**Christopher Dennies**

**Batch 1 - April 2020**

BuSiness report

NBFC Default Prediction Final notes

Contents

[**1.** **Introduction of the Business Problem** 3](#_Toc72100987)

[**1.1 Problem Statement:** 3](#_Toc72100988)

[**1.2 Business Implications of the Study:** 3](#_Toc72100989)

[**1.3 Objective:** 3](#_Toc72100990)

[**2.** **EDA – Univariate / Bi-Variate / Multi-Variate** 4](#_Toc72100991)

[**2.1 Visual Presentation of missing values** 5](#_Toc72100992)

[**2.2 Univariate – Default rate** 6](#_Toc72100993)

[**2.3 Univariate – NET\_LTV** 7](#_Toc72100994)

[**2.4 Product / Loan amount / Default** 8](#_Toc72100995)

[**2.5 City / Defaults** 9](#_Toc72100996)

[**2.6 Trend of Defaults** 10](#_Toc72100997)

[**2.7 Bivariate/Multivariate Analysis** 11](#_Toc72100998)

[2.7.1 Balance Tenure vs Foreclosure 12](#_Toc72100999)

[2.7.2 Completed Tenure vs Foreclosure 12](#_Toc72101000)

[2.7.3 Excess Available vs Foreclosure 13](#_Toc72101001)

[2.7.4 FOIR vs Foreclosure 13](#_Toc72101002)

[2.7.5 Net-Receivable vs Foreclosure 14](#_Toc72101003)

[2.7.6 Outstanding Principal vs Foreclosure 14](#_Toc72101004)

[2.7.7 Paid Principal vs Foreclosure 15](#_Toc72101005)

[2.7.8 Paid Principal vs Foreclosure 15](#_Toc72101006)

[2.7.9 Pre EMI-Due Amount vs Foreclosure 16](#_Toc72101007)

[**3.** **Data Cleaning & Preprocessing** 17](#_Toc72101008)

[**3.1 Dropping variables** 17](#_Toc72101009)

[**3.2 Correlation Plot & Dropping variables** 18](#_Toc72101010)

[**3.3 Applying VIF & Dropping variables** 22](#_Toc72101011)

[**3.4 Outlier Treatment / Univariate Analysis** 22](#_Toc72101012)

[**3.5 Derived Metrics & Insights** 22](#_Toc72101013)

[**4.** **Model Building & Model Validation** 23](#_Toc72101014)

[**4.1 Models Comparison** 24](#_Toc72101015)

[**5.** **Final Interpretation** 25](#_Toc72101016)

[**6.** **Final Recommendation** 26](#_Toc72101017)

[**7.** **Appendix** 27](#_Toc72101018)

[**7.1 Data Dictionary:** 27](#_Toc72101019)

[**7.2 Descriptive Statistics – Significant Variables:** 29](#_Toc72101020)

[**7.3 Outlier Treatment / Univariate Analysis** 29](#_Toc72101021)

[**7.4 Derived Metrics & Insights** 34](#_Toc72101022)

[**7.5 Logistic Regression Output – Model 4** 35](#_Toc72101023)

[7.5.1. LOGISTIC REGRESSION - WITH DEFAULT CUTOFF 0.5 37](#_Toc72101024)

[7.5.2. LOGISTIC REGRESSION – TRAIN DATA - WITH OPTIMUM CUTOFF 0.09 38](#_Toc72101025)

[7.5.3. LOGISTIC REGRESSION – TEST DATA - WITH OPTIMUM CUTOFF 0.09 39](#_Toc72101026)

[7.5.4. LOGISTIC REGRESSION – SMOTE DATA – TRAIN DATASET – CUTOFF – 0.09 40](#_Toc72101027)

[7.5.5. LOGISTIC REGRESSION – SMOTE DATA – TEST DATASET – CUTOFF – 0.09 41](#_Toc72101028)

[**7.6 LDA - LINEAR DISCRMINANT ANALYSIS** 42](#_Toc72101029)

[7.6.1. LDA – TRAIN DATASET – CUTOFF – 0.06 43](#_Toc72101030)

[7.6.2. LDA – TEST DATASET – CUTOFF – 0.06 44](#_Toc72101031)

[7.6.3. LDA – SMOTE DATASET – CUTOFF – 0.06 45](#_Toc72101032)

[**7.7. RANDOM FOREST MODEL** 46](#_Toc72101033)

[7.7.1. RANDOM FOREST – TRAIN DATASET 46](#_Toc72101034)

[7.7.2. RANDOM FOREST – TEST DATASET 47](#_Toc72101035)

[7.7.3. RANDOM FOREST – SMOTE DATASET 48](#_Toc72101036)

# **Introduction of the Business Problem**

## **1.1 Problem Statement:**

A Non-Banking Financial Company (NBFC) is a company 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 used as the collateral for the loan.

* Ultimate problem – High Foreclosure Costs, Legal Hassle & Loss of Customers
* The Penultimate problem is the losing customers defaults which is resulting to the foreclosure process which is costing the NBFC and thereby.
* Defaults is a Core problem.

## **1.2 Business Implications of the Study:**

* Highlighting the driving factors leading to 'FORECLOSURE' of the loan will help the NBFC to take prior actions while sanctioning the loan and during payment tenure.
* Utilization of funds are more directed to the right customers.
* Profitability of the NBFC is increased and there by keeping a tab on Non-Performing assets (NPA).

## **1.3 Objective:**

Achieve a Best Model to predict the probability of default of the existing loan accounts. Recommend important variables for NBFC to take prior action to avoid foreclosures, save costs from legal hassle and thereby avoid losing customers.

# **EDA – Univariate / Bi-Variate / Multi-Variate**

Data consists of aggregated loan transactions data of the customers and below are the observations.

* There are 20012 rows and 53 columns.
* There are no duplicated rows.
* Float – 32 Variables
* Integer – 14 Variables
* Date Time – 3 variables
* Object – 4 variables
* Methodology of collected data – Aggregated loan transaction data.
* Time - (August 2010 – December 2018) – 8 Years 4 months loan data
* Frequency – The loan data narrowed down to daily date wise.
* Renaming not required for this dataset.

## **2.1 Visual Presentation of missing values**

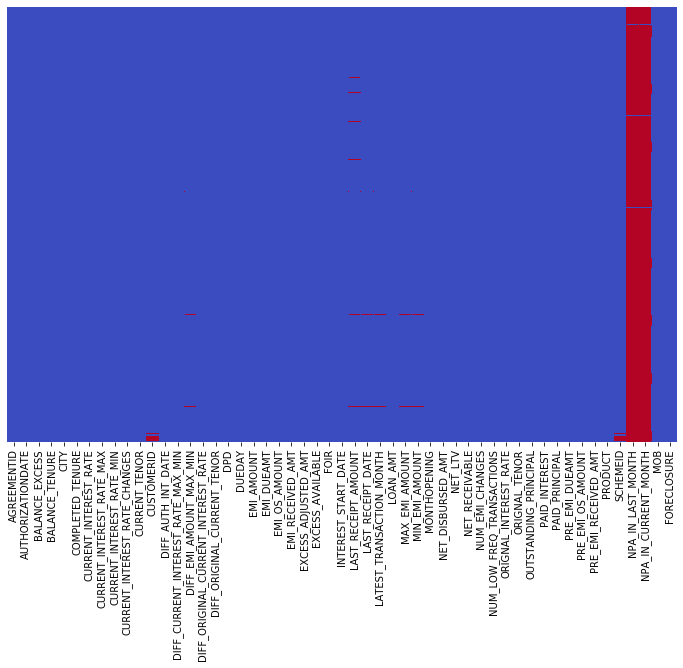


Figure 1: Visual presentation of missing variables

* There are missing values in the dataset. Below data is expressed in Percentage Missing values. Both NPA in last month and current month has 99.41% missing values. Rest all variables are negligible. I.e., < 2%

**CUSTOMERID 1.4000**

**DIFF\_EMI\_AMOUNT\_MAX\_MIN 0.4400**

**LAST\_RECEIPT\_AMOUNT 1.2300**

**LAST\_RECEIPT\_DATE 0.3700**

**LATEST\_TRANSACTION\_MONTH 0.3700**

**MAX\_EMI\_AMOUNT 0.4400**

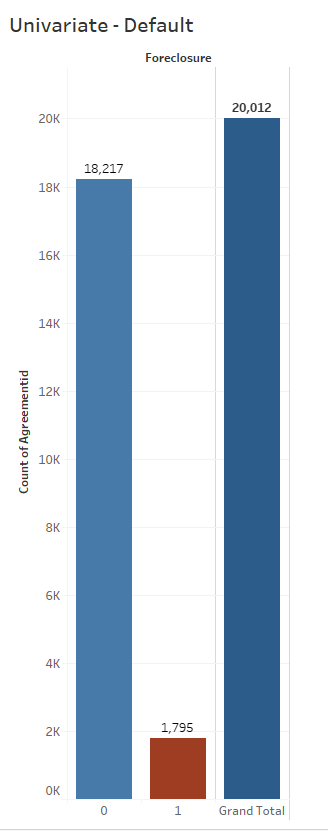
**MIN\_EMI\_AMOUNT 0.4400**

**SCHEMEID 1.4000**

**NPA\_IN\_LAST\_MONTH 99.4100**

**NPA\_IN\_CURRENT\_MONTH 99.4100**

## **2.2 Univariate – Default rate**



* Default rate @ 8.9%
* Data is imbalanced.
* From Histogram Plot, the most of the NET\_LTV lies between 30 – 75 %.

