Capstone Project : NBFC (Foreclosure)

Notes- I

Vikram Radhakrishnan

13 December 2020

NBFC

Before we move on to solving the business problem, lets look at some basic information that will help us understand this project better.

What is an NBFC?

A Non-Banking Financial Company (NBFC) is a company registered under the Companies Act, 2013 of India, engaged in the business of loans and advances, acquisition of shares, stock, bonds, hire-purchase insurance business or chit-fund business, but does not include any institution whose principal business is that of agriculture, industrial activity, purchase or sale of any goods (other than securities) or providing any services and sale/purchase/construction of immovable property.

How are they different from Banks?

While NBFCs may appear to be similar to banks, there are distinct differences that set them apart from regular Banks.

- NBFCs provide banking services to people without holding a Bank licence,
- An NBFC cannot accept Demand Deposits,
- An NBFC is not a part of the payment and settlement system
- An NBFC cannot issue Cheques drawn on itself
- Deposit insurance facility of the Deposit Insurance and Credit Guarantee Corporation is not available for NBFC depositors, unlike banks
- An NBFC is not required to maintain Reserve Ratios (CRR, SLR etc.)
- An NBFC cannot indulge primarily in agricultural or industrial activities or sale-purchase, construction of immovable property
- Foreign Investment allowed up to 100 %
- An NBFC accompanies working in Financial Body and Money handling

When or Why are they preferred over banks?

That leads to one to question – why would one approach an NBFC instead of a bank?

- Quick Disbursal of Funds
- Competitive Interest rates
- Lenient Eligibility Criteria
- Relatively minimal Paperwork and Documentation

What is a foreclosure?

Foreclosure is the legal process by which a lender (NBFC in this case) seizes and sells a home or property after a borrower is unable to fulfill his or her repayment obligation.

1) Introduction of the business problem

Defining the problem statement:

In our current case study, we are provided with aggregate data consisting of multiple parameters, of loans that have been disbursed an NBFC. Each of these entries also has a parameter describing whether that particular loan was foreclosed or not. Using this data as as a reference or learning data, one has to predict whether a particular loan taken outside this dataset would result in a default/foreclosure or remains status quo. It is therefore what we refer to in statistical terms, as a binary classification problem with a supervised learning approach.

Need of the Study:

Evaluating whether a new loan can be disbursed to an applicant is a time consuming and tedious process, if done manually. In a manual process, the disburser would usually need to take a decision based on intuition based on previous experience or apply empirical methods. This intuition or application of experience is of course, applied after checking a few limited 'financial health' parameters of the borrower such as credit scores, earning history, assets and other indicators that support his image as a future non-defaulter. As with all other human predictions, there is scope for errors in calculations and there is a need to minimize these errors too. There is therefore, a need for a better way to predict the chances of a borrower's chances of defaulting.

Understanding business/social opportunity

It is a great need for financial institutions in today's fast paced world, to not only automate the process of prediction of loan defaults but do so with higher confidence and accuracy based on patterns of transactions that have happened in the past. Loans need to be disbursed in the shortest time possible time to remain competitive and customer friendly.

Machine learning algorithms can be very useful here by 'learning' from past data and predicting the likelihood of a borrower's chance to default from a large number of parameters. This process is not only very fast but can provide better reliability and accuracy compared to human judgement when implemented correctly.

2)Data Report

a) Understanding how data was collected in terms of time, frequency and methodology

The data provided lists information related to a series of loans authorized or active from 2010 to 2018, tracks their status (on a particular date) in terms of interest rates, payments, balances e.t.c.

Based on the values of the fields provided, it appears that the loan is tracked on a monthly basis.

The data contains both static and dynamic data. Some fields such as agreement ID, authorization date, customer ID, product e.t.c will remain fixed or static every month, whereas other fields such as paid principal, paid interest, month opening, due day e.t.c will keep changing with time (dynamic).

b) Visual inspection of data (rows, columns, descriptive details)

From a quick perusal of the data, we see observe that there are 53 fields, our columns in our case. 52 of these can be considered as independent variable which influence a single dependent variable, which in this case is 'foreclosure'. Some of these independent variables will influence 'foreclosure' and some may not – for example, agreement ID or customer ID are only identifiers and may not affect the result itself. However, its possible that if a customer defaults on one loan, he may default on the others as well. Therefore its important to understand which features or columns to keep for the analysis and which can be ignored.

