Capstone Business Report: NBFC Foreclosure of Loan

Animesh Das | PGP - DSBA | April 11, 2021

1. 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.

Problem: 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 collateral for the loan.

Constraints: Foreclosure costs are high and cumbersome, so lenders want to find a suitable solution to avoid foreclosures.

Scope: Predict foreclosures based on available data and interpret the most important variables that will enable the NBFC to take required actions to retain the good customers.

Objectives: Empower the NBFC to separate the good loans from bad, thereby reduce potential losses, extend better services to the good customers, simplify future loan processing, exercise more caution prior to giving out a bad loan.

2. EDA and Business Implication

The given data contains details of loans authorized to specific customers under an agreement and scheme ID from August 29, 2010 through December 31, 2018 along with their foreclosure statuses among other details.

The data was visually inspected prior to loading onto Python for detailed analysis. Visual inspection gave out some understanding about the data particularly on how some of the variables were calculated columns, i.e., derived from other independent variables in the data itself. Columns like "DIFF_CURRENT_INTEREST_RATE_MAX_MIN", "LATEST_TRANSACTION_MONTH", "BALANCE TENURE" etc. are some columns that has a direct relationship and derives information from other columns in the dataset. Some missing values were also found to exist in the data.

***Table 1**: The NBFC Loan dataset displaying the top 10 rows with some of the columns*

AGREEMENTID	AUTHORIZATIONDATE	BALANCE_EXCESS	BALANCE_TENURE	CITY	...	SCHEMEID	FORECLOSURE
11220001	8/29/2010	0	0	MUMBAI	...	10901100	1
11220002	9/15/2010	0	99	MUMBAI	...	10901100	1
11220006	11/2/2010	0	231	MUMBAI	...	10901101	1
11220008	10/6/2010	0	0	THANE	...	10901100	1
11220010	10/26/2010	0	215	MUMBAI	...	10901101	1
11220011	10/28/2010	0	137	THANE	...	10901100	0
11220012	11/5/2010	0	294	MUMBAI	...	10901100	0
11220014	12/24/2010	0	276	MUMBAI	...	10901100	1
11220016	12/16/2010	0	145	THANE	...	10901101	1
11220017	11/25/2010	9988.42	291	THANE	...	10901116	0

Observations:

- There are 20012 rows and 53 columns in the dataset
- "Foreclosure" is the target variable that we will need to predict based on the data
    - 1 indicates a foreclosure
- The dataset contains a mix of data of types numeric, categorical as well as datetime variables

- Variable names in the data set are in proper naming convention and no renaming is necessary.
- The dataset is elaborate and contains several variables with information that can be broadly classified under customer information, EMI amount information, interest rate, dates of authorization, payments and interest start, tenor etc.

