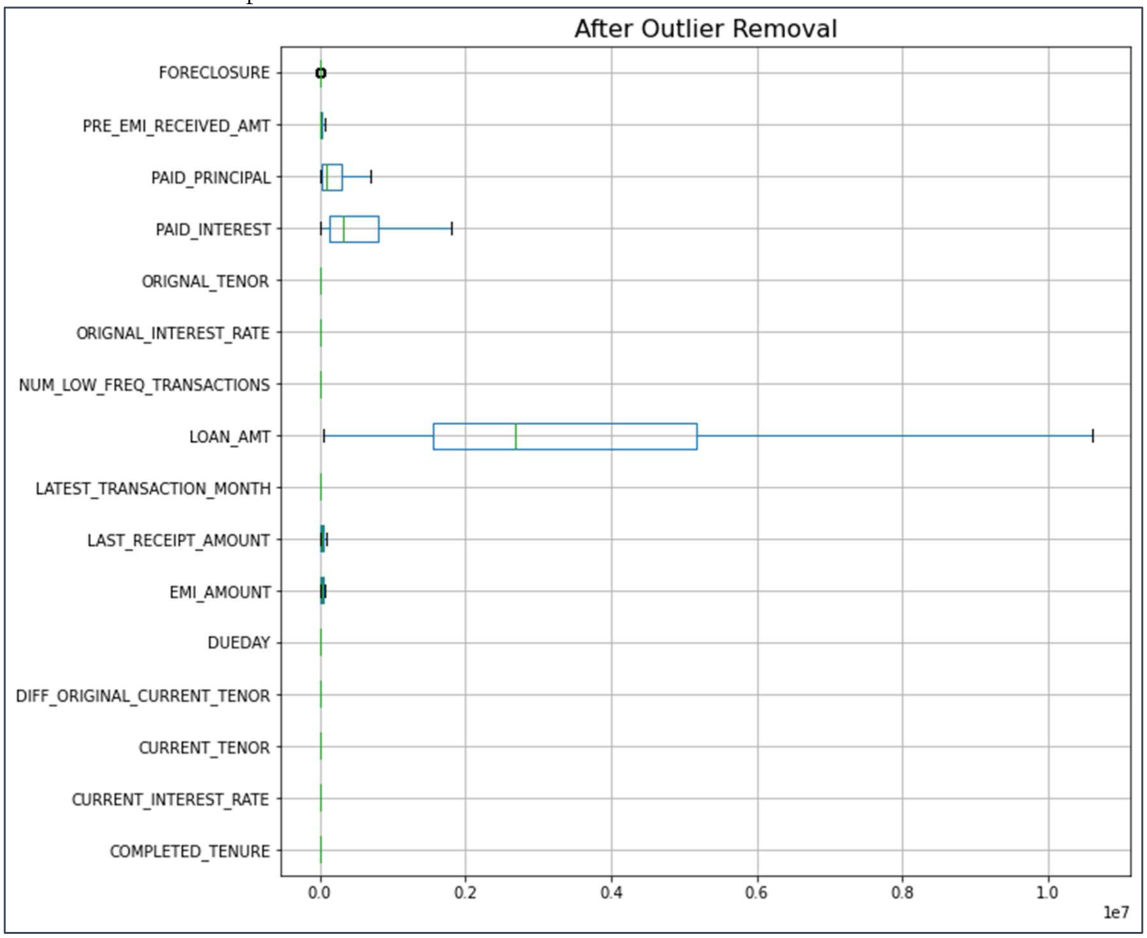

The treatment of the outliers is processed by two methods like Inter-quartile-range method and Z score. Here we have used the treatment with Inter-quartile range (IQR) methods. We have used the Inter Quartile Range method here to treat the outliers. The quartiles are divided into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

- Q1 is the "middle" value in the first half of the rank-ordered data set
- Q2 is the median value in the set.
- Q3 is the "middle" value in the second half of the rank-ordered data set.

