Christopher Dennies

Batch 1 - April 2020

Abstract

**Highlighting the Important driving factors which can help NBFC ‘s to take a prior action to avoid foreclosure and thereby reducing the cost incurred by the Foreclosure Process.**

BuSiness report

NBFC Foreclosure Prediction Notes 1

**Table of Contents**

**I. Introduction of the Business Problem**

**II. Data Report**

**III. Missing value Treatment & Dropping variables**

**IV. Correlation Plot & Dropping variables**

**V. Applying VIF & Dropping variables**

**VI. Outlier Treatment / Univariate Analysis**

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

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

**X. Business Insights from EDA**

**I. Introduction of the Business Problem**

**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.

Because of the High Foreclosures costs, the lenders are looking forward to a solution in which they can avoid the cumbersome process.

**Business Implications of the Study:**

Prediction of driving factors leading to 'FORECLOSURE' of the loan will help the NBFC to take prior actions while sanctioning and during payment tenure, thereby ensuring to avoid a Foreclosure process.

By identifying and implementing measures from the study, the business outcome of an NBFC is far more beneficial in cutting down the cost and at the same time retaining the customers in the long run.

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).

**II. Data Report**

**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 |

**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.
* 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. Ie < 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**

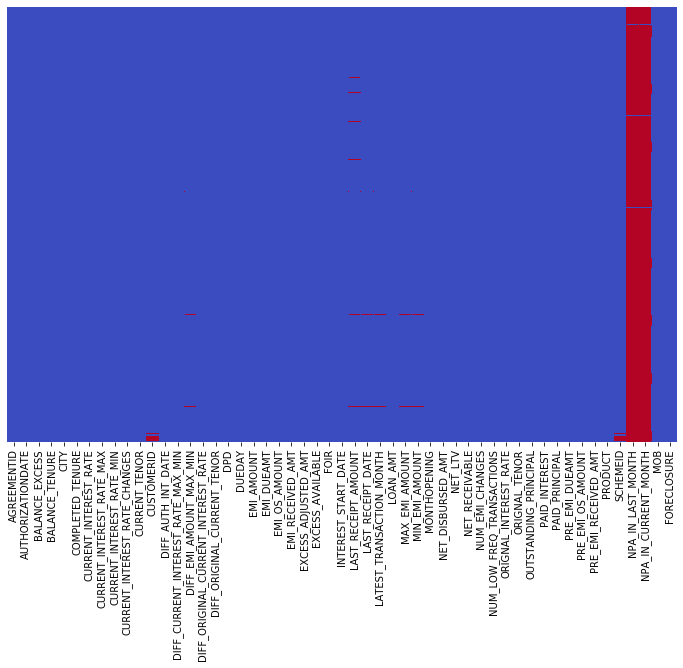
**MIN\_EMI\_AMOUNT 0.4400**

**SCHEMEID 1.4000**

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

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

Figure 1 : Visual Presentation of missing values



**III. Missing value Treatment & Dropping variables**

* **Agreement Id** variable is retained because it holds the distinct count of Foreclosure accounts.
* **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 foreclosure prediction, which is dropped.

* **MOB** is an internal code and the data has no extra information,

which will not help in foreclosure prediction, which is dropped.

* **NPA\_IN\_LAST\_MONTH** variable has 99.41 missing values and only 2 Foreclosures of 15 NPA's, which is not a good predictor will drop this variable. Refer Below table: **Table 1:**

| **FORECLOSURE** | **0** | **1** | **All** |
| --- | --- | --- | --- |
| **NPA\_IN\_LAST\_MONTH** |  |  |  |
| **0** | 69 | 33 | 102 |
| **#N/** | 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: **Table 2:**