Figure 2: Default Rate

## **2.3 Univariate – NET\_LTV**

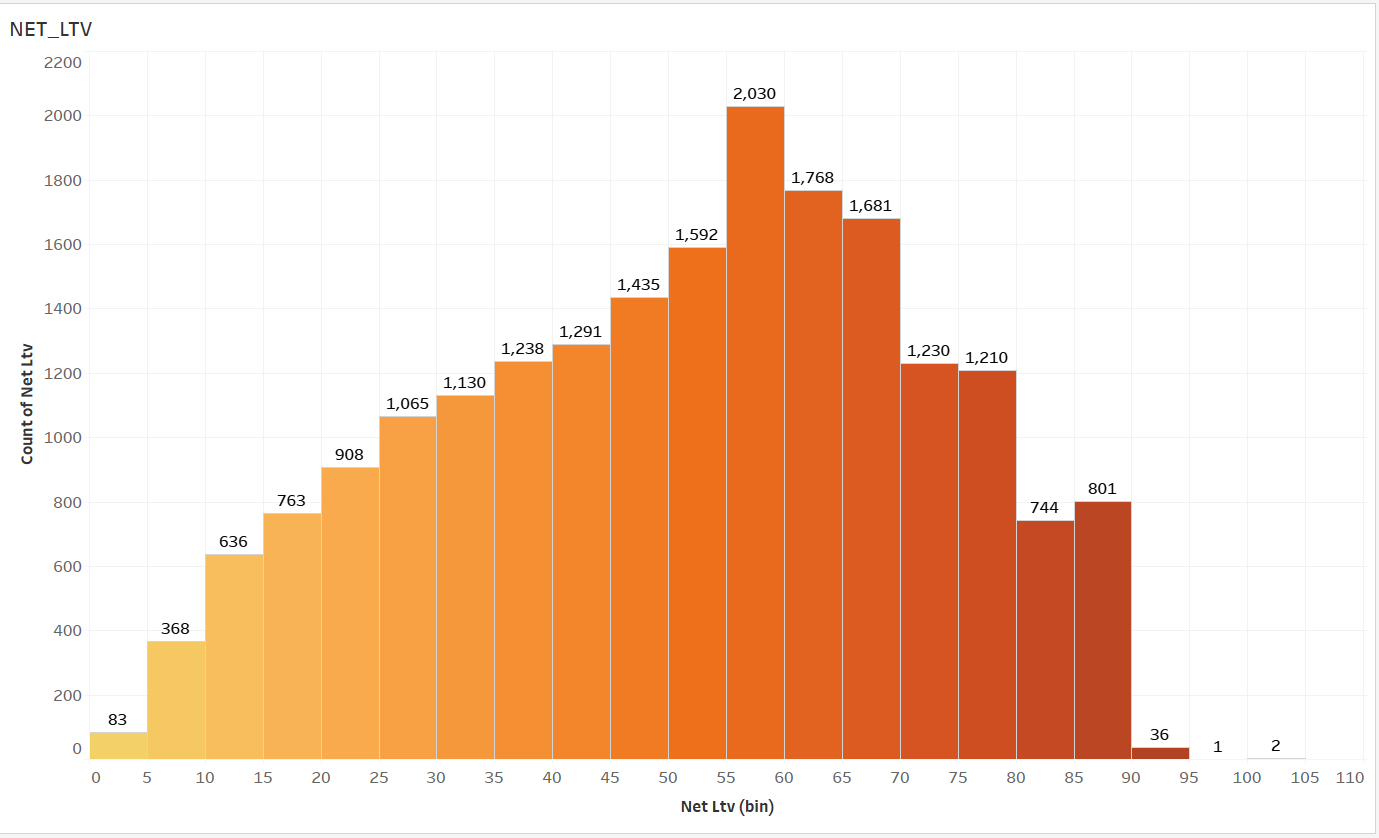


Figure 3: Net – Loan to Value (LTV)

* Highest selling product is STHL -Small Ticket Home loans, Followed by LAP – loan against property, HL – Home loan & STLAP – Small ticket loan against property.
* HL – Highest defaults.

## **2.4 Product / Loan amount / Default**

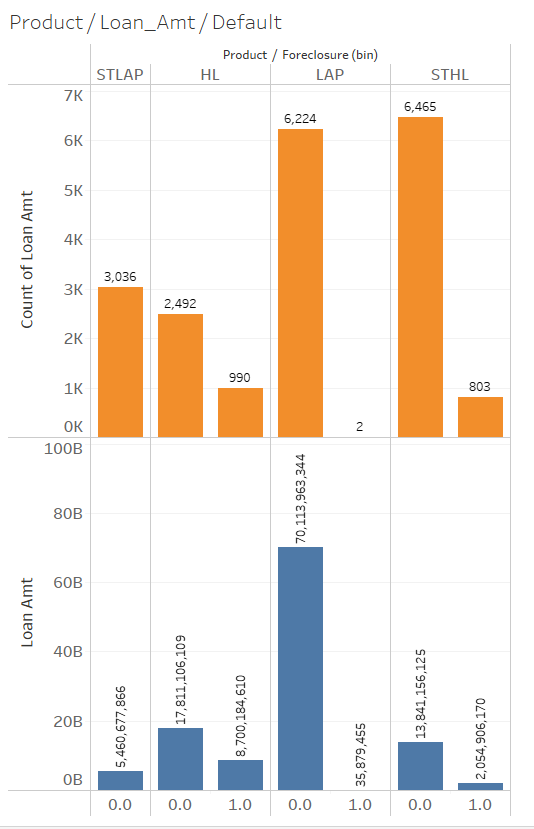


Figure 4. Product / Loan Amount vs. Default

* LAP – Highest loan amounts disbursed.

## **2.5 City / Defaults**

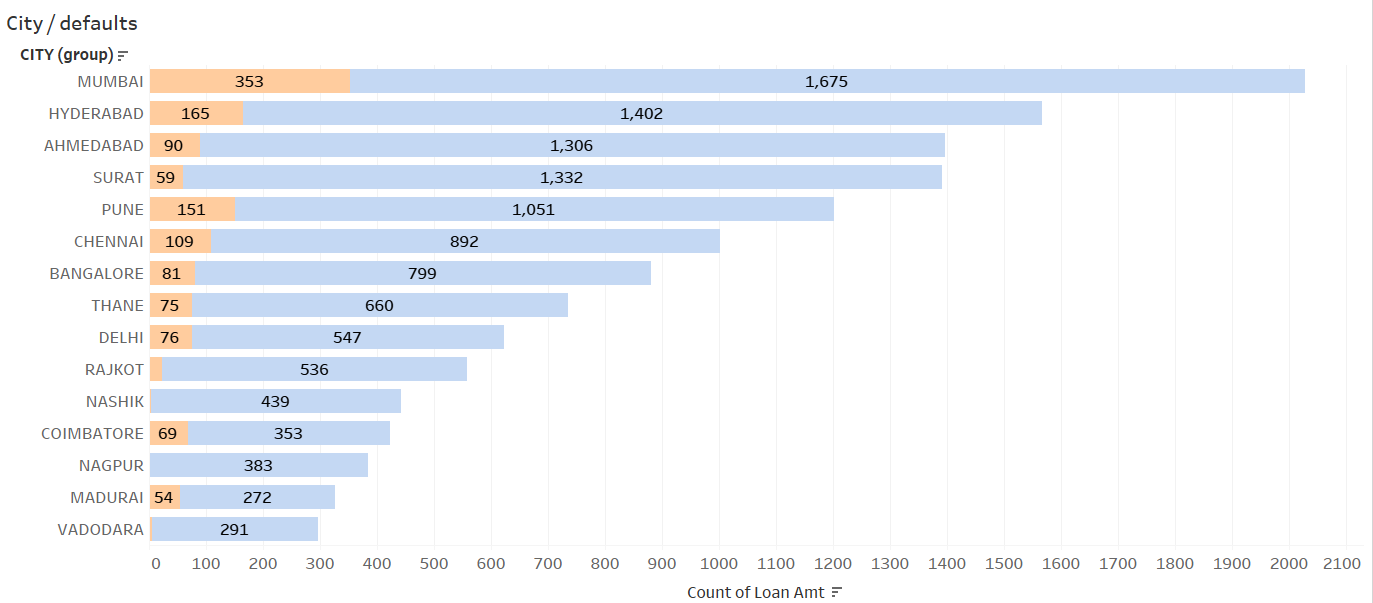
****

Figure 5. City vs. Default

* Mumbai – Highest loan disbursements and Highest defaults @ 353.
* Top 15 cities and their Defaults Vs Non- Defaults.

## **2.6 Trend of Defaults**

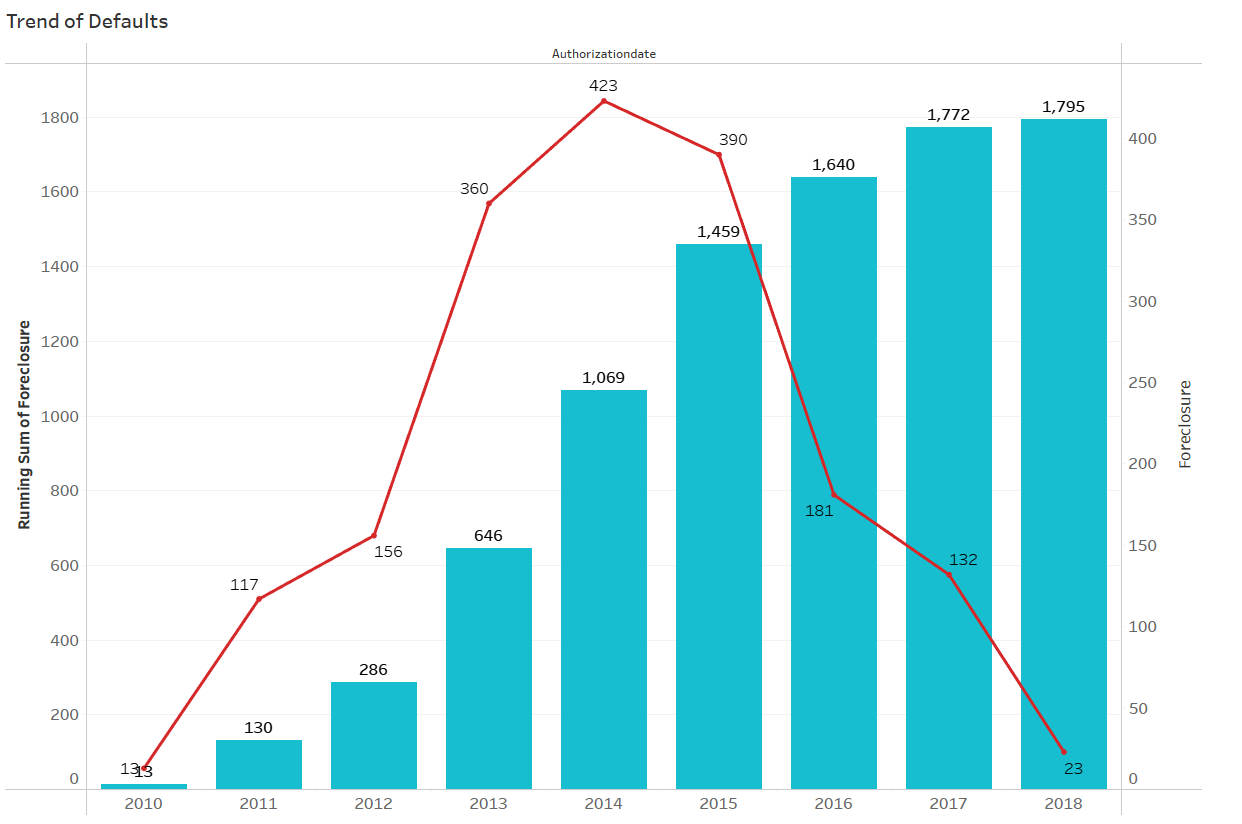
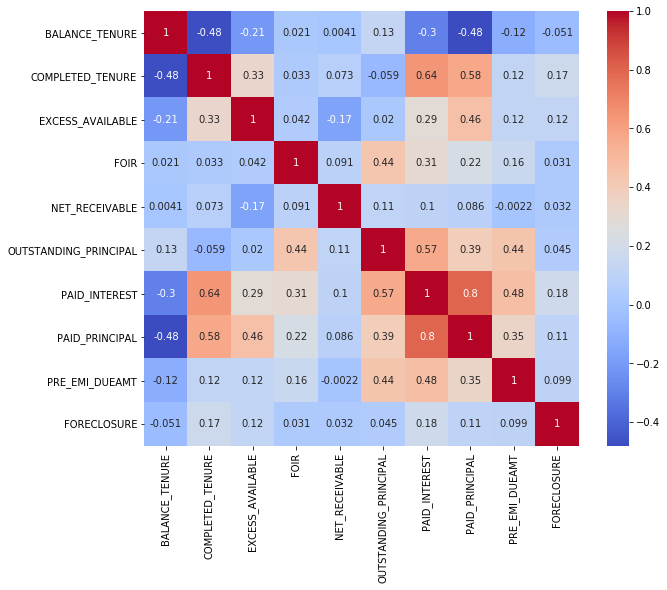


Figure 6. Default Trend

* The Trend of Foreclosures are trending up from 2010 to 2014 & trending down from 2014 to 2018.
* Year 2014 recorded highest number of Foreclosures.