There are three 'date-time' fields, 4 'object' or string fields for city, product and NPA. The rest of the fields are numerical in the form of either float or integers.

There are 20012 records or rows in all. More details will follow in EDA.

c) Understanding of attributes (variable info, renaming if required)

The data is read in and basic information on the data types is queried. As mentioned earlier, there are three ‘date-time’ fields, 4 ‘object’ or string fields for city, product and NPA. The rest of the fields are numerical in the form of either float or integers.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20012 entries, 0 to 20011
Data columns (total 53 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   AGREEMENTID                           20012 non-null  int64
 1   AUTHORIZATIONDATE                     20012 non-null  datetime64[ns]
 2   BALANCE_EXCESS                        20012 non-null  float64
 3   BALANCE_TENURE                        20012 non-null  int64
```

Though there are minor spelling mistakes in the names of the fields, we are not renaming them. However, if required, we could do so.

3) Exploratory data analysis

Basic statistics of the data is also taken out. Information such as mean, median, standard deviation e.t.c are extracted for a quick understanding.

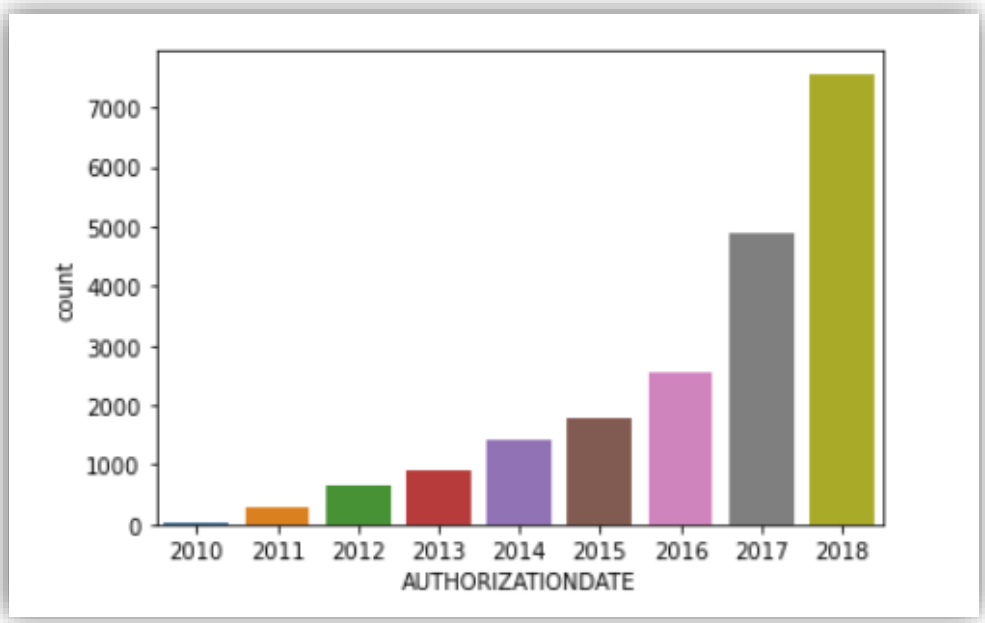