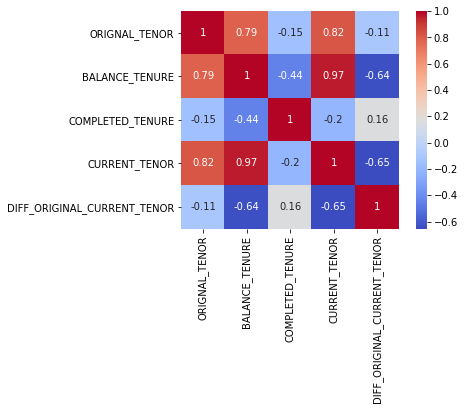
| **FORECLOSURE** | **0** | **1** | **All** |
| --- | --- | --- | --- |
| **NPA\_IN\_CURRENT\_MONTH** |  |  |  |
| **0** | 70 | 33 | 103 |
| **Yes** | 14 | 2 | 16 |
| **All** | 84 | 35 | 119 |

* **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.

**IV. Correlation Plot & Dropping variables**

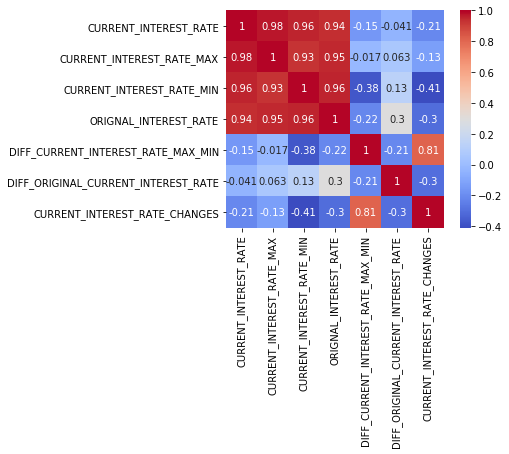
* From the below correlation graph Figure 2, Original Tenor, Balance Tenor and Current Tenor are highly correlated, Balance Tenor will be retained along with Completed Tenor, with domain understanding. Dropping Original Tenor, Current Tenor & difference between original and current tenor.

**Figure 2 :**



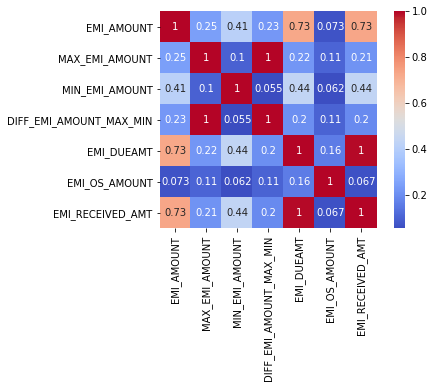
* From Below **Figure 3 :** Current Interest rate is highly correlated with other version of available interest rates(Max, Min & Original), Current Interest rate will be retained, others dropped.
* Difference between Current max and Current min, Difference between Original and Current Interest Rate & Current interest rate changes dropped as no insights derived from it.

​ **Figure 3 :**

****

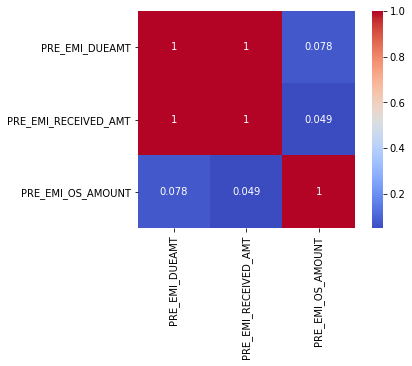
* From **Figure 4:** EMI Amount and Outstanding EMI amount and Received amount are more intuitive to use when compared to other variation of EMI variables. Rest other variables dropped.

**Figure 4:**

****

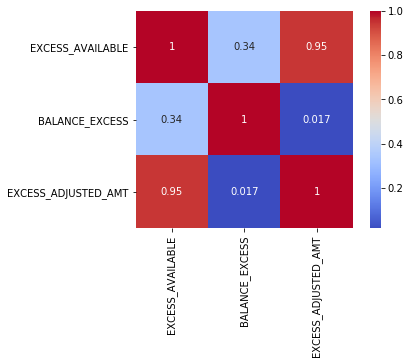
* From **Figure 5:** Pre-EMI Due amount & Pre EMI Received Amount are perfectly highly correlated, in the context of foreclosure the 'pre emi due amount' will be retained along with 'Pre Emi OS amount'.