## **2.7** [**Bivariate/Multivariate Analysis**](https://www.kaggle.com/ab9bhatia/complete-eda-for-loan-analysis#bivariate)

* Below is the pair plot of the significant variables which clearly shows that there is no clear relationship between each other, i.e.. There is no Multicollinearity.

****

**Figure 7: Pair Plot of Significant variables**

2.7.1 Balance Tenure vs Foreclosure

* Median of foreclosure is less than non-foreclosure median with less margin. Balance Tenure, Unlikely to be a strong predictor.

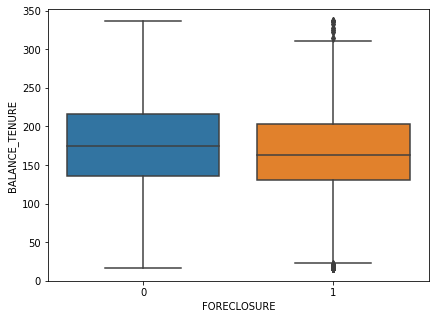
****

Figure 8 – Balance Tenure vs. Foreclosure

2.7.2 Completed Tenure vs Foreclosure

* Foreclosure and Non-Foreclosure population distribution is different and distinct, completed tenure likely to be a Strong Predictor.

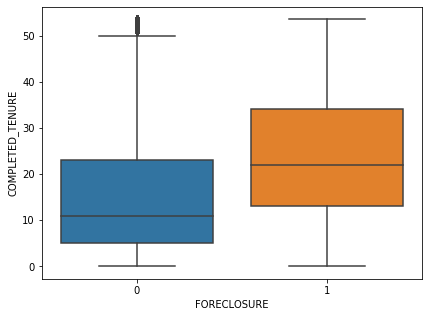
****

Figure 9 – Completed Tenure vs. Foreclosure

2.7.3 Excess Available vs Foreclosure

* Distributions are not similar, excess available highly likely to be strong predictor.

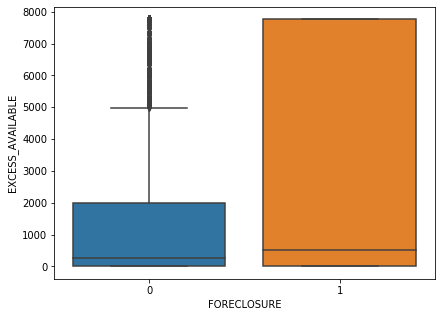
****

Figure 10 – Excess Available vs Foreclosure

2.7.4 FOIR vs Foreclosure

* Distributions are similar like to be a weak predictor.

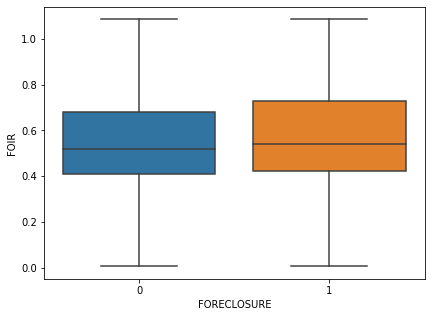
****

Figure 11 – FOIR vs. Foreclosure

2.7.5 Net-Receivable vs Foreclosure

* Distribution between foreclosure and non-foreclosure distributions are not similar, likely to be strong predictor.

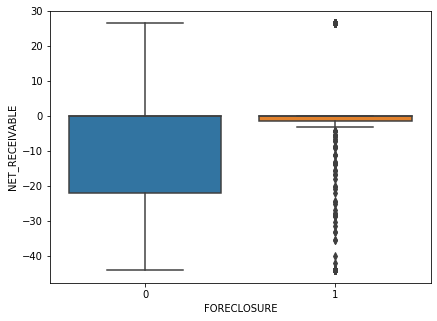
****

Figure 12 – Net-Receivable vs Foreclosure

2.7.6 Outstanding Principal vs Foreclosure

* Higher outstanding principle are likely head to Foreclosure, likely to be a strong predictor.

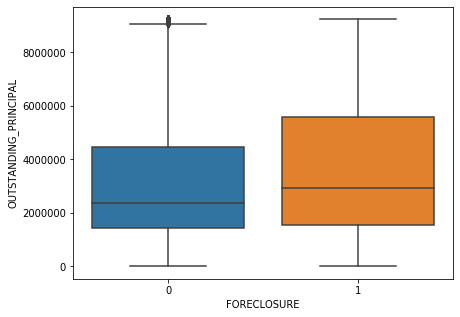
****

Figure 13 – Outstanding Principal vs Foreclosure

2.7.7 Paid Principal vs Foreclosure

* Customer paying more interest are likely to Foreclosure, could be an important variable in the final model.

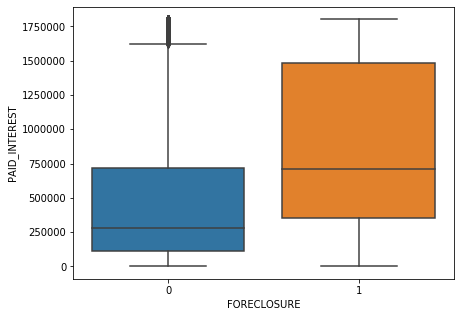
****

Figure 14 – Paid Interest vs Foreclosure

2.7.8 Paid Principal vs Foreclosure

* Paid Principal variable is contrary to business understanding, as per the distributions.

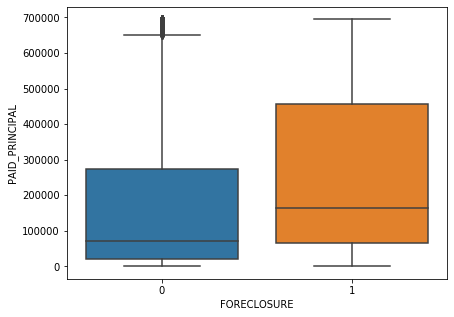
****

Figure 15 – Paid Interest vs Foreclosure

2.7.9 Pre EMI-Due Amount vs Foreclosure

* Distribution is quite distinctive in nature; Pre Emi-Due amount likely to be a strong predictor.

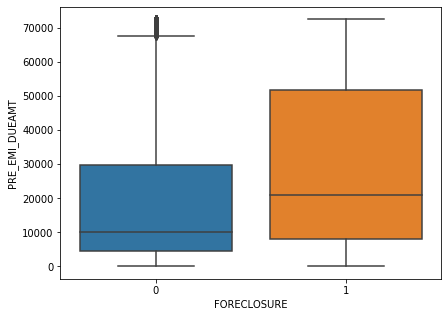
****

Figure 16 – Pre EMI-Due Amount vs Foreclosure

# **Data Cleaning & Preprocessing**

## **3.1 Dropping variables**

The variables are dropped based on data understanding

* **Agreement Id** variable holds the distinct count of Foreclosure accounts. It is dropped.
* **Customer Id** has few missing values, and the data is unique at an agreement id level which will not help in foreclosure prediction, which is dropped.
* **Scheme Id** has few missing values, and the data has no extra information, which will not help in default prediction, which is dropped.
* **MOB** is an internal code, and the data has no extra information, which will not help in default prediction, which is dropped.
* The dataset has variables explaining the same subject which requires to be dropped by domain understanding.
* Dropping variables as per high number of missing values and treatment for the missing variables.
* **NPA\_IN\_LAST\_MONTH** variable has 99.41 missing values and only 2 Foreclosures of 15 NPA's, which is not a good predictor thereby this variable is dropped.

Refer Below table 1:

| **FORECLOSURE** | **0** | **1** | **All** |
| --- | --- | --- | --- |
| **NPA\_IN\_LAST\_MONTH** |  |  |  |
| **0** | 69 | 33 | 102 |
| **#N/A** | 2 | 0 | 2 |
| **Yes** | 13 | 2 | 15 |
| **All** | 84 | 35 | 119 |

* **NPA\_IN\_CURRENT\_MONTH** variable has 99.41 missing values and only 2 Foreclosures of 16 NPA's, which is not a good predictor will drop this variable.

Refer below table 2:

| **FORECLOSURE** | **0** | **1** | **All** |
| --- | --- | --- | --- |
| **NPA\_IN\_CURRENT\_MONTH** |  |  |  |
| **0** | 70 | 33 | 103 |
| **Yes** | 14 | 2 | 16 |
| **All** | 84 | 35 | 119 |

* **Min & Max & Min Max Difference Emi Amount, Latest transaction month, Last received amount** Variables imputed with median as these have extreme values.
* **Last receipt date** Variable imputed with mode as it has high frequency.

## **3.2 Correlation Plot & Dropping variables**

* From the below correlation graph Figure 17, Original Tenor, Balance Tenor and Current Tenor are highly correlated, Balance Tenor will be retained along with Completed Tenor, with domain understanding. Dropping Original Tenor, Current Tenor & difference between original and current tenor.

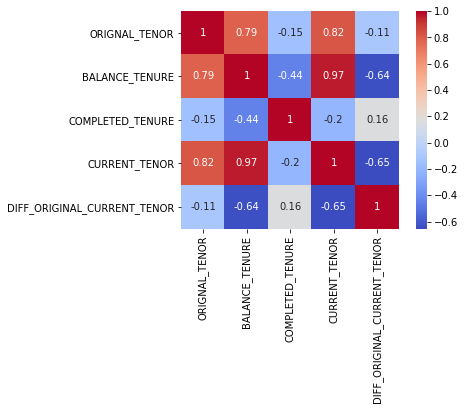


Figure 17 – Correlation Plot

* From Below **Figure 18:** Current Interest rate is highly correlated with other version of available interest rates (Max, Min & Original), Current Interest rate will be retained, others dropped.
* Difference between Current max and Current min, Difference between Original and Current Interest Rate & Current interest rate changes dropped as no insights derived from it.

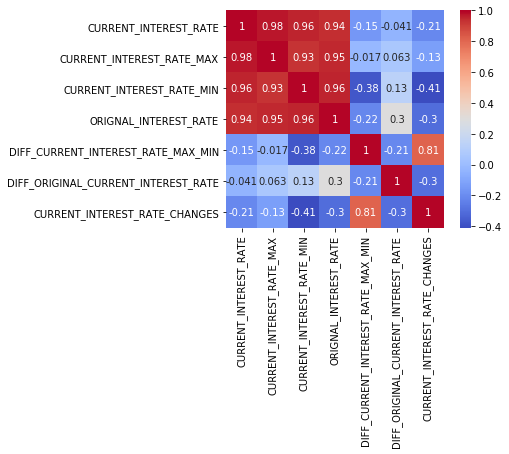
****

Figure 18 – Correlation Plot

* From **Figure 19:** EMI Amount and Outstanding EMI amount and Received amount are more intuitive to use when compared to other variation of EMI variables. Rest other variables dropped.