The interquartile range is equal to Q3 minus Q1

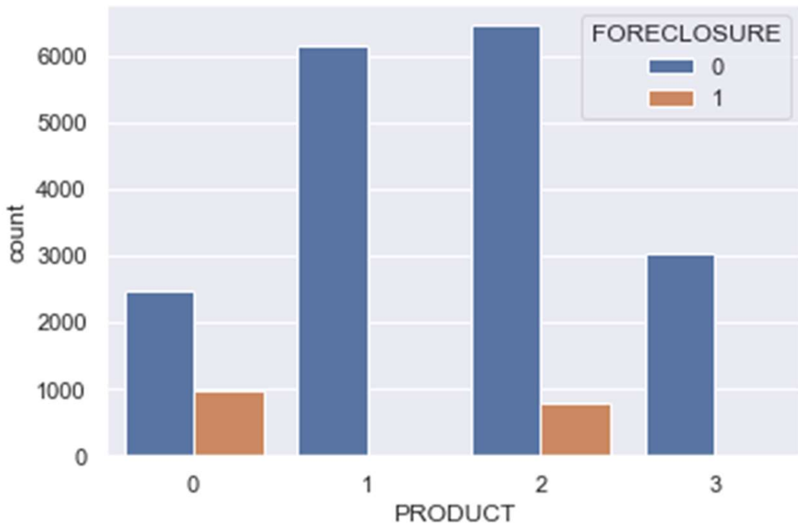
The data set is highly multicollinearity, hence specifically the variable "Paid Principal" the outlier is very high compared with the other features. Hence to treat them, outlier treatment is highly recommended.

Once the treatment is completed, once again check plot the boxplot to check if any outliers are still present in the data. There are no outliers present in the data after the treatment.




## Variable transformation (if applicable)

We have to encode the variable Product from categorical to numeric variable. The product variable in the dataset was divided into four slots as HL (Home Loan), LAP (Loan against Property), STHL (Short-term Home Loan) and STLAP (Short term loan against property). They are encoded as 0,1,2 and 3 respectively. The modelling and the predictions are done based on the products like HL and STHL.




The four products – HL (Home loan) is 0, Loan against property (LAP) is 1, STHL is 2 and STLAP is 4. As per the data shared to us the highest volume of products that NBFC moves is STHL and LAP. Where the data on foreclosure along with the products shows that the maximum foreclosure is happened for HL and STHL. Hence, we will be giving the prediction to NBFC for these two features. 90% of the data are defaulters and 9% are non-defaulters.

# 4. Model Building

The dataset has to be split into train and test dataset. The splitting of train and test are done with the help of package from sklearn. The prediction is done for product class specifically for HL and STHL. Hence, the data is split into Train and test with a ratio of 70:30.

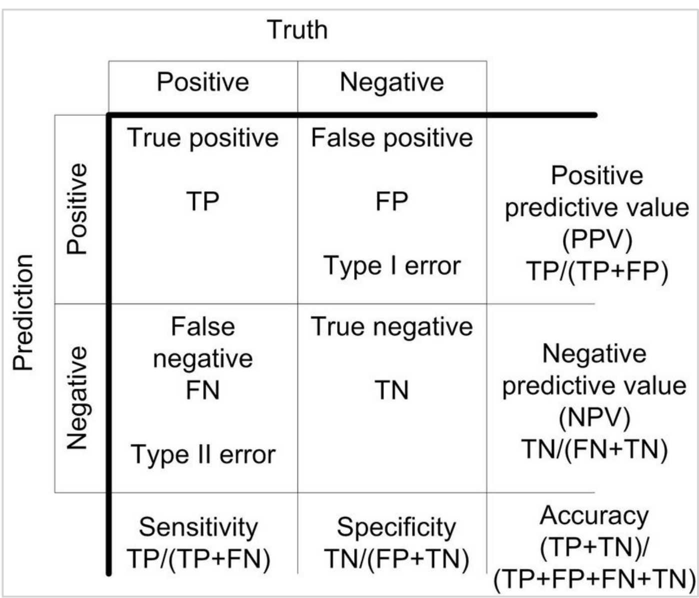
X is classified as the significant featured variables and Y is classified as the target variables that is "Foreclosure".