**Figure 5:**

****

* From **Figure 6:** Excess Available and Excess Adjusted Amount are highly correlated, 'Excess Available' will be retained along with 'Balance Excess'.

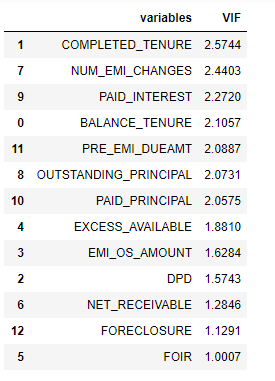
**Figure 6:**

****

**V. Applying VIF & Dropping variables**

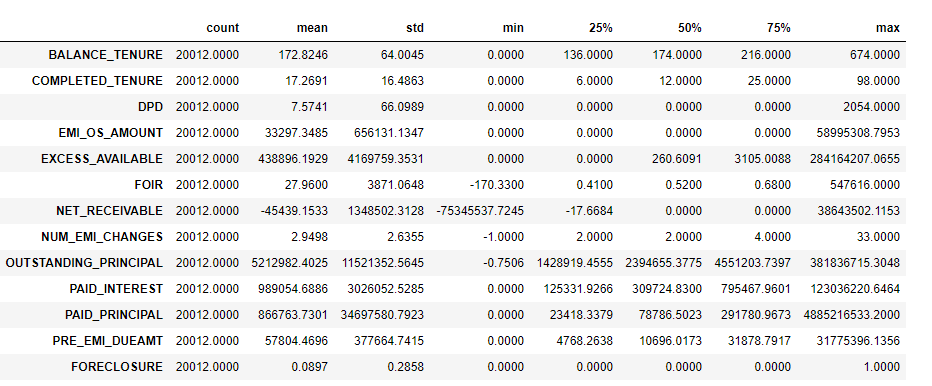
* Variance inflation factor applied to 26 variables with a cut off below 5 , dropped to 12 significant variables excluding the target variable Foreclosure.

**Table 3 :**



* From the below 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.

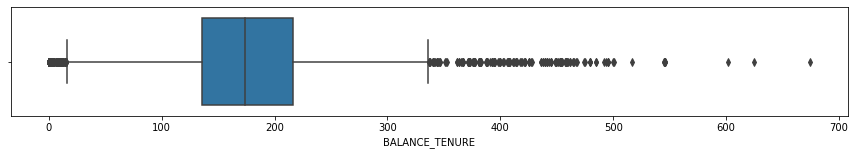
**Table 4 :**

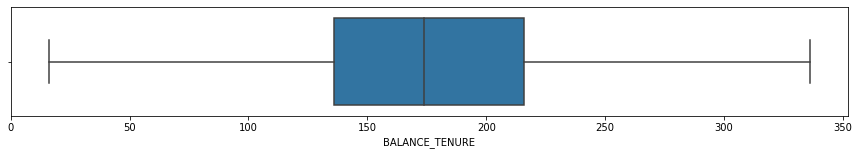


**VI. Outlier Treatment / Univariate Analysis**

* Outlier treatment applied to 9 variables.
* 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.