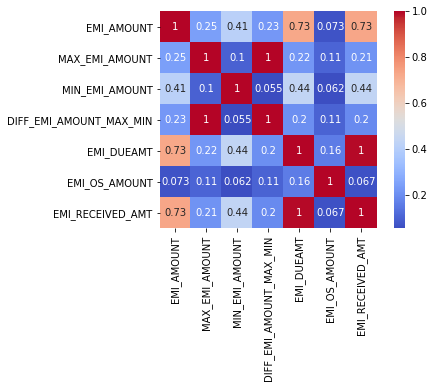
****

Figure 19 – Correlation Plot

* From **Figure 20:** Pre-EMI Due amount & Pre EMI-Received Amount are perfectly highly correlated, in the context of foreclosure the pre emi due amount will be retained along with 'Pre Emi OS amount'.

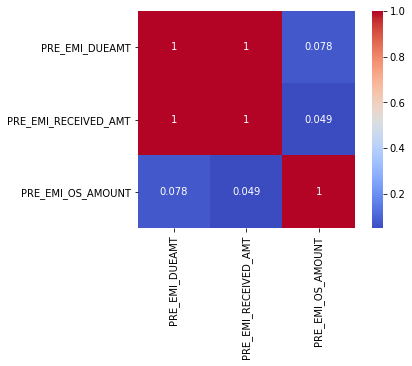
****

Figure 20 – Correlation Plot

* From **Figure 21:** Excess Available and Excess Adjusted Amount are highly correlated, 'Excess Available' will be retained along with 'Balance Excess'.

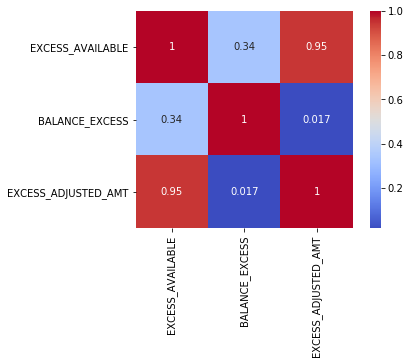
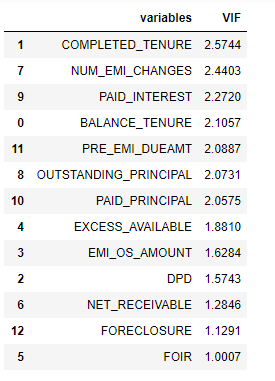
****

Figure 21 – Correlation Plot

## **3.3 Applying VIF & Dropping variables**

* Variance inflation factor applied to 26 variables with a cut off below 5, dropped to 12 significant variables excluding the target variable Foreclosure.

**Table 3:**



* Refer **Table 4** in Appendix, descriptive statistics of the significant variables, to increase the discriminatory power DPD, EMI OS amt & Number of Emi Changes will be binned, and rest continuous variables will do outlier treatment.

## **3.4 Outlier Treatment / Univariate Analysis**

* Outlier treatment applied to 9 variables.
* Refer Figure 22,23,24,25,26,27,28,29,30 in Appendix.

## **3.5** [**Derived Metrics**](https://www.kaggle.com/ab9bhatia/complete-eda-for-loan-analysis#derived) **& Insights**

* Refer Table 5,6 & 7 in Appendix.
* Included City & Product categorical variables – Converted to integers using cat codes.

Note: Finally, 16 variables are selected to run the appropriate models.

# **Model Building & Model Validation**

Supervised Learning Approach is used as the labels are provided.

* As it’s a Binary Classification problem, under Classification – linear models such as Logistic regression & Linear Discriminant analysis is carried out.
* Under Non-Linear Classification Models Random Forest is Chosen.
* As the data is imbalanced, Smote was used to improve the metrics.
* The overall accuracy will give a wrong picture. Rather Recall and Precision of that class needs attention and tuned to get a best Model.
* 16 Significant variables were applied to logistic regression, 3 variables were dropped as “P” value being greater than significance value.
* One Variable NET\_LTV, though the P value is greater, retained as per domain understanding.
* The List is arrived Basis Domain Knowledge, Correlation Plots, Variation inflation factor and finally on the P-values.
* Stas Model Library was used to build a logistic model.
* Logit Regression Results – Refer - **Table 8** in Appendix.

## **4.1 Models Comparison**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **Dataset** | **Precision** | **Recall** | **F1-Score** | **Accuracy** | **AUC** |
| [**Logistic Regression with Default Cut-Off**](#_7.5_Logistic_Regression) | **Train** | **0.46** | **0.05** | **0.09** | **0.91** | **0.52** |
| [**Logistic Regression with Optimal Cut-Off**](#_7.5_Logistic_Regression) | **Train** | **0.19** | **0.69** | **0.30** | **0.72** | **0.70** |
| [**Logistic Regression with Optimal Cut-Off**](#_7.5_Logistic_Regression) | **Test** | **0.19** | **0.66** | **0.30** | **0.72** | **0.69** |
| [**Logistic Regression on SMOTE Train data**](#_7.5_Logistic_Regression) | **SMOTE Train** | **0.66** | **0.74** | **0.70** | **0.72** | **0.81** |
| [**Logistic Regression on SMOTE Test data**](#_7.5_Logistic_Regression) | **SMOTE Test** | **0.66** | **0.73** | **0.69** | **0.72** | **0.81** |
| [**Linear Discriminant Analysis - LDA**](#_7.6_LDA_-) | **Train** | **0.39** | **0.12** | **0.18** | **0.90** | **0.79** |
| [**Linear Discriminant Analysis - LDA**](#_7.6_LDA_-) | **Test** | **0.37** | **0.10** | **0.16** | **0.90** | **0.77** |
| [**Linear Discriminant Analysis with Optimal Cut-Off**](#_7.6_LDA_-) | **Train** | **0.17** | **0.79** | **0.28** | **0.64** | **0.71** |
| [**Linear Discriminant Analysis with Optimal Cut-Off**](#_7.6_LDA_-) | **Test** | **0.17** | **0.75** | **0.27** | **0.64** | **0.69** |
| [**Linear Discriminant Analysis - LDA on SMOTE**](#_7.6_LDA_-) | **SMOTE Train** | **0.63** | **0.84** | **0.72** | **0.72** | **0.74** |
| [**Random Forest Model on Train**](#_7.7._RANDOM_FOREST) | **Train** | **0.84** | **0.38** | **0.52** | **0.94** | **0.69** |
| [**Random Forest Model on Test**](#_7.7._RANDOM_FOREST) | **Test** | **0.77** | **0.31** | **0.44** | **0.93** | **0.65** |
| **[Random Forest Model on SMOTE](#_7.7._RANDOM_FOREST)** | **SMOTE Train** | **0.94** | **0.92** | **0.93** | **0.94** | **0.94** |

**Table 8.1: Comparison of Models**

* Random forest is a high-performance model, but it is a black box model lacking insight. Though the variable importance is achieved it lacks magnitude unlike logistic regression.
* SMOTE was used to balance the data and thereby it helped to fine tune the model. By fine Tuning, Random forest model achieved the maximum accuracy compared to all the models.
* Logistic Regression Model is preferred over other models, as it give enormous information on the variables which could easily be interpreted and understood by probabilities.
* Random Forest only helps in identifying the order of Importance among the variables but not the magnitude. Refer Table 8.2.

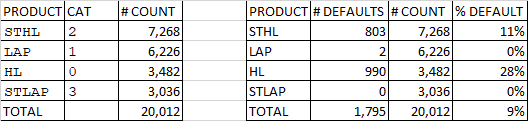
# **Final Interpretation**

|  |  |  |  |
| --- | --- | --- | --- |
| **VARIABLES** | **COEFFICENT** | **Exp (Coeff) - ODD ratio** | **Percent** |
| **Intercept** | **-0.285200000000** | **0.751863867** | **-24.81%** |
| **NUM\_EMI\_CHANGES\_RANGE\_CAT** | **0.130300000000** | **1.139170083** | **13.92%** |
| **NET\_RECEIVABLE** | **0.003000000000** | **1.003004505** | **0.30%** |
| **NET\_LTV** | **0.002300000000** | **1.002302647** | **0.23%** |
| **EXCESS\_AVAILABLE** | **0.000060840000** | **1.000060842** | **0.01%** |
| **PRE\_EMI\_DUEAMT** | **0.000011210000** | **1.00001121** | **0.00%** |
| **PAID\_INTEREST** | **0.000001540000** | **1.00000154** | **0.00%** |
| **LOAN\_AMT** | **-0.000000025480** | **0.999999975** | **0.00%** |
| **OUTSTANDING\_PRINCIPAL** | **-0.000000116700** | **0.999999883** | **0.00%** |
| **PAID\_PRINCIPAL** | **-0.000002954000** | **0.999997046** | **0.00%** |
| **BALANCE\_TENURE** | **-0.003900000000** | **0.996107595** | **-0.39%** |
| **CITY\_NEW** | **-0.017700000000** | **0.982455725** | **-1.75%** |
| **FOIR** | **-0.868400000000** | **0.419622408** | **-58.04%** |
| **PRODUCT** | **-0.982800000000** | **0.374261698** | **-62.57%** |

**Table 8.3**

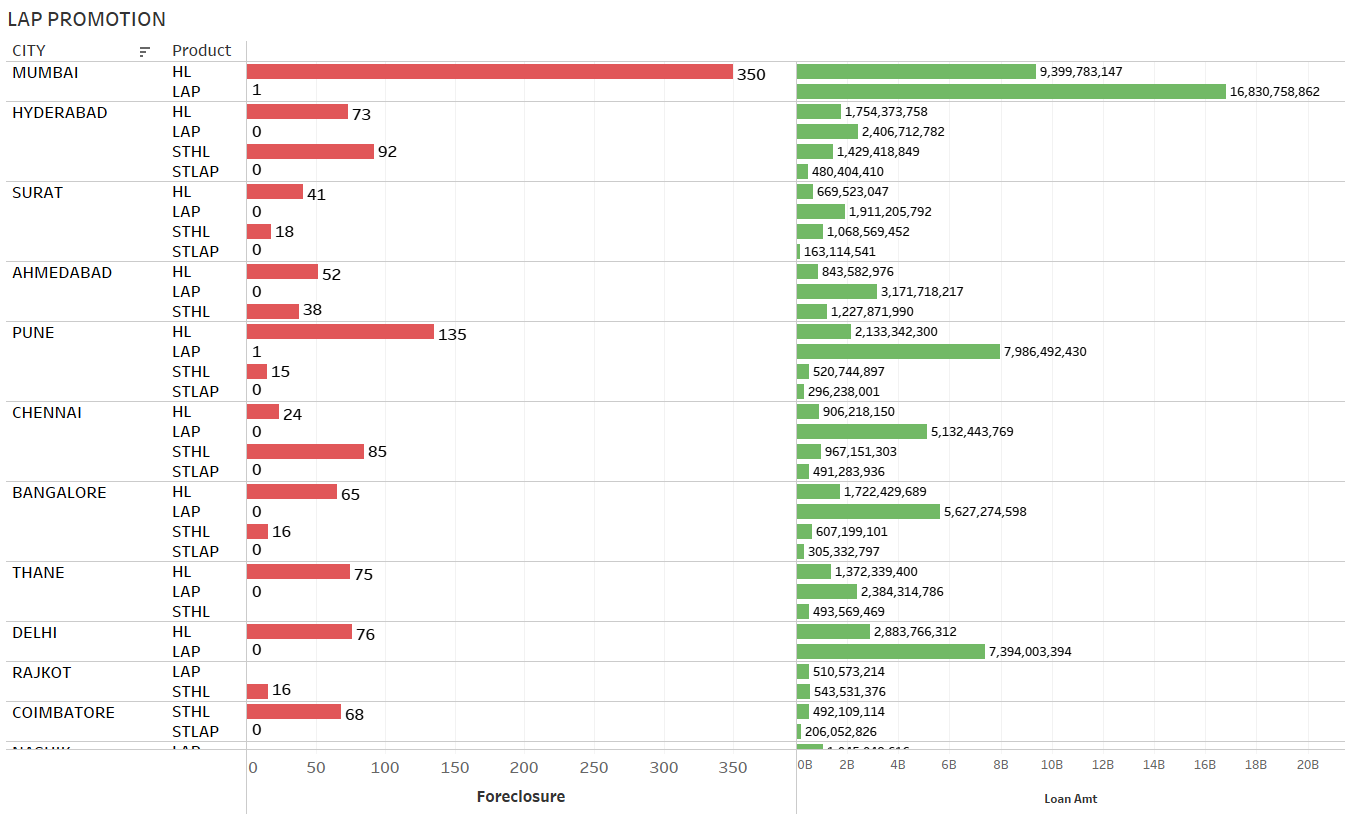
* For a unit change in No. EMI Changes, the odds ratio for an account to default is 1.13.
* For a unit change in No EMI Changes, there is 13 percent increase in odds for an account to default.