***Table 2**: Descriptive Statistics of the NBFC Loan data*

Variable	count	mean	std	min	25%	50%	75%	max
AGREEMENTID	20012	-	-	-	-	-	-	-
BALANCE_EXCESS	20012	78996	1348636	0	0	0	57	75556000
BALANCE_TENURE	20012	173	64	0	136	174	216	674
COMPLETED_TENURE	20012	17	16	0	6	12	25	98
CURRENT_INTEREST_RATE	20012	14.78	2.49	9.9	12.8	14.55	16.23	25.10
CURRENT_INTEREST_RATE_MAX	20012	14.9	2.48	10.43	13.11	14.67	16.54	37.46
CURRENT_INTEREST_RATE_MIN	20012	14.3	2.68	-5.06	12.42	13.73	16.17	24.03
CURRENT_INTEREST_RATE_CHANGES	20012	0.76	1.13	0	0	0	2	9.00
CURRENT_TENOR	20012	190	59	6	166	180	228	713
CUSTOMERID	19731	-	-	-	-	-	-	-
DIFF_AUTH_INT_DATE	20012	0	1	-17	0	0	0	70
DIFF_CURRENT_INTEREST_RATE_MAX_MIN	20012	0.6	0.97	0	0	0	1.19	24.35
DIFF_EMI_AMOUNT_MAX_MIN	19923	115209	967082	0	10207	19885	42466	84968250
DIFF_ORIGINAL_CURRENT_INTEREST_RATE	20012	-0.38	0.88	-7.18	-1.19	0	0	10.32
DIFF_ORIGINAL_CURRENT_TENOR	20012	-7	34	-461	-14	0	0	234
DPD	20012	8	66	0	0	0	0	2054
DUEDAY	20012	6	3	1	5	5	5	15
EMI_AMOUNT	20012	43610	113132	0	10685	18938	36424	4879479
EMI_DUEAMT	20012	1991553	6838394	0	204022	545065	1481417	354610400
EMI_OS_AMOUNT	20012	33297	656131	0	0	0	0	58995310
EMI_RECEIVED_AMT	20012	1958256	6762984	0	202094	537658	1456414	354610400
EXCESS_ADJUSTED_AMT	20012	359900	3923346	0	0	0	261	284164200
EXCESS_AVAILABLE	20012	438896	4169759	0	0	261	3105	284164200
FOIR	20012	27.96	3871.06	-170.33	0.41	0.52	0.68	547616.00
LAST_RECEIPT_AMOUNT	19765	80674	808403	1	11061	19642	38219	84968810
LATEST_TRANSACTION_MONTH	19937	11	3	1	12	12	12	12
LOAN_AMT	20012	5897355	12985661	37532	1558947	2684572	5233436	424566500
MAX_EMI_AMOUNT	19923	122254	970452	13	13318	23600	49361	84968810
MIN_EMI_AMOUNT	19923	7045	43425	0	118	133	3334	3156965
MONTHOPENING	20012	5447511	11838513	0	1483752	2503694	4791778	381836700
NET_DISBURSED_AMT	20012	5847666	12911932	37532	1544083	2640779	5186725	424566500
NET_LTV	20012	51.19	21.11	0.38	35.16	53.3	66.77	100.00
NET_RECEIVABLE	20012	-45439	1348502	-75345538	-18	0	0	38643500
NUM_EMI_CHANGES	20012	3	3	-1	2	2	4	33
NUM_LOW_FREQ_TRANSACTIONS	20012	3	3	0	1	2	3	30

Variable	count	Mean	std	min	25%	50%	75%	max
ORIGNAL_INTEREST_RATE	20012	14.4	2.6	9.65	12.49	13.73	16.17	27.78
ORIGNAL_TENOR	20012	183	45	14	180	180	228	300
OUTSTANDING_PRINCIPAL	20012	5212982	11521353	-1	1428919	2394655	4551204	381836700
PAID_INTEREST	20012	989055	3026053	0	125332	309725	795468	123036200
PAID_PRINCIPAL	20012	866764	34697581	0	23418	78787	291781	4885217000
PRE_EMI_DUEAMT	20012	57804	377665	0	4768	10696	31879	31775400
PRE_EMI_OS_AMOUNT	20012	259	10967	0	0	0	0	1074264
PRE_EMI_RECEIVED_AMT	20012	57545	376972	0	4755	10679	31805	31775400
SCHEMEID	19731	-	-	-	-	-	-	-
MOB	20012	19	17	0	7	13	26	98
FORECLOSURE	20012	0	0	0	0	0	0	1

Observations:

- As we can see from the “count” column from the descriptive table above, there is an indication of missing values in the dataset
- The different variables in the dataset contain values that are in varied ranges
- The minimum value in “DIFF_AUTH_INT_DATE” variable indicates that the interest start date is prior to authorization date
- The average loan amount is 5897355 while the median is 5233436
- Current interest rate minimum goes down as low as -5%
- The average original interest rate is 14.4% and maximum is seen to be close to 28%
- The highest original tenor is of 300 periods while the lowest is just 14, with an average of 45
- The statistical values arising from the IDs have been removed as they are not actual numbers
- Below we can see in the variable info table above, most of the variables are of numeric nature, 3 are timestamps and 4 are of string type.
- A lot of null values can be noticed in the “NPA_IN_LAST_MONTH” and “NPA_IN_CURRENT_MONTH” column

A check has also been performed to detect duplicate records. There were no duplicate records found in the dataset.