Duplicate records are checked for. No duplicates are found.

A. UNIVARIATE ANALYSIS

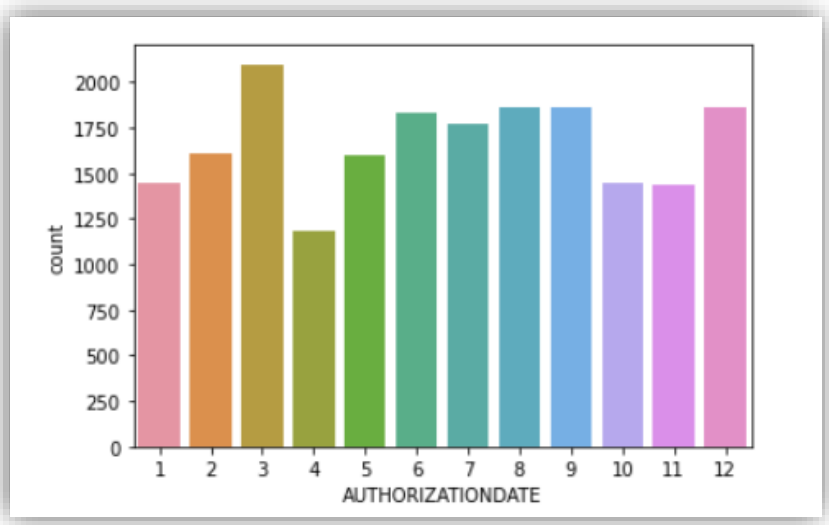
Univariate analysis is done on all columns of the dataset. In this report, we provide a small section of those analyses that might provide better insights.

AUTHORIZATION DATE

We observe that 2018 has the highest number of loans authorized and the amount of authorizations has increased exponentially or steadily Y-O-Y.



Highest number of loans appear to be taken in the month of March and the lowest in April.



CITY

The number of unique cities is found to be **272.**

The top 10 cities with the highest number of loans taken are the following:

MUMBAI	2028
HYDERABAD	1567
AHMEDABAD	1396
SURAT	1391
PUNE	1202
CHENNAI	1001
BANGALORE	880
THANE	735
DELHI	623
RAJKOT	558

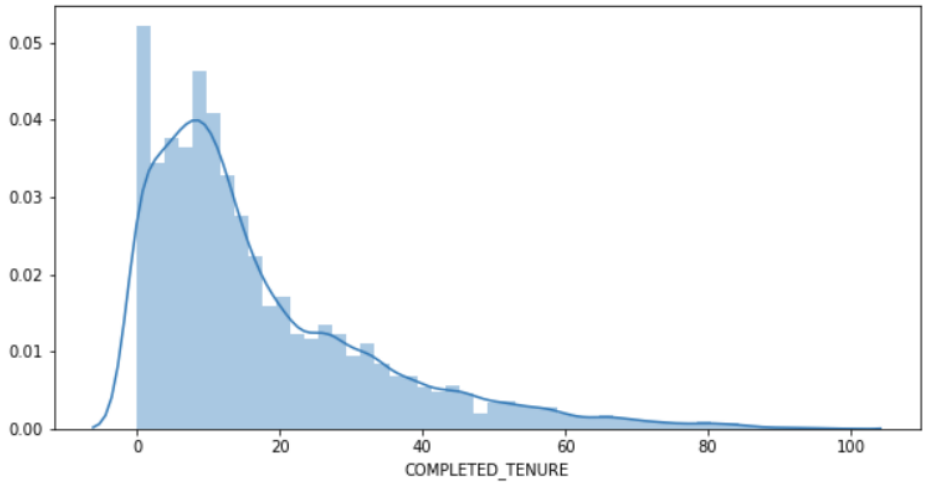
In terms of foreclosures, Mumbai again leads the pack.

MUMBAI	353
HYDERABAD	165
PUNE	151
CHENNAI	109
AHMEDABAD	90
BANGALORE	81
DELHI	76
THANE	75
COIMBATORE	69
SURAT	59

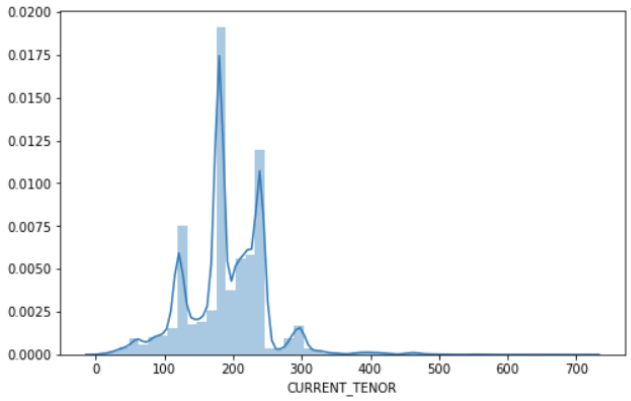
When seen as a **ratio of defaults to loans applied per city (*ratios below in figure*)**, it is seen that smaller cities are likely to default more. In terms of big cities (metros), Mumbai seems to be high on the list of likely default.

NATHAM	1.000000
VEPPAMPATTU	1.000000
CHIDAMBARAM	1.000000
SABARKANTHA	0.250000
DHARMAPURI	0.250000