**Table 8.4**



* For a unit change in product i.e., HL to LAP, LAP to STHL, STHL to STLAP there is a 62 percent decrease in odd for an account to default.

# [**Final Recommendatio**](https://www.kaggle.com/ab9bhatia/complete-eda-for-loan-analysis#derived)**n**

**Figure 41:**

* High Defaults are seen in Home loan category.
* When customer opts for an EMI change in Home Loan category there is 13% chance that an account might default.
* The NBFC should have a dedicated follow up on customers who opt for EMI change.
* High value ticket HL loans are seen in Mumbai, Hyderabad, Pune, The NBFC Should be more vigilant in these cities for EMI changes.
* Promote more LAP loans as there are less defaults observed.
* Promote more LAP loans in Hyderabad, Surat, Ahmedabad, Rajkot, Coimbatore.

# **Appendix**

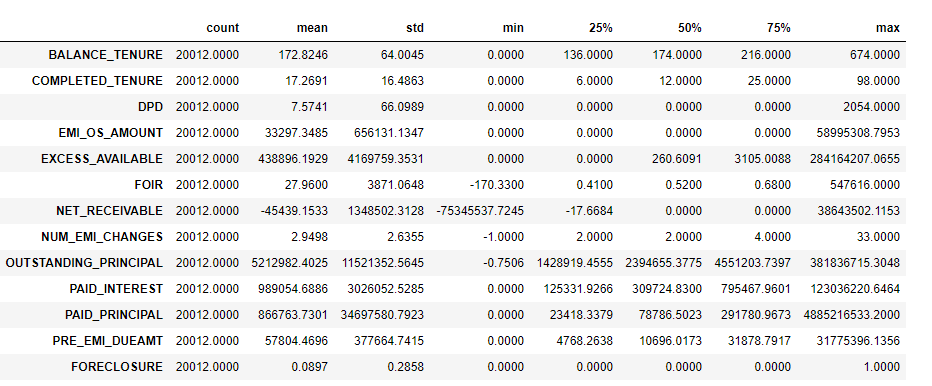
## **7.1 Data Dictionary:**

Variables are sorted as per understanding.

|  |  |
| --- | --- |
| **COLUMN NAME** | **DESCRIPTION** |
| AGREEMENTID | Agreement ID of the loan account ( a customer can have multiple loans) |
| CUSTOMERID | Unique Customer ID given to each customer |
| SCHEMEID | Scheme ID under which loan was given |
| MOB | Internal code |
| AUTHORIZATIONDATE | Authorization date of the loan |
| INTEREST\_START\_DATE | Interest start date on the loan |
| DIFF\_AUTH\_INT\_DATE | Difference between authorization and interest start date |
| DUEDAY | Next due date of the loan |
| ORIGNAL\_TENOR | Original tenor of the loan (when the loan was sanctioned) |
| CURRENT\_TENOR | Current tenor of the loan |
| DIFF\_ORIGINAL\_CURRENT\_TENOR | Difference in original and current tenor (ORIGNAL\_TENOR - CURRENT\_TENOR) |
| COMPLETED\_TENURE | Completed tenure |
| BALANCE\_TENURE | Remaining tenure |
| DPD | Days past due |
| ORIGNAL\_INTEREST\_RATE | Original rate of interest on the loan (when the loan was sanctioned). Renamed field (Old Name: ORIGNAL\_ROI) |
| CURRENT\_INTEREST\_RATE | Current rate of interest on the loan. Renamed field (Old Name: CURRENT\_ROI ) |
| DIFF\_ORIGINAL\_CURRENT\_INTEREST\_RATE | Difference in original ROI and current ROI (ORIGNAL\_ROI - CURRENT\_ROI) |
| CURRENT\_INTEREST\_RATE\_MAX | Maximum value of the CURRENT ROI across transactions |
| CURRENT\_INTEREST\_RATE\_MIN | Minimum value of the CURRENT ROI across transactions |
| DIFF\_CURRENT\_INTEREST\_RATE\_MAX\_MIN | Difference between the maximum and minimum interest rate per agreement |
| CURRENT\_INTEREST\_RATE\_CHANGES | Number of times the CURRENT ROI has changed |
| LOAN\_AMT | Loan amount which was sanctioned |
| NET\_DISBURSED\_AMT | Amount that was disbursed |
| OUTSTANDING\_PRINCIPAL | Outstanding principal |
| PAID\_INTEREST | Paid interst |
| PAID\_PRINCIPAL | Paid principal |
| PRE\_EMI\_DUEAMT | Pre EMI due amount for the loan |
| PRE\_EMI\_RECEIVED\_AMT | Pre EMI that was received |
| PRE\_EMI\_OS\_AMOUNT | Pre EMI Outstanding amount |
| NUM\_EMI\_CHANGES | Number of different values in the receipts amount |
| NUM\_LOW\_FREQ\_TRANSACTIONS | Number of transactions done in less than 28 days |
| BALANCE\_EXCESS | Balance of excess amount |
| EMI\_AMOUNT | Mode of the receipt amount |
| MAX\_EMI\_AMOUNT | Maximum receipt amount |
| MIN\_EMI\_AMOUNT | Minimum receipt amount |
| DIFF\_EMI\_AMOUNT\_MAX\_MIN | Difference between maximum and minimum EMI AMOUNT |
| EMI\_DUEAMT | EMI due amount |
| EMI\_RECEIVED\_AMT | EMI received amount |
| EMI\_OS\_AMOUNT | EMI outstanding amount |
| EXCESS\_ADJUSTED\_AMT | Excess adjusted amount |
| EXCESS\_AVAILABLE | Excess received |
| NET\_RECEIVABLE | Net receivable (EMI\_DUEAMT - EMI\_RECEIVED\_AMT = EMI\_OS\_AMOUNT) + (EXCESS\_AVAILABLE - EXCESS\_ADJUSTED\_AMT = BALANCE\_EXCESS) = NET\_RECEIVABLE) |
| LATEST\_TRANSACTION\_MONTH | Month of last receipt date. In case account is Foreclosed, it will be month of Foreclosure |
| LAST\_RECEIPT\_DATE | Last receipt date |
| LAST\_RECEIPT\_AMOUNT | Last receipt amount |
| FOIR | Fixed obligation to income ratio (Value should range from 0-1 – Derived variable) |
| NET\_LTV | Net Loan to Value ratio (Value ranges from 0-100 (in %) – Derived variable) |
| MONTHOPENING | Month of opening |
| CITY | City of origination |
| PRODUCT | Loan product |
| NPA\_IN\_LAST\_MONTH | Whether NPA in last month |
| NPA\_IN\_CURRENT\_MONTH | Whether NPA in current month |
| FORECLOSURE | Labelled Field |

## **7.2 Descriptive Statistics – Significant Variables:**

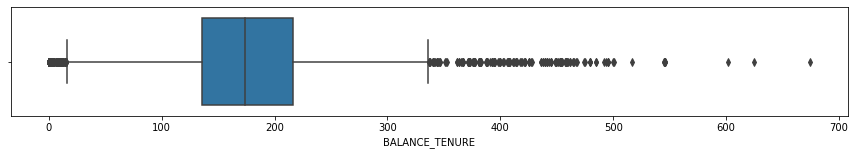
**Table 4:**

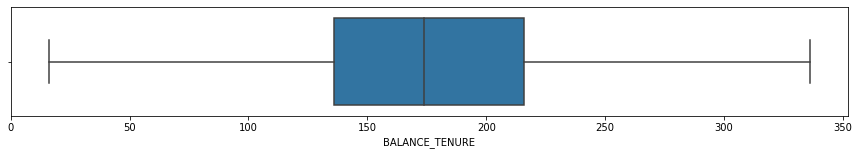


## **7.3 Outlier Treatment / Univariate Analysis**

* Balance tenure – Before outlier treatment, balance tenure had extreme outliers to 674 months. After treatment most of the values lie approximately between130 to 220 months.

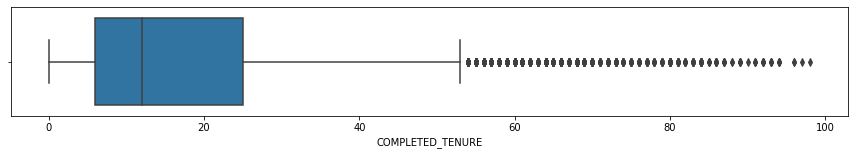
**Figure 22:**

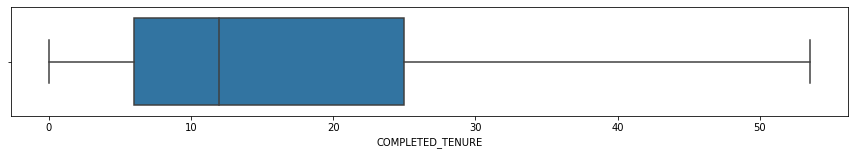




* Completed tenure – Before outlier treatment, completed tenure had extreme outliers to 98 months. After treatment most of the values lie approximately between 7 to 25 months.

**Figure 23:**

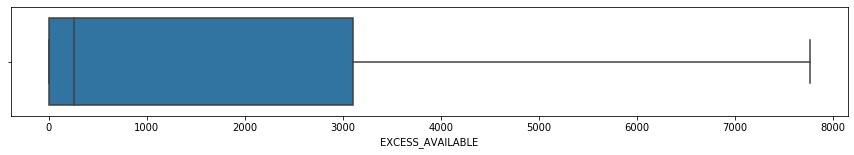




* Excess available – Before outlier treatment, Excess available had extreme outliers to 28 cr odd. After treatment most of the values lie approximately between 0 to 3k.