**Modelling Methods used**:
In this study we have used the below modelling techniques: -

1. Gaussian Naïve Bayes Model
2. Logistic Regression
3. Decision Tree
4. KNN (K-nearest neighbor)
5. SVM (Support Vector Machine)
6. Linear Discriminant Analysis (LDA)
7. Ada Boosting
8. XG Boost
9. Cross Validation

| Measures/Model | Sensitivity | Specificity | Precision | Negative Predictive Value | False Positive Rate | False Discovery Rate | False Negative Rate | Accuracy | F1 Score | Matthews Correlation Coefficient | AUC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gaussian Naïve Bayse_Train | 0.88 | 0.40 | 0.88 | 0.40 | 0.60 | 0.12 | 0.12 | 0.80 | 0.88 | 0.28 | 0.75 |
| Gaussian Naïve Bayse_Test | 0.88 | 0.43 | 0.89 | 0.40 | 0.57 | 0.11 | 0.12 | 0.81 | 0.88 | 0.30 | 0.74 |
| Logistic Regression Train | 0.85 | 0.29 | 0.91 | 0.18 | 0.71 | 0.09 | 0.15 | 0.79 | 0.88 | 0.11 | 0.55 |
| Logistic Regression Test | 0.84 | 0.29 | 0.92 | 0.16 | 0.71 | 0.08 | 0.16 | 0.79 | 0.88 | 0.10 | 0.55 |
| Decision Tree Train | 0.93 | 0.62 | 0.92 | 0.65 | 0.38 | 0.08 | 0.07 | 0.87 | 0.92 | 0.56 | 0.79 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Decision Tree Test** | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| **KNN Train** | 0.90 | 0.71 | 0.96 | 0.44 | 0.29 | 0.04 | 0.10 | 0.88 | 0.93 | 0.50 | 0.91 |
| **KNN Test** | 0.87 | 0.52 | 0.94 | 0.33 | 0.48 | 0.06 | 0.13 | 0.83 | 0.90 | 0.32 | 0.75 |
| **SVM Train** | 0.86 | 0.55 | 0.97 | 0.17 | 0.45 | 0.03 | 0.14 | 0.84 | 0.91 | 0.24 | |
| **SVM Test** | 0.85 | 0.56 | 0.97 | 0.17 | 0.44 | 0.03 | 0.15 | 0.84 | 0.91 | 0.24 | |
| **Linear Discriminant Analysis Train** | 0.88 | 0.57 | 0.95 | 0.37 | 0.43 | 0.05 | 0.12 | 0.85 | 0.91 | 0.38 | 0.86 |
| **Linear Discriminant Analysis Test** | 0.88 | 0.61 | 0.95 | 0.38 | 0.39 | 0.05 | 0.12 | 0.85 | 0.92 | 0.41 | 0.86 |
| **AdaBoost Train** | 0.93 | 0.83 | 0.97 | 0.63 | 0.17 | 0.03 | 0.07 | 0.92 | 0.95 | 0.68 | 0.94 |
| **AdaBoost Test** | 0.92 | 0.83 | 0.97 | 0.61 | 0.17 | 0.03 | 0.08 | 0.91 | 0.95 | 0.66 | 0.93 |
| **XGBoost Train** | 0.94 | 0.84 | 0.97 | 0.67 | 0.16 | 0.03 | 0.06 | 0.92 | 0.96 | 0.71 | 0.95 |
| **XGBoost Test** | 0.94 | 0.84 | 0.97 | 0.70 | 0.16 | 0.03 | 0.06 | 0.93 | 0.96 | 0.73 | 0.95 |




We use confusion matrix (error matrix), specified table layout that visualized the performance of the algorithms. The two-dimension table (actual and predicted) and identical classes in both dimensions. The matrix defines the class of matrix as defaulters and non- defaulters. Here in this dataset, we are predicting the probability of foreclosure for non- defaulter that is 1 category listed in target variable.

As per our analysis, we have XG Booster as the best model compared with others. The train data and the test data give an accuracy of 94%. The data says that 94% of the data is predicted correctly as defaulters and 16% of the data has predicted incorrectly as defaulters when they fall as non-defaulters. 97% of the defaulters are predicted correctly and 84% of the data have been correctly predicted as non-defaulters.