SRIPERUMBUDUR	0.250000
TIRUCHIRAPPALLI	0.236025
MEDAK	0.200000
NADIAD	0.200000
ERODE	0.191111
MUMBAI	0.174063
GHAZIABAD	0.173913
THANJAVUR	0.171053
KUMBAKONAM	0.166667
MADURAI	0.165644
COIMBATORE	0.163507
BARODA	0.161290
TIRUNELVELI	0.160000
KOVILPATTI	0.153846
CHENGALPET	0.148148

COMPLETED TENURE



The completed tenured histogram is left skewed which shows that a large number of active loans have been taken in the last few years.



If we observe the current tenure, it is seen that the largest density is found in loans that are around 10,15 and 20 years.

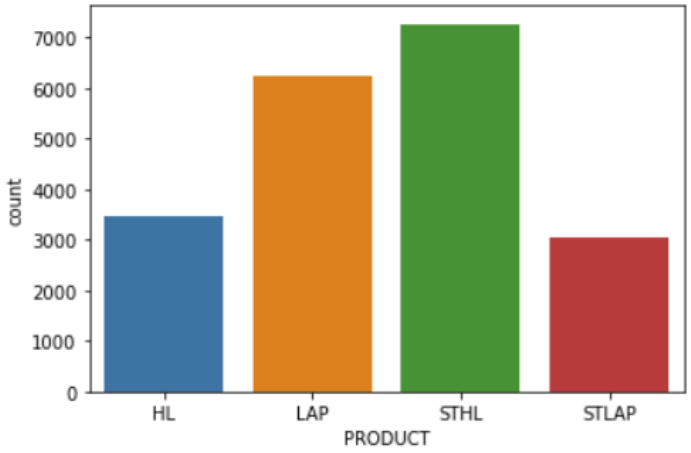
DIFF_AUTH_INT_DATE

For a large number of loans, the authorization of the loan and interest start date are the same.

0	19926
-1	49
5	8
2	4
3	4
-4	4
-2	3
1	2
7	2
6	1
4	1
-17	1
70	1
15	1
11	1
12	1
-3	1
14	1
9	1

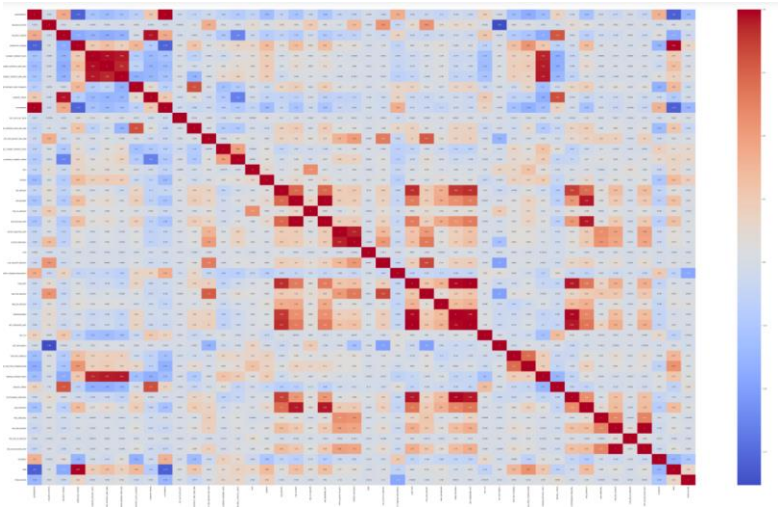
PRODUCT

The best selling product in terms of just number of products sold appears to be STHL and he lowest is STLAP.



B. BIVARIATE ANALYSIS

To start with, a heat map is drawn to find out the correlation between the variables. However, it is difficult to view the heatmap in the report in a single page.



Instead of viewing the heatmap graphically with multiple variables, the highly correlated feature pairs are printed instead. In this case, we print values that have a correlation value of more than 0.95.