***Table 3**: Variable info of NBFC Loan data*

#	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
4	CITY	20012 non-null	object
5	COMPLETED_TENURE	20012 non-null	int64
6	CURRENT_INTEREST_RATE	20012 non-null	float64
7	CURRENT_INTEREST_RATE_MAX	20012 non-null	float64
8	CURRENT_INTEREST_RATE_MIN	20012 non-null	float64
9	CURRENT_INTEREST_RATE_CHANGES	20012 non-null	int64
10	CURRENT_TENOR	20012 non-null	int64
11	CUSTOMERID	19731 non-null	float64
12	DIFF_AUTH_INT_DATE	20012 non-null	int64
13	DIFF_CURRENT_INTEREST_RATE_MAX_MIN	20012 non-null	float64
14	DIFF_EMI_AMOUNT_MAX_MIN	19923 non-null	float64
15	DIFF_ORIGINAL_CURRENT_INTEREST_RATE	20012 non-null	float64
16	DIFF_ORIGINAL_CURRENT_TENOR	20012 non-null	int64
17	DPD	20012 non-null	int64
18	DUEDAY	20012 non-null	int64
19	EMI_AMOUNT	20012 non-null	float64
20	EMI_DUEAMT	20012 non-null	float64
21	EMI_OS_AMOUNT	20012 non-null	float64
22	EMI_RECEIVED_AMT	20012 non-null	float64
23	EXCESS_ADJUSTED_AMT	20012 non-null	float64
24	EXCESS_AVAILABLE	20012 non-null	float64
25	FOIR	20012 non-null	float64
26	INTEREST_START_DATE	20012 non-null	datetime64[ns]
27	LAST_RECEIPT_AMOUNT	19765 non-null	float64
28	LAST_RECEIPT_DATE	19937 non-null	datetime64[ns]
29	LATEST_TRANSACTION_MONTH	19937 non-null	float64
30	LOAN_AMT	20012 non-null	float64
31	MAX_EMI_AMOUNT	19923 non-null	float64
32	MIN_EMI_AMOUNT	19923 non-null	float64
33	MONTHOPENING	20012 non-null	float64
34	NET_DISBURSED_AMT	20012 non-null	float64
35	NET_LTV	20012 non-null	float64
36	NET_RECEIVABLE	20012 non-null	float64
37	NUM_EMI_CHANGES	20012 non-null	int64
38	NUM_LOW_FREQ_TRANSACTIONS	20012 non-null	int64
39	ORIGNAL_INTEREST_RATE	20012 non-null	float64
40	ORIGNAL_TENOR	20012 non-null	int64
41	OUTSTANDING_PRINCIPAL	20012 non-null	float64
42	PAID_INTEREST	20012 non-null	float64
43	PAID_PRINCIPAL	20012 non-null	float64
44	PRE_EMI_DUEAMT	20012 non-null	float64
45	PRE_EMI_OS_AMOUNT	20012 non-null	float64
46	PRE_EMI_RECEIVED_AMT	20012 non-null	float64
47	PRODUCT	20012 non-null	object
48	SCHEMEID	19731 non-null	float64
49	NPA_IN_LAST_MONTH	119 non-null	object
50	NPA_IN_CURRENT_MONTH	119 non-null	object
51	MOB	20012 non-null	int64
52	FORECLOSURE	20012 non-null	int64