## XG Boost Models:

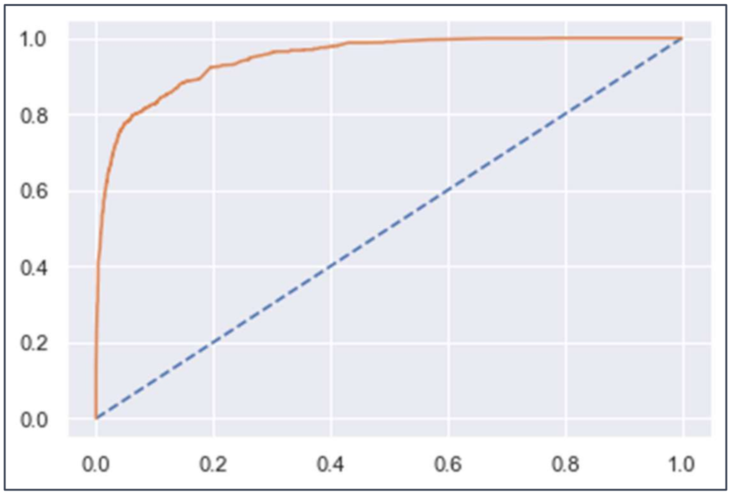
XGBoost or Xtreme Gradient Boosting is an improvised version of Gradient Boosting. Gradient boosting combines the predictions from multiple decision trees to generate the final prediction. The trees in XG Boost are build sequent trying to correct the errors of the previous trees. The below are the results driven from our models.

**Train Dataset:**

| | 0 | 1 |
|---|---|---|
| 0 | 6109 | 161 |
| 1 | 406 | 840 |

Accuracy of the model is 0.924
Specificity of the model is 0.84
Negative Predicted Value is 0.67
F1 score is 0.96 and F1 score for non-defaulters is 0.75

AUC & ROC




AUC: 0.948

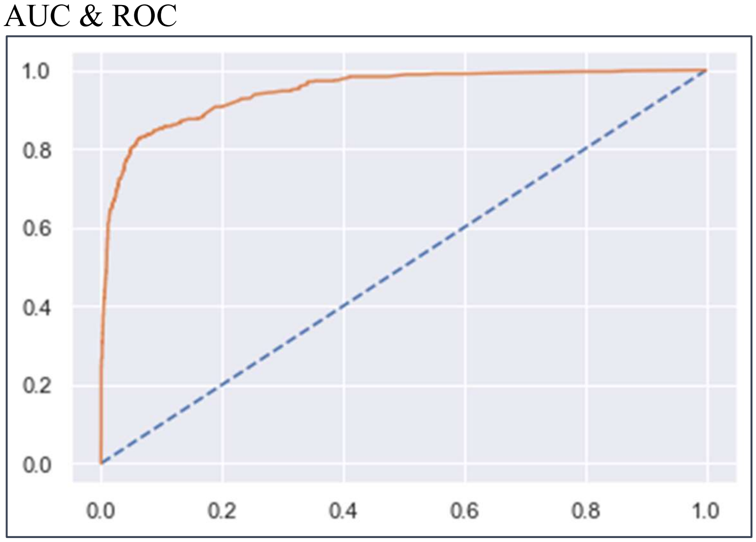
The model is steeper and gives a strong model

Test Data:

| | 0 | 1 |
|---|---|---|
| 0 | 2603 | 72 |
| 1 | 164 | 383 |

The Accuracy of the model is 0.93. The specificity of the model is 0.84. The precision of the model is 0.97.The correlation of the model is 0.73. Correlation shows that 0 been the worst and towards 1 the best.

AUC & ROC




AUC: 0.947

The model is steeper and gives a strong model. When compared with the train model, both the model gives the same AUC and ROC curve. Hence, we can confirm that the train model and the test model has predicted perfectly.

**Model Tuning Measures:**

*Tuning* is the process of maximizing a model's performance without overfitting or creating too high of a variance. As our dataset is imbalance, we use SMOTE (Synthetic Minority Oversampling Technique) and Cross Validation for the models.

*Cross Validation* is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting.

While taking up the scores of Cross Validation we have got the accuracy of the model as 100%, this shows that the model is overfitted and the correction of this is done using SMOTE. We will rebuild all the models by taking SMOTE and will regenerate the confusion matrix for all the resampled data using SMOTE.