```
('AGREEMENTID', 'CUSTOMERID')
('BALANCE_TENURE', 'CURRENT_TENOR')
('COMPLETED_TENURE', 'MOB')
('CURRENT_INTEREST_RATE', 'CURRENT_INTEREST_RATE_MAX')
('CURRENT_INTEREST_RATE', 'CURRENT_INTEREST_RATE_MIN')
('CURRENT_INTEREST_RATE_MAX', 'CURRENT_INTEREST_RATE')
('CURRENT_INTEREST_RATE_MAX', 'ORIGNAL_INTEREST_RATE')
('CURRENT_INTEREST_RATE_MIN', 'CURRENT_INTEREST_RATE')
('CURRENT_INTEREST_RATE_MIN', 'ORIGNAL_INTEREST_RATE')
('CURRENT_TENOR', 'BALANCE_TENURE')
('CUSTOMERID', 'AGREEMENTID')
('EMI_DUEAMT', 'EMI_RECEIVED_AMT')
('EMI_DUEAMT', 'PAID_INTEREST')
('EMI_RECEIVED_AMT', 'EMI_DUEAMT')
('EMI_RECEIVED_AMT', 'PAID_INTEREST')
('LOAN_AMT', 'MONTHOPENING')
('LOAN_AMT', 'NET_DISBURSED_AMT')
('LOAN_AMT', 'OUTSTANDING_PRINCIPAL')
('MONTHOPENING', 'LOAN_AMT')
('MONTHOPENING', 'NET_DISBURSED_AMT')
('MONTHOPENING', 'OUTSTANDING_PRINCIPAL')
('NET_DISBURSED_AMT', 'LOAN_AMT')
('NET_DISBURSED_AMT', 'MONTHOPENING')
('NET_DISBURSED_AMT', 'OUTSTANDING_PRINCIPAL')
('ORIGNAL_INTEREST_RATE', 'CURRENT_INTEREST_RATE_MAX')
('ORIGNAL_INTEREST_RATE', 'CURRENT_INTEREST_RATE_MIN')
('OUTSTANDING_PRINCIPAL', 'LOAN_AMT')
('OUTSTANDING_PRINCIPAL', 'MONTHOPENING')
('OUTSTANDING_PRINCIPAL', 'NET_DISBURSED_AMT')
('PAID_INTEREST', 'EMI_DUEAMT')
('PAID_INTEREST', 'EMI_RECEIVED_AMT')
('PRE_EMI_DUEAMT', 'PRE_EMI_RECEIVED_AMT')
('PRE_EMI_RECEIVED_AMT', 'PRE_EMI_DUEAMT')
('MOB', 'COMPLETED_TENURE')
```

If we adjust the value of correlation threshold to 0.98, the following values are observed to be well correlated.