**Figure 24:**

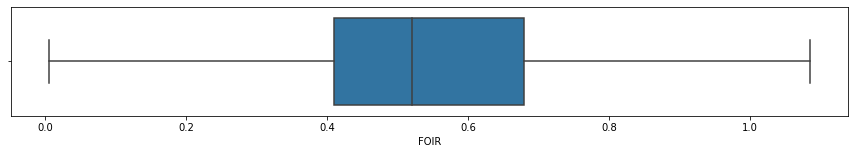




* FOIR – Before outlier treatment, FOIR available had negative value. After treatment most of the values lie approximately between 0.4 to 0.7 which is ideal range ( 0 – 1 ).

**Figure 25:**





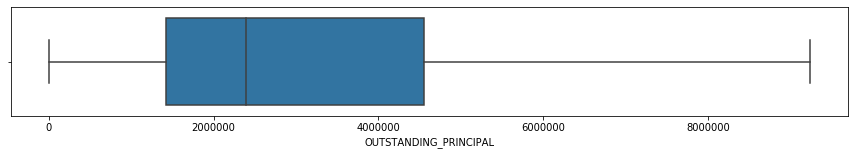
* Net receivable – Before outlier treatment, Net receivable had extreme outliers on both positive and negative ends. After treatment most of the values lie approximately between -18 to 0 lacs(mostly on the negative end ). Which is good predictor for foreclosure.

**Figure 26:**

* Outstanding principal – Before outlier treatment, outstanding principal had extreme outliers to 38 cr. After treatment most of the values lie approximately between 17 to 45 lacs.

**Figure 27:**

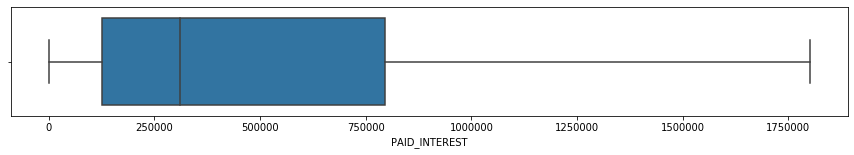




* Paid Interest – Before outlier treatment, paid interest had extreme outliers to 12.3 cr. After treatment most of the values lie approximately between 2 to 7.7 lacs.

**Figure 28:**

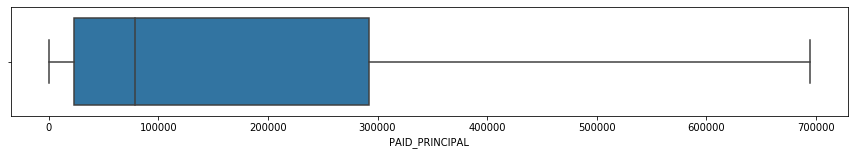




* Paid Principal – Before outlier treatment, Paid principal had extreme outliers to 488 cr. After treatment most of the values lie approximately between 40k to 2.9 lacs.

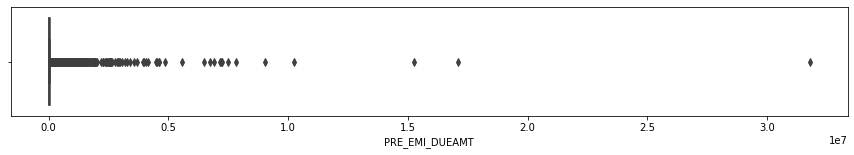
**Figure 29:**

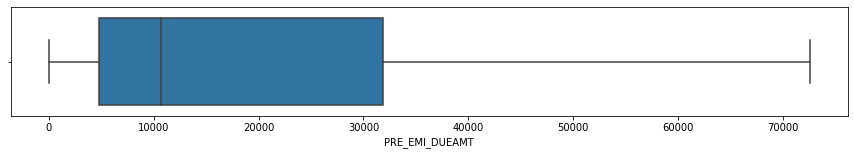




* Pre Emi-Due amt – Before outlier treatment, Pre Emi Due amt had extreme outliers to 3.1 cr. After treatment most of the values lie approximately between 5k to 32k.

**Figure 30:**





## **7.4** [**Derived Metrics**](https://www.kaggle.com/ab9bhatia/complete-eda-for-loan-analysis#derived) **& Insights**

To increase the discriminatory power of the model, variables DPD, EMI OS amt & Number of Emi Changes was binned.

New variable names – DPD\_RANGE, EMI\_OSAMT\_RANGE & NUM\_EMI\_CHANGES\_RANGE.

**Table 5: As days past due increases the probability of foreclosure is high. The binning technique will help us assign more Foreclosure weights to the higher segment.**

**FORECLOSURE 0 1 All Per %**

**DPD\_RANGE**

**0-1 17113 1657 18770 9**

**1-30 546 59 605 10**

**30-60 217 22 239 9**

**60-90 148 26 174 15**

**90 and above 193 31 224 14**

**All 18217 1795 20012**

**Table 6: The % Foreclosure seen across for EMI OS bins are distinctive, hence would improve the discriminatory power of the model.**

**FORECLOSURE 0 1 All Per %**

**EMI\_OSAMT\_RANGE**

**0-10k 17153 1623 18776 5**

**10k-50k 346 62 408 15.2**

**50k-300K 492 79 571 13.8**

**300k and above 226 31 257 12.1**

**All 18217 1795 20012 14.0**

**Table 7: The %Foreclosures have a monotonically increasing trend as customers opt for more EMI changes**

**FORECLOSURE 0 1 All Per %**

**NUM\_EMI\_CHANGES\_RANGE**

**-5-2# 10880 916 11796 8**

**2-5# 5276 583 5859 10**

**5 and above 2061 296 2357 13**

**All 18217 1795 20012**

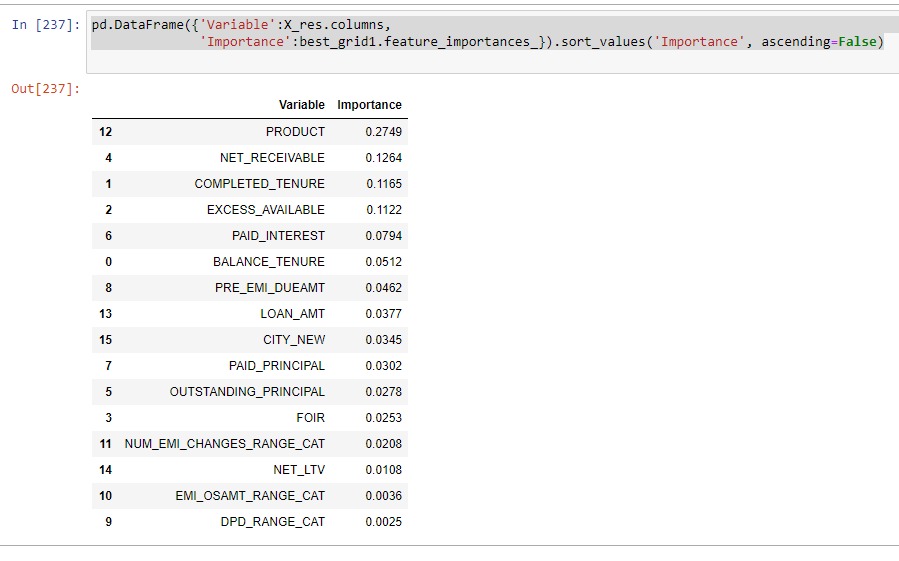
## **7.5 Logistic Regression Output – Model 4**

**Table 8**

|  |  |  |  |
| --- | --- | --- | --- |
| Logit Regression Results | | | |
| **Dep. Variable:** | FORECLOSURE | **No. Observations:** | 13408 |
| **Model:** | Logit | **Df Residuals:** | 13394 |
| **Method:** | MLE | **Df Model:** | 13 |
| **Date:** | Sun, 25 Apr 2021 | **Pseudo R-squ.:** | 0.1515 |
| **Time:** | 19:33:30 | **Log-Likelihood:** | -3426.5 |
| **converged:** | True | **LL-Null:** | -4038.5 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 1.172e-253 |

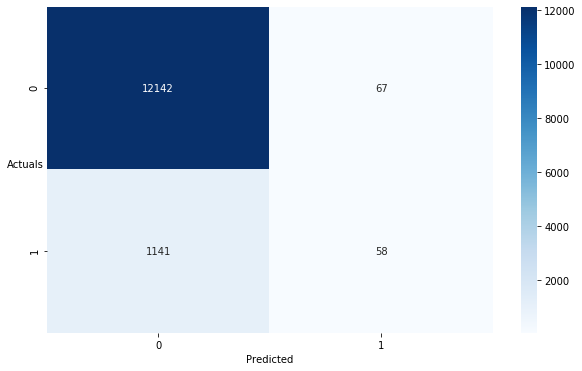
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | -0.2852 | 0.205 | -1.390 | 0.164 | -0.687 | 0.117 |
| **BALANCE\_TENURE** | -0.0039 | 0.001 | -5.152 | 0.000 | -0.005 | -0.002 |
| **EXCESS\_AVAILABLE** | 6.084e-05 | 1.04e-05 | 5.828 | 0.000 | 4.04e-05 | 8.13e-05 |
| **FOIR** | -0.8684 | 0.149 | -5.820 | 0.000 | -1.161 | -0.576 |
| **NET\_RECEIVABLE** | 0.0030 | 0.002 | 1.791 | 0.073 | -0.000 | 0.006 |
| **OUTSTANDING\_PRINCIPAL** | -1.167e-07 | 1.93e-08 | -6.047 | 0.000 | -1.55e-07 | -7.89e-08 |
| **PAID\_INTEREST** | 1.54e-06 | 9.8e-08 | 15.718 | 0.000 | 1.35e-06 | 1.73e-06 |
| **PAID\_PRINCIPAL** | -2.954e-06 | 2.92e-07 | -10.133 | 0.000 | -3.53e-06 | -2.38e-06 |
| **PRE\_EMI\_DUEAMT** | 1.121e-05 | 1.5e-06 | 7.462 | 0.000 | 8.26e-06 | 1.42e-05 |
| **NUM\_EMI\_CHANGES\_RANGE\_CAT** | 0.1303 | 0.050 | 2.596 | 0.009 | 0.032 | 0.229 |
| **PRODUCT** | -0.9828 | 0.045 | -22.011 | 0.000 | -1.070 | -0.895 |
| **LOAN\_AMT** | -2.548e-08 | 5.89e-09 | -4.327 | 0.000 | -3.7e-08 | -1.39e-08 |
| **NET\_LTV** | 0.0023 | 0.002 | 1.423 | 0.155 | -0.001 | 0.006 |
| **CITY\_NEW** | -0.0177 | 0.008 | -2.202 | 0.028 | -0.033 | -0.002 |

Table 8.2



7.5.1. LOGISTIC REGRESSION - WITH DEFAULT CUTOFF 0.5

**Figure 31:**

****

precision recall f1-score support

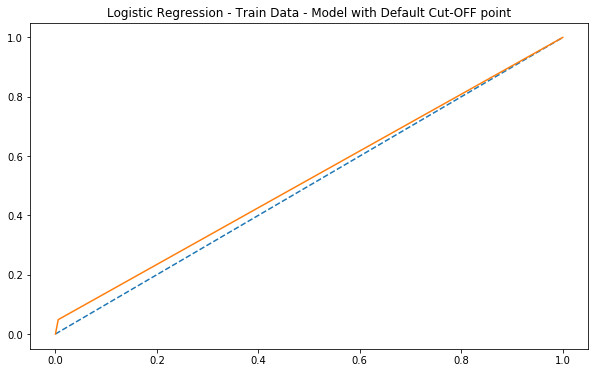
0 0.914 0.995 0.953 12209

1 0.464 0.048 0.088 1199

accuracy 0.910 13408

macro avg 0.689 0.521 0.520 13408

weighted avg 0.874 0.910 0.875 13408

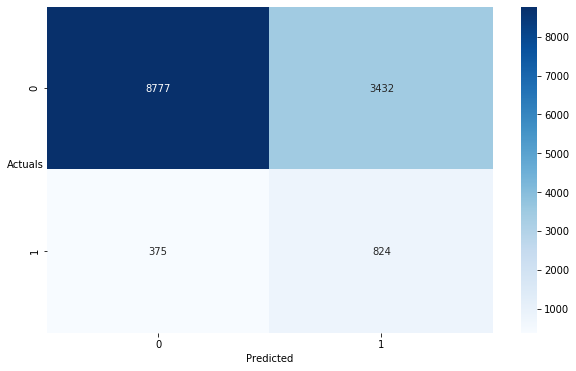


**Inference :** Recall at 4.8 percent and precision at 46.4 percent which only 4.8% defaults predicted correctly with a default cutoff 0.5. But Specificity 99 percent indicates that the most loan accounts are showing as non default.