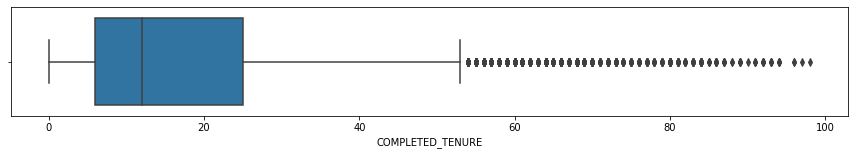
**Figure 7:**

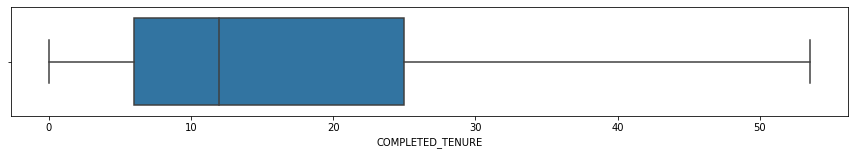
****

****

* 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 8:**

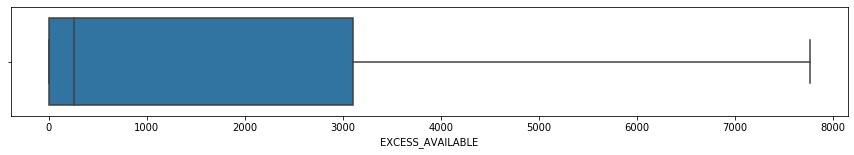




* Excess available – Before outlier treatment, Excess available had extreme outliers to 28 cr odd. After treatment most of the values lie approximately between 0 to 3k.

**Figure 9:**

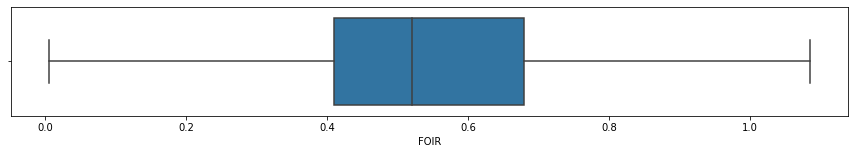
****

****

* 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 10:**

****

****

* 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 11:**

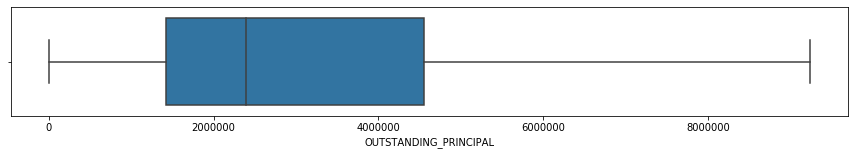
****

****

* 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 12:**

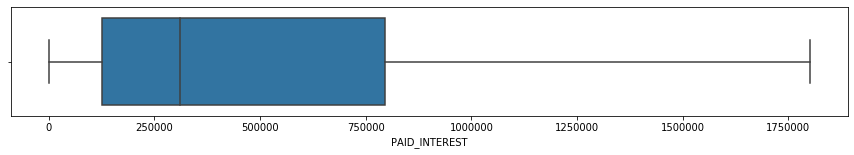
****

****

* 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 13:**

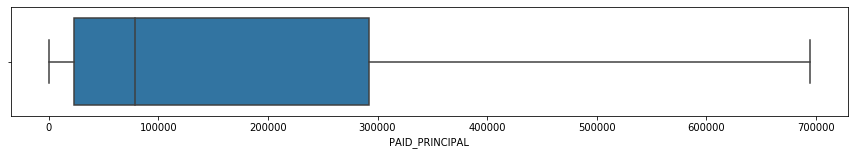
****

****

* 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.