```
('AGREEMENTID', 'CUSTOMERID')
('COMPLETED_TENURE', 'MOB')
('CUSTOMERID', 'AGREEMENTID')
('EMI_DUEAMT', 'EMI_RECEIVED_AMT')
('EMI_RECEIVED_AMT', 'EMI_DUEAMT')
('LOAN_AMT', 'NET_DISBURSED_AMT')
('MONTHOPENING', 'NET_DISBURSED_AMT')
('MONTHOPENING', 'OUTSTANDING_PRINCIPAL')
('NET_DISBURSED_AMT', 'LOAN_AMT')
('NET_DISBURSED_AMT', 'MONTHOPENING')
('OUTSTANDING_PRINCIPAL', 'MONTHOPENING')
('PRE_EMI_DUEAMT', 'PRE_EMI_RECEIVED_AMT')
('PRE_EMI_RECEIVED_AMT', 'PRE_EMI_DUEAMT')
('MOB', 'COMPLETED_TENURE')
```

a) **Removal of unwanted variables:**

Certain columns which do not have sufficient data such as NPA_IN_LAST_MONTH & NPA_IN_CURRENT_MONTH are removed. Next, we can look at variables that are highly correla ted and then proceed to remove some features that may not be necessary. CUSTOMER_ID is an other column that may not be necessary as there is already a primary identification in the form of AGREEMENT_ID. Datetime values such as 'AUTHORIZATIONDATE', 'INTEREST_START_DATE' c an be removed since their difference is already captured under a variable DIFF_AUTH_INT_DAT E.

b) **Null/Missing Value treatment :**

A null value check is done and the following features have null values. The number of null values in each features has also been provided in the table.

CUSTOMERID	281
DIFF_EMI_AMOUNT_MAX_MIN	89
LAST_RECEIPT_AMOUNT	247
LAST_RECEIPT_DATE	75
LATEST_TRANSACTION_MONTH	75
MAX_EMI_AMOUNT	89
MIN_EMI_AMOUNT	89
SCHEMEID	281
NPA_IN_LAST_MONTH	19893
NPA_IN_CURRENT_MONTH	19893

Treatment of null values:

Treatment of missing values by imputation or removal is important in performing logistic regression. CART may be able to handle missing data since it has in-built algorithms to impute missing values with surrogate variables. We will we look at treatment further during modeling process. However a few points on a couple of features:

CUSTOMER ID : The null values in this column cannot be treated using the usual ways of imputation since these are not statistical values. There are two options here – Either remove this column or keep the column as it is for later reference. We choose the second option.

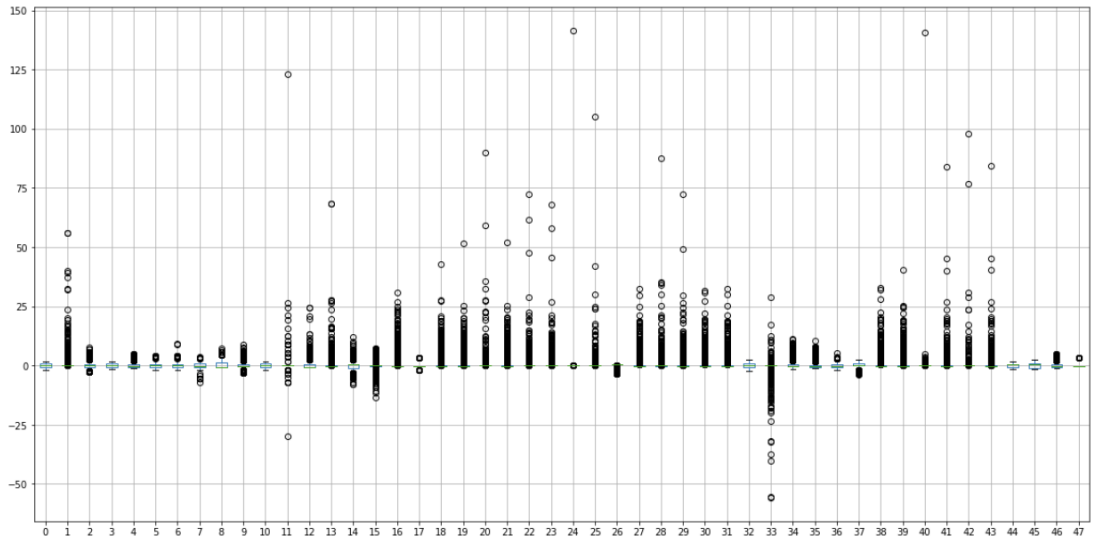
DIFF_EMI_AMOUNT_MAX_MIN: This shows the difference between two columns "MAX_EMI_AMOUNT" and "MIN_EMI_AMOUNT". Therefore the missing values can be derived from these two columns, as long as they have values.

We do not drop any columns or any impute any values at the moment.

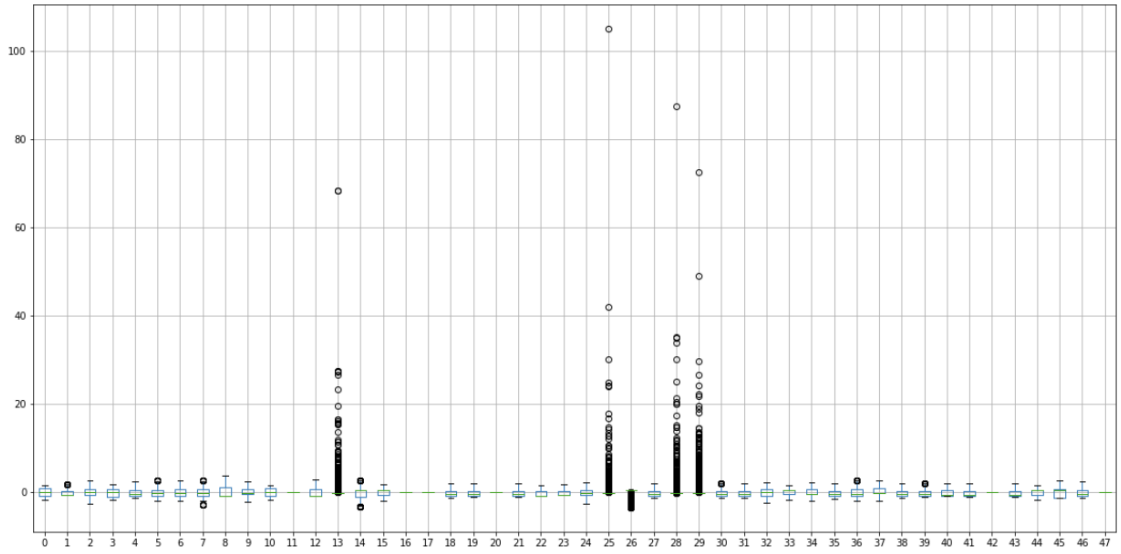
c) **Outlier treatment :**

The 1.5 IQR technique can be used to treat outliers, where 1.5xIQR is subtracted from the 1st quartile and 1.5 x IQR is added to the 3rd quartile value. Any number outside these limits is treated as an outlier and moved to the nearest outlier limit.

BEFORE TREATMENT



AFTER OUTLIER TREATMENT



D) Variable transformation (if applicable)