SMOTE:

| Measures/Model | Sensitivity | Specificity | Precision | Negative Predictive Value | False Positive Rate | False Discovery Rate | False Negative Rate | Accuracy | F1 Score | Matthews Correlation Coefficient |
|---|---|---|---|---|---|---|---|---|---|---|
| **SMOTE Naïve Bayse_Train** | 0.62 | 0.75 | 0.83 | 0.49 | 0.25 | 0.17 | 0.38 | 0.66 | 0.71 | 0.34 |
| **SMOTE Naïve Bayse_Test** | 0.89 | 0.39 | 0.84 | 0.51 | 0.61 | 0.16 | 0.11 | 0.78 | 0.87 | 0.32 |
| **SMOTE KNN Train** | 0.96 | 0.84 | 0.82 | 0.97 | 0.16 | 0.18 | 0.04 | 0.89 | 0.88 | 0.79 |
| **SMOTE KNN Test** | 0.91 | 0.34 | 0.75 | 0.64 | 0.66 | 0.25 | 0.09 | 0.73 | 0.82 | 0.31 |
| **SMOTE SVM Train** | 0.71 | 0.74 | 0.76 | 0.68 | 0.26 | 0.24 | 0.29 | 0.72 | 0.73 | 0.44 |
| **SMOTE SVM Test** | 0.92 | 0.36 | 0.76 | 0.66 | 0.64 | 0.24 | 0.08 | 0.74 | 0.83 | 0.34 |
| **SMOTE LDA Train** | 0.85 | 0.81 | 0.79 | 0.86 | 0.19 | 0.21 | 0.15 | 0.83 | 0.82 | 0.65 |
| **XG Boost Train** | 0.93 | 0.91 | 0.91 | 0.93 | 0.09 | 0.09 | 0.07 | 0.92 | 0.92 | 0.84 |
| **XG Boost Test** | 0.95 | 0.59 | 0.89 | 0.78 | 0.41 | 0.11 | 0.05 | 0.87 | 0.92 | 0.60 |

Here we have built the model for the below techniques:

1. Naïve Bayes Model
2. KNN
3. SVM
4. LDA
5. XG Boost

Comparing with all the model we, have better correlation coefficient is for XG Boost model. We have 0.84 correlation ecoefficiency on train model. When the same was implicated on the test model we have 0.60 correlation. We could find that all the models have train models are getting better results while compared to test.

**XG Boost Model:**

**Train Data**

| | 0 | 1 |
|---|---|---|
| 0 | 5679 | 591 |
| 1 | 445 | 5825 |

Accuracy of the train data is 0.92 where the specificity is 0.91 and negative predicted value is 0.93. This shows that our prediction of default is correct. F1 score is 0.92 bringing correlation of 0.84

**Test Data**:

| | 0 | 1 |
|---|---|---|
| 0 | 2377 | 298 |
| 1 | 123 | 424 |

Accuracy of the train data is 0.87 where the sensitivity is 0.95 and negative predicted value is 0.78. This shows that our prediction of default is correct. F1 score is 0.92 bringing correlation of 0.60

**Cross Validation:**

Cross validation is a technique for assessing how the statistical analysis generalizes to an independent data set. It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data. Using cross-validation, there are high chances that we can detect over-fitting with ease.

There are several cross validation techniques such as :-
1. K-Fold Cross Validation
2. Leave P-out Cross Validation
3. Leave One-out Cross Validation
4. Repeated Random Sub-Sampling Method
5. Holdout Method

Here we use the K-Fold Cross validation techniques for the below models that we have performed while processing the predictive models by using SMOTE.