AUC – 52

7.5.2. LOGISTIC REGRESSION – TRAIN DATA - WITH OPTIMUM CUTOFF 0.09

**Figure 32:**

****

precision recall f1-score support

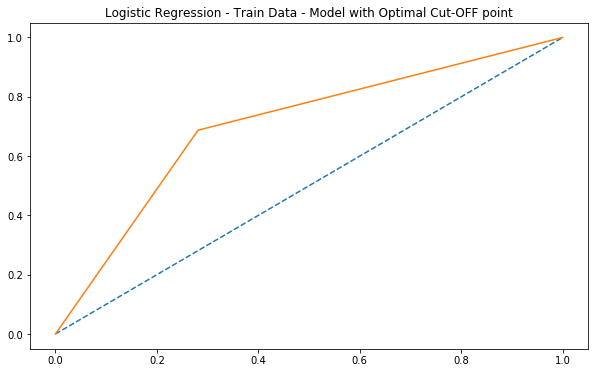
0 0.959 0.719 0.822 12209

1 0.194 0.687 0.302 1199

accuracy 0.716 13408

macro avg 0.576 0.703 0.562 13408

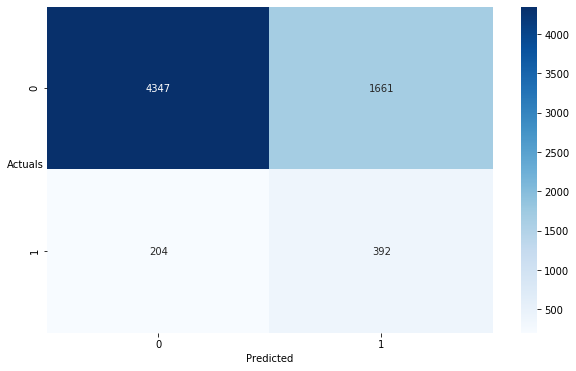
weighted avg 0.891 0.716 0.775 13408



**Inference :** Recall at 68 percent and precision at 19.4 percent is lowest, with 68% defaults is predicted correctly with a optimum cutoff 0.09. Specificity 71.9 percent. AUC – 70

7.5.3. LOGISTIC REGRESSION – TEST DATA - WITH OPTIMUM CUTOFF 0.09

**Figure 33:**

****

precision recall f1-score support

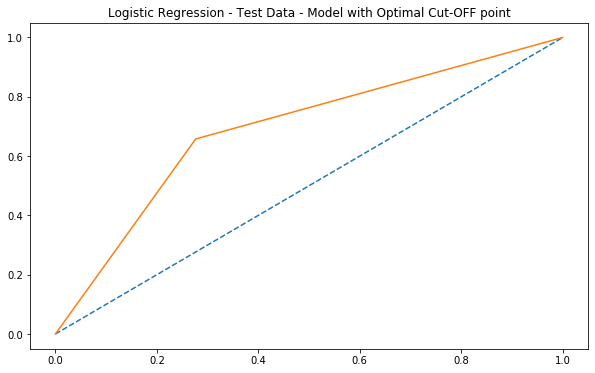
0 0.955 0.724 0.823 6008

1 0.191 0.658 0.296 596

accuracy 0.718 6604

macro avg 0.573 0.691 0.560 6604

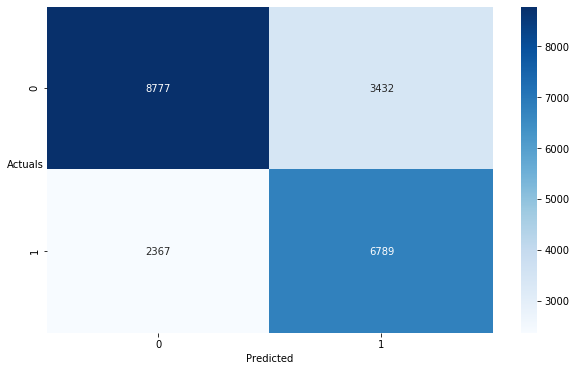
weighted avg 0.886 0.718 0.776 6604



**Inference :** Test setRecall reduced to 65.8 percent and precision at 19.1 percent is lowest, with 65.8% defaults is predicted correctly with a optimum cutoff 0.09. Specificity 72.4 percent. AUC – 69

7.5.4. LOGISTIC REGRESSION – SMOTE DATA – TRAIN DATASET – CUTOFF – 0.09

**Figure 34:**

****

precision recall f1-score support

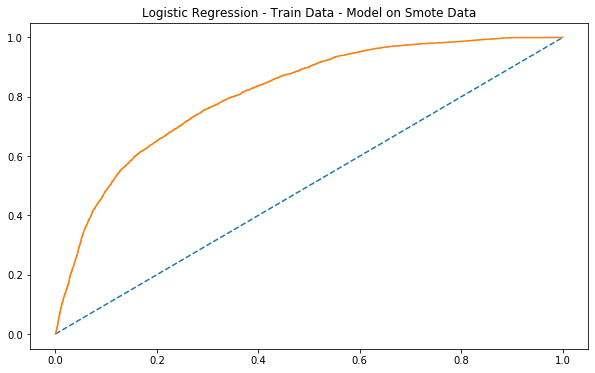
0 0.788 0.719 0.752 12209

1 0.664 0.741 0.701 9156

accuracy 0.729 21365

macro avg 0.726 0.730 0.726 21365

weighted avg 0.735 0.729 0.730 21365

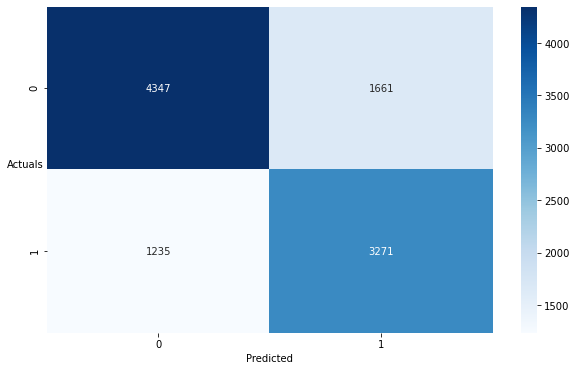


**Inference:** Recall at 74 percent and precision at 66 percent which 74% of defaults predicted correctly with a optimum cutoff 0.09 is a very good model when smote is applied. Recall is at maximum compared to past 3 summary of logistic regression. Both Recall and precision are high with a regularized data.

AUC- 81.

7.5.5. LOGISTIC REGRESSION – SMOTE DATA – TEST DATASET – CUTOFF – 0.09

**Figure 34.1**



precision recall f1-score support

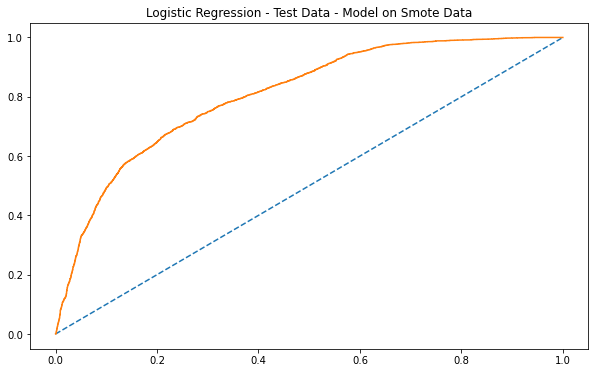
0 0.779 0.724 0.750 6008

1 0.663 0.726 0.693 4506

accuracy 0.725 10514

macro avg 0.721 0.725 0.722 10514

weighted avg 0.729 0.725 0.726 10514



**Inference:** Recall at 73 percent and precision at 66 percent which 73% of defaults predicted correctly with an optimum cutoff 0.09 is a very good model when smote is applied on test data. Recall is equal to the train data of logistic regression. Both Recall and precision are high with a regularized data.

AUC- 81.

Both Smote Train data and Smote Test Data have similar metrics, hence it is a good model.

## **7.6 LDA - LINEAR DISCRMINANT ANALYSIS**

**LDA With default values for both train and test datasets.**

**Table 9**

precision recall f1-score support

0 0.92 0.98 0.95 12209

1 0.39 0.12 0.18 1199

accuracy 0.90 13408

macro avg 0.66 0.55 0.57 13408

weighted avg 0.87 0.90 0.88 13408

**Table 10**

precision recall f1-score support

0 0.92 0.98 0.95 6008

1 0.37 0.10 0.16 596

accuracy 0.90 6604

macro avg 0.64 0.54 0.55 6604

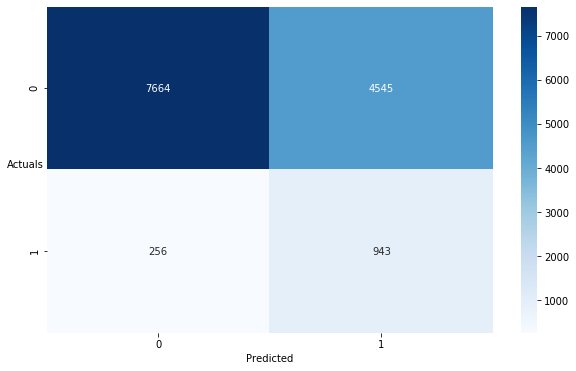
weighted avg 0.87 0.90 0.88 6604

**Inference:**

Recall for both train and test data for LDA model with default values show poor recall scores of 12 & 10 percent and having precision being lowest. Prediction of loan defaults correctly at 10 percent levels is very poor metrics.