In our case, we have made only a few transformations to variables such as converting few categorical variables into numerical variables through encoding.

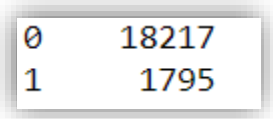
E) Addition of new variables (if required)

Addition of new variables has not been done at this stage and can be carried out in the modeling stage to reduce the number of features and extract more information.

4. Business insights from EDA

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

In order to check if the data is balanced, we compare the number of records which show default/foreclosure [1] to the number of records that do not [0].



It is observed that less than 10% of records show foreclosure, while more than 90% of records have not defaulted. This leads to an imbalance in training data due to the need of more records that show foreclosure. In the real world, its sometimes difficult to get this data.

We can therefore consider performing over-sampling or under-sampling of the data.

Over-Sampling is when some of the default case records are duplicated or copied in order in order to have a number of records close to that of non-default cases.

Under-Sampling is when some of the records of the non-default cases are removed in order to have close to an equal number of cases for both results.

In reality though, we should not simply perform over- or under-sampling on our training data and then run the model. We need to account for cross-validation and perform over- or under-sampling on each fold independently to get an honest estimate of model performance.

Other than over- and under-sampling, there are hybrid methods that combine under-sampling with the generation of additional data. Two of the most popular techniques used are ROSE and SMOTE.

b) Any business insights using clustering (if applicable)

Cluster is done on selected features in the dataset (which are inturn obtained by feature selection. 4 clusters are drawn up and K-means clustering is used to determine the clusters.

One such cluster is based on current interest rates. Each cluster has a different average interest rate.

Another such cluster is based on cities. For example, cluster 2 has many 'Mumbai' entries, with other cities such as Pune. Cluster '0' has cities like Delhi and Navi Mumbai. Cluster 1 has smaller towns like Rajkot and cluster 0 has cities like Nagpur.

Other clusters have been formed automatically based on the features chosen.

c) Any other business insights

It is seen that the maximum number of foreclosures are seen in Mumbai. Smaller towns are also susceptible to foreclosures. Higher the interest rate, more the likelihood for foreclosures. The number of loans given keeps increasing every year. Accuracy of prediction will increase as the NBFCs sees more foreclosures. However, in its best interests, predictions are done to reduce foreclosures.