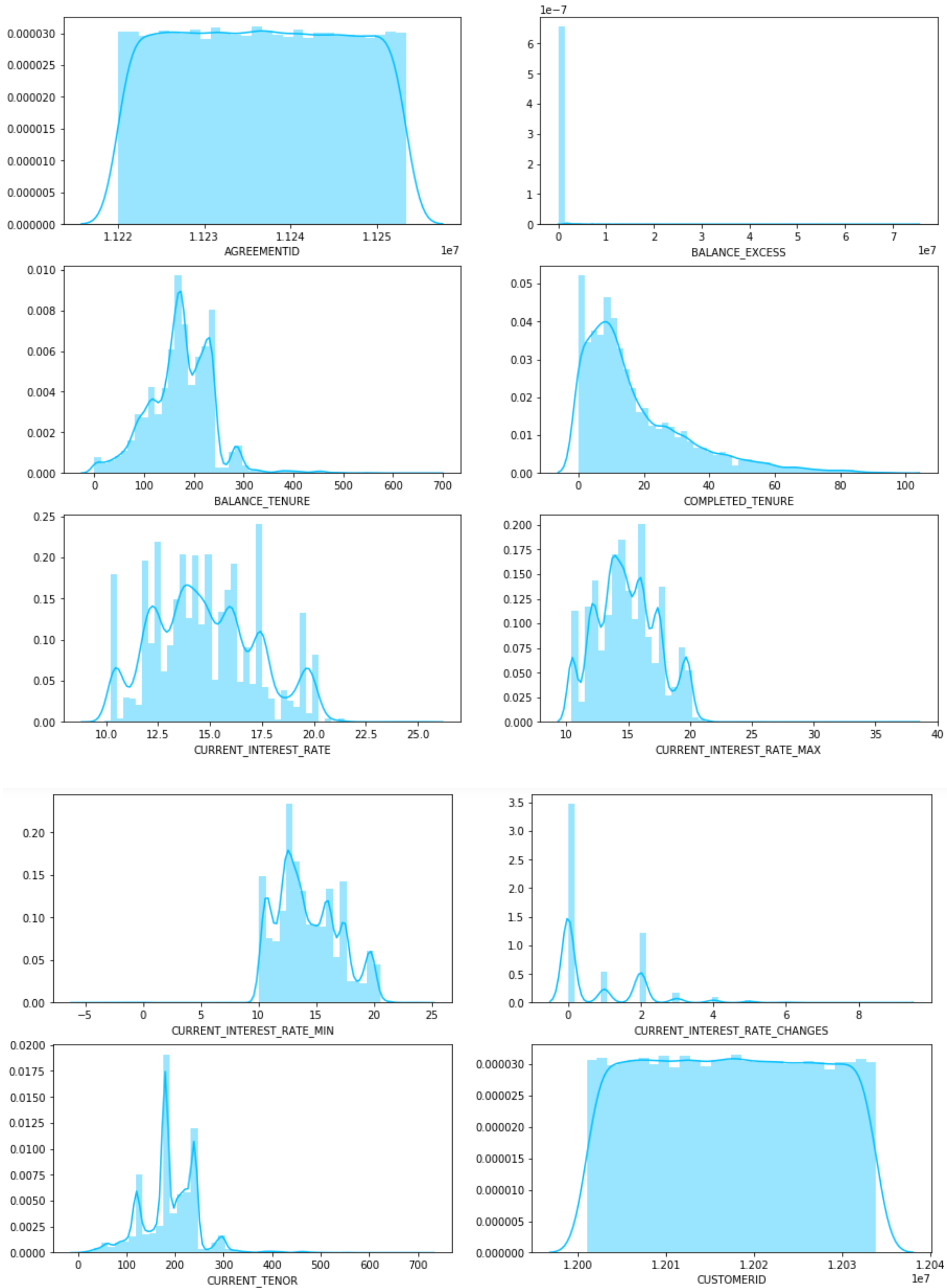
dtypes: datetime64[ns](3), float64(32), int64(14), object(4)

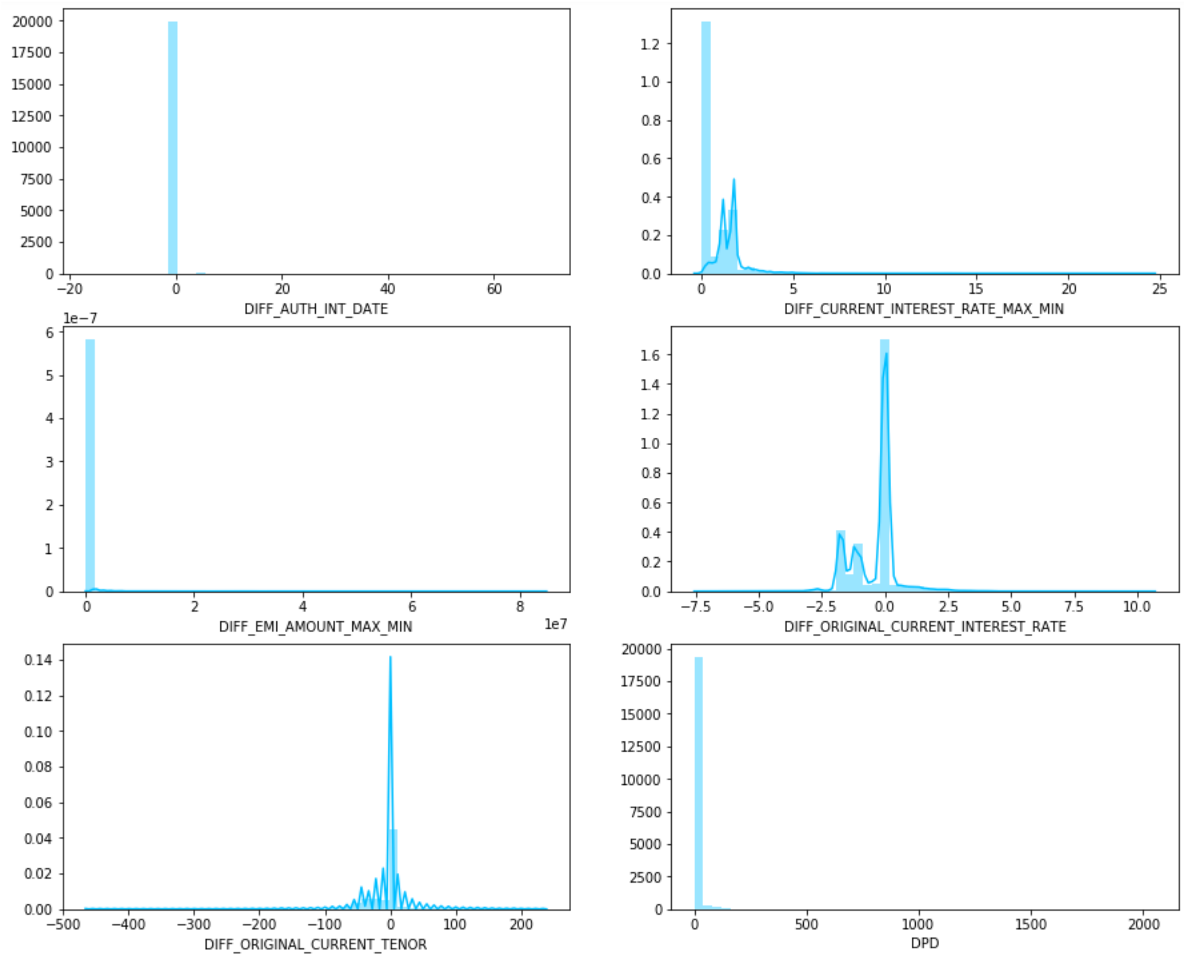
Univariate/ bivariate/ multivariate analysis was done in order to understand relationship b/w variables and uncover hidden information that would be help the business in understanding more about the anomalies and the facts present in the data.