1. Naïve Bayes Model
2. KNN
3. SVM
4. XG Boost

Comparing the results of the all models:

| Measures/Model | Std Dev | Variance | Final Mean |
|---|---|---|---|
| **SMOTE Naïve Bayse_Train** | 1.039 | 0.0001079 | 66.196 |
| **SMOTE Naïve Bayse_Test** | 2.687 | 0.0007221 | 80.356 |
| **SMOTE KNN Train** | 1.156 | 0.0001336 | 84.378 |
| **SMOTE KNN Test** | 1.193 | 0.0001423 | 83.582 |
| **SMOTE SVM Train** | 1.363 | 0.0001857 | 72.049 |
| **SMOTE SVM Test** | 1.604 | 0.0002573 | 83.583 |
| **SMOTE LDA Train** | 1.635 | 0.0002672 | 82.48 |
| **SMOTE LDA Test** | 1.847 | 0.0003413 | 84.824 |
| **XG Boost Train** | 3.388 | 0.001148 | 90.271 |
| **XG Boost Test** | 1.544 | 0.0002384 | 89.852 |

We would be considering XG boost as the best model as we have the best scores in K fold model.

Train set:

```
Accuracy: Final mean:90.271%, Final standard deviation:(3.388%)
Accuracies from each of the 5 folds using XGB_SM_model: [0.83333333 0.83971292 0.9146
7305 0.91786284 0.91866029 0.91307815
 0.9354067  0.92583732 0.92105263 0.90749601]
Variance of XGB_SM_model accuracies: 0.0011479692416484155
```

Test Set:

```
Accuracy: Final mean:89.852%, Final standard deviation:(1.544%)
Accuracies from each of the 5 folds using XGB_SM_model: [0.86996904 0.89164087 0.9006
2112 0.89751553 0.88509317 0.89130435
 0.93167702 0.90372671 0.90372671 0.90993789]
Variance of XGB_SM_model accuracies: 0.00023842245563906726
```

## 5. Model validation

We have evaluated the model on the basis of their Correlation coefficient score, Sensitivity, Recall, F1 Scores, Precision as well as the cross-validation scores.

Evaluation of the model are performed as below:

| | |
|---|---|
| Sensitivity | TPR = TP / (TP + FN) |
| Specificity | SPC = TN / (FP + TN) |
| Precision | PPV = TP / (TP + FP) |
| Negative Predictive Value | NPV = TN / (TN + FN) |
| False Positive Rate | FPR = FP / (FP + TN) |
| False Discovery Rate | FDR = FP / (FP + TP) |
| False Negative Rate | FNR = FN / (FN + TP) |
| Accuracy | ACC = (TP + TN) / (P + N) |
| F1 Score | F1 = 2TP / (2TP + FP + FN) |
| Matthews Correlation Coefficient | TP*TN - FP*FN / sqrt((TP+FP) *(TP+FN)*(TN+FP)*(TN+FN)) |

**Matthews Correlation Coefficient** is more reliable statistical rate which produce high score only if the prediction is obtained good results in all of the four-confusion matrix category. Analysis of the correlation is taken in a 0 to 1 point scale. 0 is the worst and 1 been the best model. This is one of the important factor considered other than Accuracy and AUC and ROC curve.

When we do a comparison between all the models, we have XG Boosting technique been outstanding among other models in terms of correlation as well as accuracy, F1 score, Precision, etc.