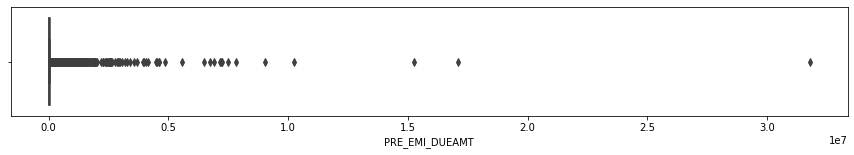
**Figure 14:**

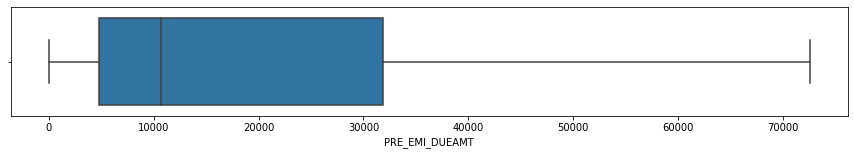
****

****

* 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 15:**

****

****

**VII.** [**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**

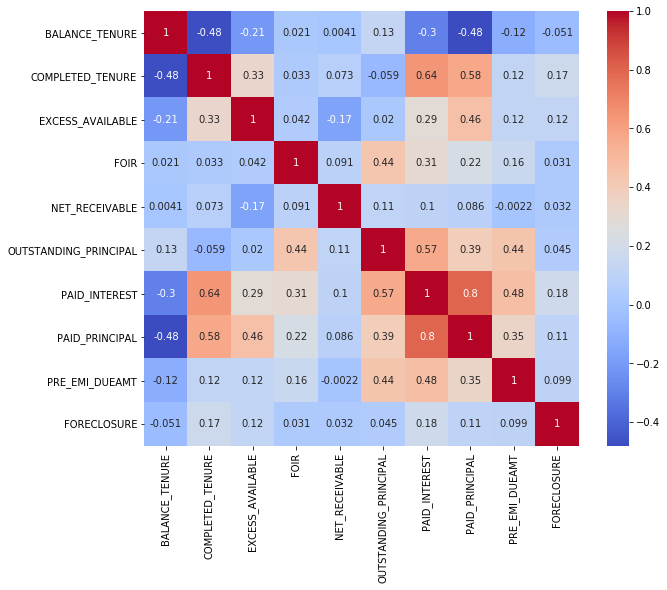
**5 and above 2061 296 2357 13**

**All 18217 1795 20012**

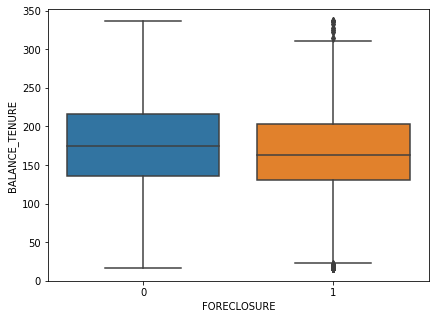
[**VIII. Bivariate/Multivariate Analysis**](https://www.kaggle.com/ab9bhatia/complete-eda-for-loan-analysis#bivariate)

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

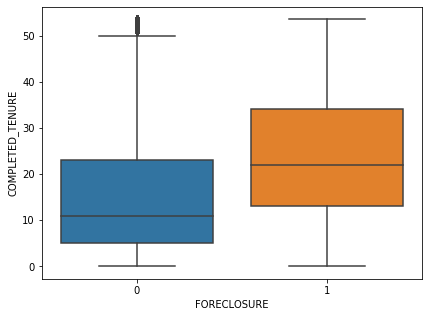
**Figure 16:**

****

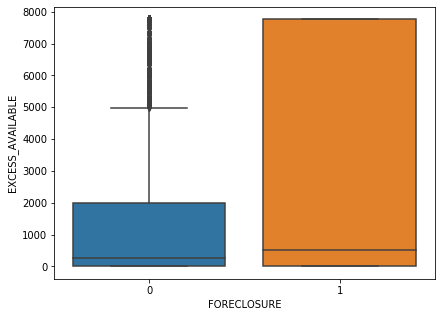
**Figure 17: Foreclosure and non-foreclosure distribution is similar. Unlikely to be a strong predictor.**

****

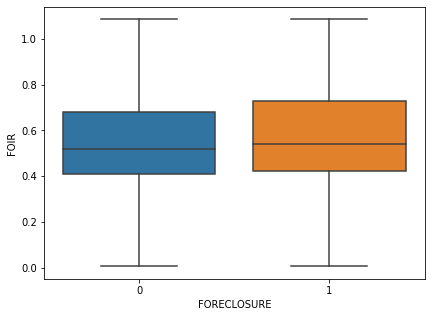
**Figure 18: Foreclosure and Non-Foreclosure population distribution is different and distinct, likely to be a Strong Predictor**

****

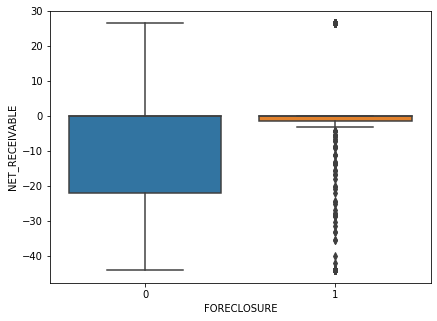
**Figure 19: Distributions are not similar and very likely to be strong predictor**