7.6.1. LDA – TRAIN DATASET – CUTOFF – 0.06

**Figure 35**

****

precision recall f1-score support

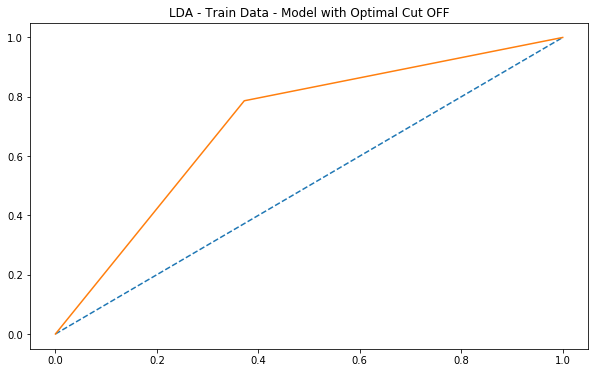
0 0.968 0.628 0.761 12209

1 0.172 0.786 0.282 1199

accuracy 0.642 13408

macro avg 0.570 0.707 0.522 13408

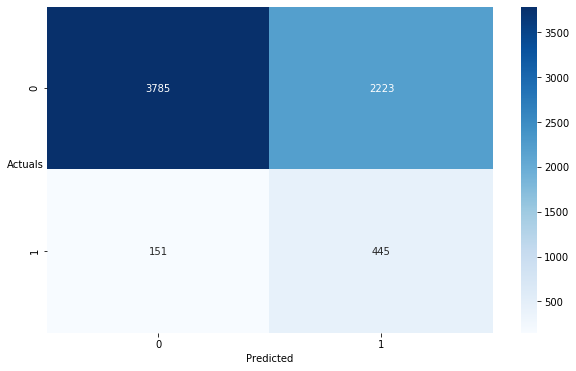
weighted avg 0.897 0.642 0.719 13408



**Inference:** Recall at 78 percent and precision at 17 percent which 78% of defaults predicted correctly with a optimum cutoff 0.06 is a very good but precision being low. AUC- 70.

7.6.2. LDA – TEST DATASET – CUTOFF – 0.06

**Figure 36**

****

precision recall f1-score support

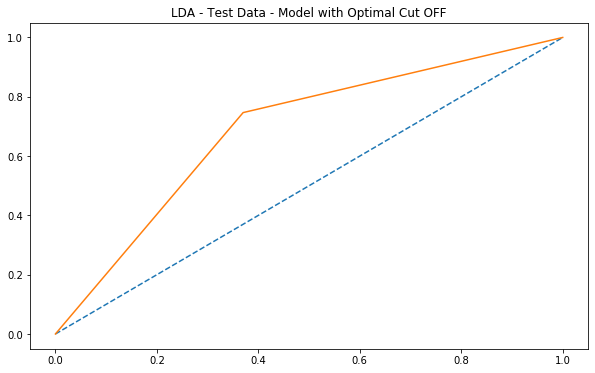
0 0.962 0.630 0.761 6008

1 0.167 0.747 0.273 596

accuracy 0.641 6604

macro avg 0.564 0.688 0.517 6604

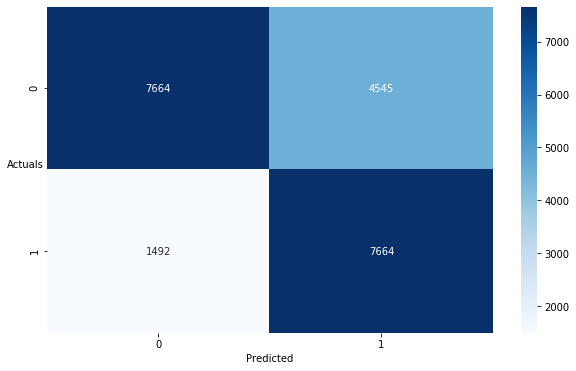
weighted avg 0.890 0.641 0.717 6604



**Inference:** Recall reduced to 74 percent on test data and precision at 16 percent which 74% of defaults predicted correctly with a optimum cutoff 0.06 is a very good but precision being low. AUC- 68.

7.6.3. LDA – SMOTE DATASET – CUTOFF – 0.06

**Figure 37**

****

precision recall f1-score support

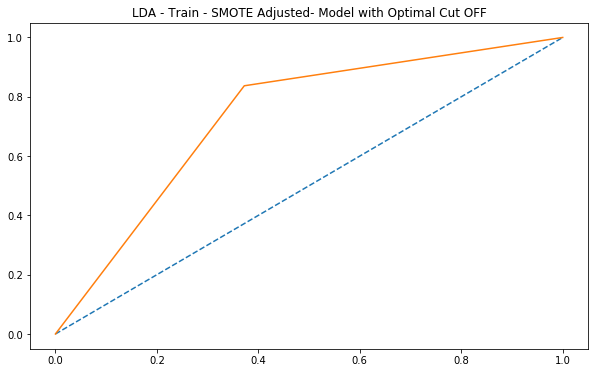
0 0.837 0.628 0.717 12209

1 0.628 0.837 0.717 9156

accuracy 0.717 21365

macro avg 0.732 0.732 0.717 21365

weighted avg 0.747 0.717 0.717 21365



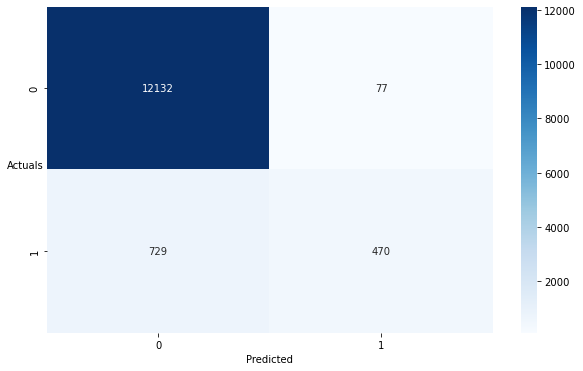
**Inference:** Recall at 83 percent and precision at 62 percent which 83% of loan defaults predicted correctly with a optimum cutoff 0.06 is a very good model when smote is applied. Recall is at maximum compared to past 3 summary of LDA. Both Recall and precision are high with a regularized data.

AUC- 73.

## **7.7. RANDOM FOREST MODEL**

7.7.1. RANDOM FOREST – TRAIN DATASET

**Figure 38**

****

precision recall f1-score support

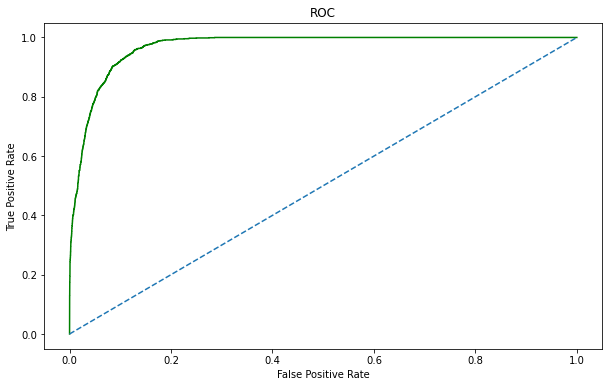
0 0.943 0.994 0.968 12209

1 0.859 0.392 0.538 1199

accuracy 0.940 13408

macro avg 0.901 0.693 0.753 13408

weighted avg 0.936 0.940 0.929 13408

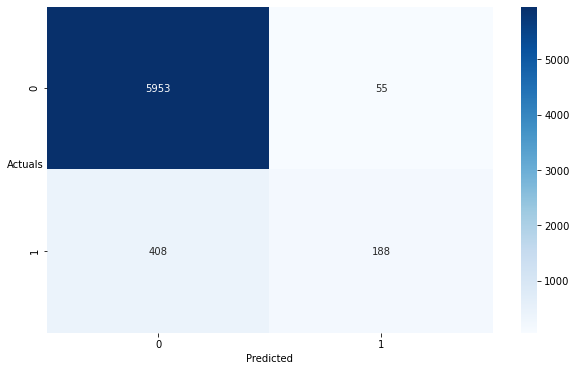


**Inference:** Recall at 39 percent and precision at 85 percent which 39% of loan defaults predicted correctly which is very low.

AUC- 69.

7.7.2. RANDOM FOREST – TEST DATASET

**Figure 39**

****

precision recall f1-score support

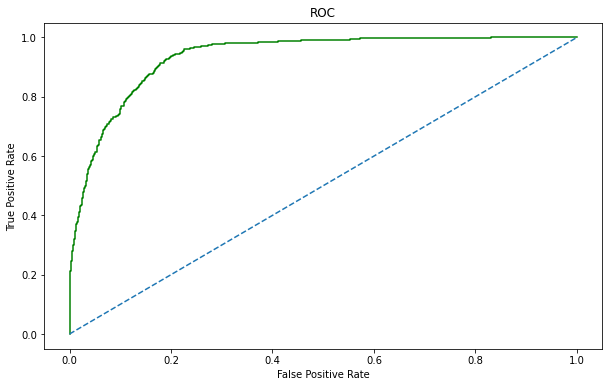
0 0.936 0.991 0.963 6008

1 0.774 0.315 0.448 596

accuracy 0.930 6604

macro avg 0.855 0.653 0.705 6604

weighted avg 0.921 0.930 0.916 6604

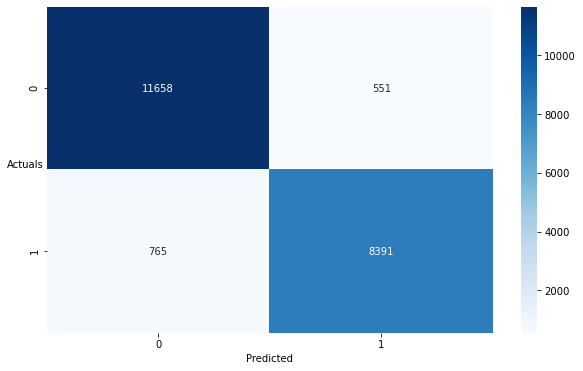


**Inference:** Recall reduced to 31 percent and precision at 77 percent which 31% of loan defaults predicted correctly which is very low.

AUC- 65.

7.7.3. RANDOM FOREST – SMOTE DATASET

**Figure 40**

****

precision recall f1-score support

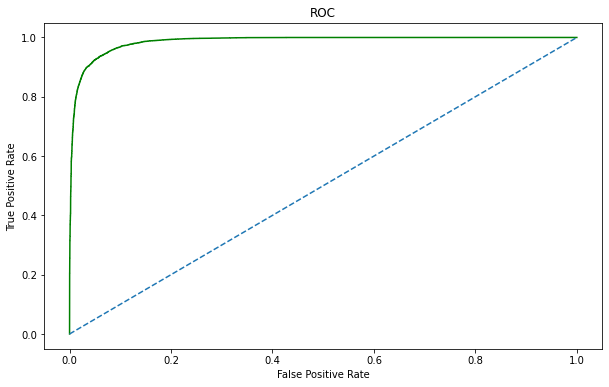
0 0.938 0.955 0.947 12209

1 0.938 0.916 0.927 9156

accuracy 0.938 21365

macro avg 0.938 0.936 0.937 21365

weighted avg 0.938 0.938 0.938 21365



**Inference:** Recall drastically increased to 91 percent and precision at 93 percent which 91% of loan defaults predicted correctly with a optimum best parameters is a very good model when smote is applied. Recall is at maximum compared to all models. Both Recall and precision are high with a regularized data. AUC- 93.