**Comparison of all the models**:

| Measures/Model | Sensitivity | Specificity | Precision | Negative Predictive Value | False Positive Rate | False Discovery Rate | False Negative Rate | Accuracy | F1 Score | Matthews Correlation Coefficient |
|---|---|---|---|---|---|---|---|---|---|---|
| Gaussian Naïve Bayse_Train | 0.88 | 0.40 | 0.88 | 0.40 | 0.60 | 0.12 | 0.12 | 0.80 | 0.88 | 0.28 |
| Gaussian Naïve Bayse_Test | 0.88 | 0.43 | 0.89 | 0.40 | 0.57 | 0.11 | 0.12 | 0.81 | 0.88 | 0.30 |
| Logistic Regression_Train | 0.85 | 0.29 | 0.91 | 0.18 | 0.71 | 0.09 | 0.15 | 0.79 | 0.88 | 0.11 |
| Logistic Regression_Test | 0.84 | 0.29 | 0.92 | 0.16 | 0.71 | 0.08 | 0.16 | 0.79 | 0.88 | 0.10 |
| Decision Tree_Train | 0.93 | 0.62 | 0.92 | 0.65 | 0.38 | 0.08 | 0.07 | 0.87 | 0.92 | 0.56 |
| Decision Tree_Test | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| KNN_Train | 0.90 | 0.71 | 0.96 | 0.44 | 0.29 | 0.04 | 0.10 | 0.88 | 0.93 | 0.50 |
| KNN_Test | 0.87 | 0.52 | 0.94 | 0.33 | 0.48 | 0.06 | 0.13 | 0.83 | 0.90 | 0.32 |
| SVM_Train | 0.86 | 0.55 | 0.97 | 0.17 | 0.45 | 0.03 | 0.14 | 0.84 | 0.91 | 0.24 |
| SVM_Test | 0.85 | 0.56 | 0.97 | 0.17 | 0.44 | 0.03 | 0.15 | 0.84 | 0.91 | 0.24 |
| Linear Discriminant Analysis_Train | 0.88 | 0.57 | 0.95 | 0.37 | 0.43 | 0.05 | 0.12 | 0.85 | 0.91 | 0.38 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Linear Discriminant Analysis_Test** | 0.88 | 0.61 | 0.95 | 0.38 | 0.39 | 0.05 | 0.12 | 0.85 | 0.92 | 0.41 |
| **AdaBoost_Train** | 0.93 | 0.83 | 0.97 | 0.63 | 0.17 | 0.03 | 0.07 | 0.92 | 0.95 | 0.68 |
| **AdaBoost_Test** | 0.92 | 0.83 | 0.97 | 0.61 | 0.17 | 0.03 | 0.08 | 0.91 | 0.95 | 0.66 |
| **XGBoost_Train** | 0.94 | 0.84 | 0.97 | 0.67 | 0.16 | 0.03 | 0.06 | 0.92 | 0.96 | 0.71 |
| **XGBoost_Test** | 0.94 | 0.84 | 0.97 | 0.70 | 0.16 | 0.03 | 0.06 | 0.93 | 0.96 | 0.73 |
| **SMOTE_Naïve Bayse_Train** | 0.62 | 0.75 | 0.83 | 0.49 | 0.25 | 0.17 | 0.38 | 0.66 | 0.71 | 0.34 |
| **SMOTE_Naïve Bayse_Test** | 0.89 | 0.39 | 0.84 | 0.51 | 0.61 | 0.16 | 0.11 | 0.78 | 0.87 | 0.32 |
| **SMOTE_KNN_Train** | 0.96 | 0.84 | 0.82 | 0.97 | 0.16 | 0.18 | 0.04 | 0.89 | 0.88 | 0.79 |
| **SMOTE_KNN_Test** | 0.91 | 0.34 | 0.75 | 0.64 | 0.66 | 0.25 | 0.09 | 0.73 | 0.82 | 0.31 |
| **SMOTE_SVM_Train** | 0.71 | 0.74 | 0.76 | 0.68 | 0.26 | 0.24 | 0.29 | 0.72 | 0.73 | 0.44 |
| **SMOTE_SVM_Test** | 0.92 | 0.36 | 0.76 | 0.66 | 0.64 | 0.24 | 0.08 | 0.74 | 0.83 | 0.34 |
| **SMOTE_LDA_Train** | 0.85 | 0.81 | 0.79 | 0.86 | 0.19 | 0.21 | 0.15 | 0.83 | 0.82 | 0.65 |
| **XGBoost_Train** | 0.93 | 0.91 | 0.91 | 0.93 | 0.09 | 0.09 | 0.07 | 0.92 | 0.92 | 0.84 |
| **XGBoost_Test** | 0.95 | 0.59 | 0.89 | 0.78 | 0.41 | 0.11 | 0.05 | 0.87 | 0.92 | 0.60 |

We have the highlighted models been the ones that has given a good result. As we have a data that are multicollinearity, we perform the tuning of the model, where the results are better on XGBoost when applied on the test model.