****

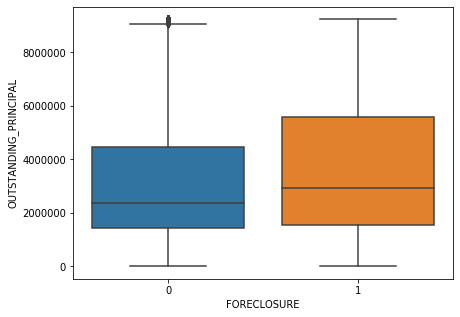
**Figure 20: FOIR – distributions are fairly similar like to be a weak predictor**

****

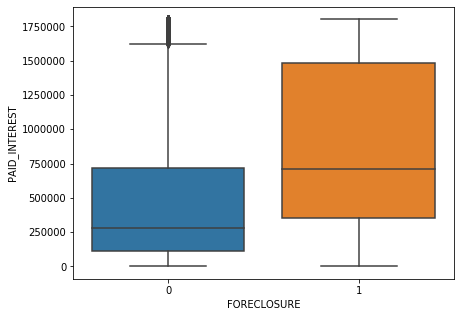
**Figure 21: Net-Receivable distribution between foreclosure and non-foreclosure distributions are not similar, likely to be strong predictor**

****

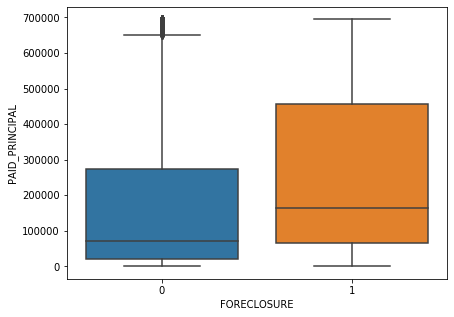
**Figure 22: Higher the outstanding principal likely the customers to foreclose, likely to be a strong predictor**

****

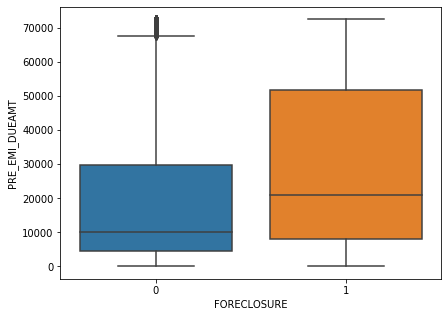
**Figure 23 : Customer paying more interest are likely to Foreclose, could be an important variable in the final model**

****

**Figure 24: This variable is contrary to business understanding; yet the distributions are different. Likely to be removed in further analysis**

****

**Figure 25: Distribution are quite distinctive in nature; likely to be a strong predictor**

****

**X. Business Insights from EDA**

**a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business**

Yes, the data is unbalanced.

Smote is a technique which can done to regularize the data. Imbalance of data will create a biased model. Best technique to avoid the curse of imbalanced data is by under-sampling the larger classified dataset and by oversampling the less classified dataset. So that, the final dataset will have a balanced/equal amount of data among all the labels

In the Context of business, the model created for the NBFC should neither be underfit or overfit as to generalize in real world conditions as we are in state of flux. ie. Constantly changing.

**b) Any business insights using clustering (if applicable)**

The clustering was performed on scaled and unscaled data on the final variables after exploratory data analysis (12 variables), the clustering results are not definitive, and the scree plot are not helpful in identifying the number of clusters required. However, predictive modelling solution is the advised here with the given data of 9% foreclosures.

**c) Any other business insights**

As observed in the derived variables section,

* As the days past due increases the probability of foreclosures are high. The binning technique will help us assign more Foreclosure weights to the higher DPD segments as we observe a slight monotonically increasing trend
* The % of Foreclosure seen across for EMI OS bins are quite distinctive, hence the binning would improve the discriminatory power of the model
* The % of Foreclosures have a monotonically increasing trend as customers opt for more EMI changes are vulnerable to foreclosure behavior