**Cross Validation Results**:

| Measures/Model | Std Dev | Variance | Final Mean |
|---|---|---|---|
| SMOTE Naïve Bayse_Train | 1.039 | 0.0001079 | 66.196 |
| SMOTE Naïve Bayse_Test | 2.687 | 0.0007221 | 80.356 |
| SMOTE KNN Train | 1.156 | 0.0001336 | 84.378 |
| SMOTE KNN Test | 1.193 | 0.0001423 | 83.582 |
| SMOTE SVM Train | 1.363 | 0.0001857 | 72.049 |
| SMOTE SVM Test | 1.604 | 0.0002573 | 83.583 |
| SMOTE LDA Train | 1.635 | 0.0002672 | 82.48 |
| SMOTE LDA Test | 1.847 | 0.0003413 | 84.824 |
| XGBoost Train | 3.388 | 0.001148 | 90.271 |
| XGBoost Test | 1.544 | 0.0002384 | 89.852 |

Hence, we conclude by considering the model as XG boost prediction to NBFC.

## 6. Final interpretation / recommendation

NBFC can take certain measures like below to stop the foreclosure if a customer is been on Notice of Default.

- **Foreclosure Workout**: This would work out effectively as the lender would come up for a compromise. The lender will definitely try out with mortgage than going for foreclosure.
- **Short Sale**: If the lender is been notified as defaulter the bank can issue a short sale and can schedule an auction of the property.
- **Bankruptcy**: If a bank declares the lender as bankrupted, this can stop the process of foreclosure.
- **Deed in Lieu**: The lender can voluntarily surrender the property to the NBFC. NBFC can take further action on the property and can close the due amount of the lender.

On part of customer, there are several reasons where they fall in as a defaulter and further move on to foreclosure.

- Unemployment

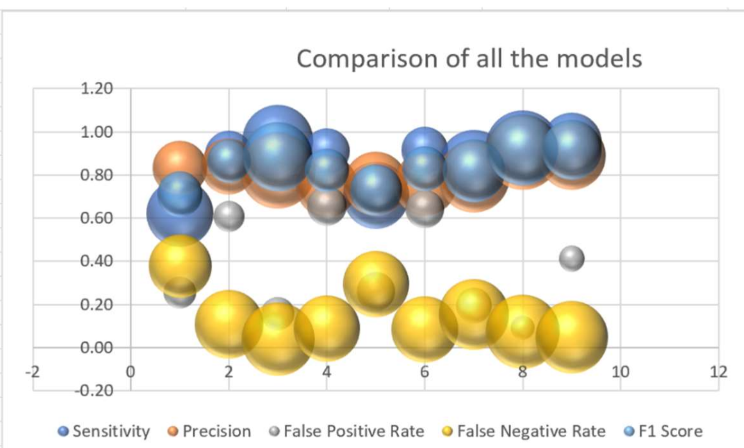
- Emergency Illness
- Death of the lender & no secondary income
- Excessive debt obligation
- Salary of the lender

If the lender has been insured with home loan with the mentioned above reasons, and this can reduce the lender falling to foreclosure procedures. Banks should promote the customers and educate them the benefits availed under the insurance taken. The premium on part of the residential property should be availed through tax redemption.

**Recommendation:**

- NBFC should implement special schemes to borrowers in Mumbai, Hyderabad, Pune and Chennai as we have the highest foreclosure cities
- As RBI has banned NBFC in charging foreclosure penalties from borrowers, we should identify our defaulters' group by revising their documents collected and should forecast the borrowers CBIL scores earlier.
- Trend of foreclosure is only for Home Loans and Short term home loan.
- Net Loan to Value ratio is to be maintained strictly at 50% for HL products and 75%for LAP.
- As NBFC plays a vital role in economic growth of the nation. In the coming period NBFC has to raise the financial support to customers nearly by 40% not by involving any banks.
- NBFC has to change their standard loan dispensing norms and offer highly innovative and tailored credit offering to borrowers.
- NBFC has to concentrate on loans taken from rural places where the volume of the loans taken and the foreclosure, the facts are shocking that rural is ratio of foreclosure is high if compared to urban places.

**Interpretation**




While comparing with all the modelling techniques, Tuning methods and Cross Validation methods we come with a conclusion that XG Boosting is the best model that gives the prediction as 93% of the test data have correctly predicted the defaulters' rate and 91% of the data has correctly identified the non-defaulters, where their accuracy of foreclosure is 92%. As per our analysis we have predicted 5825 customers can go for foreclosure.