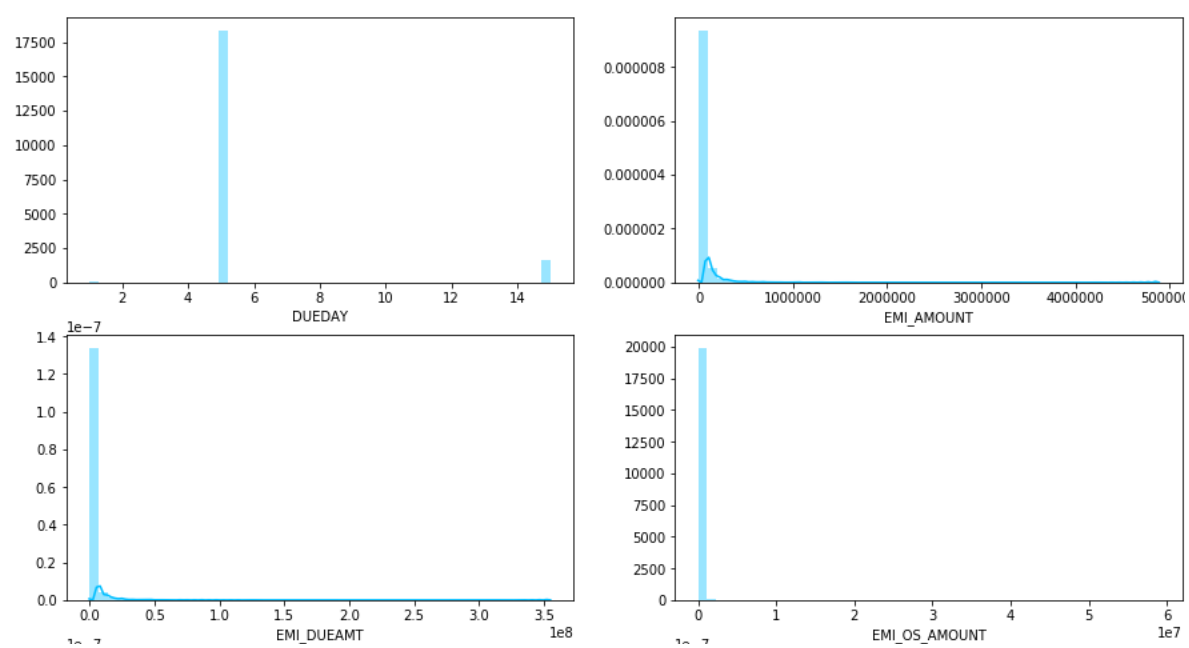
Univariate Analysis

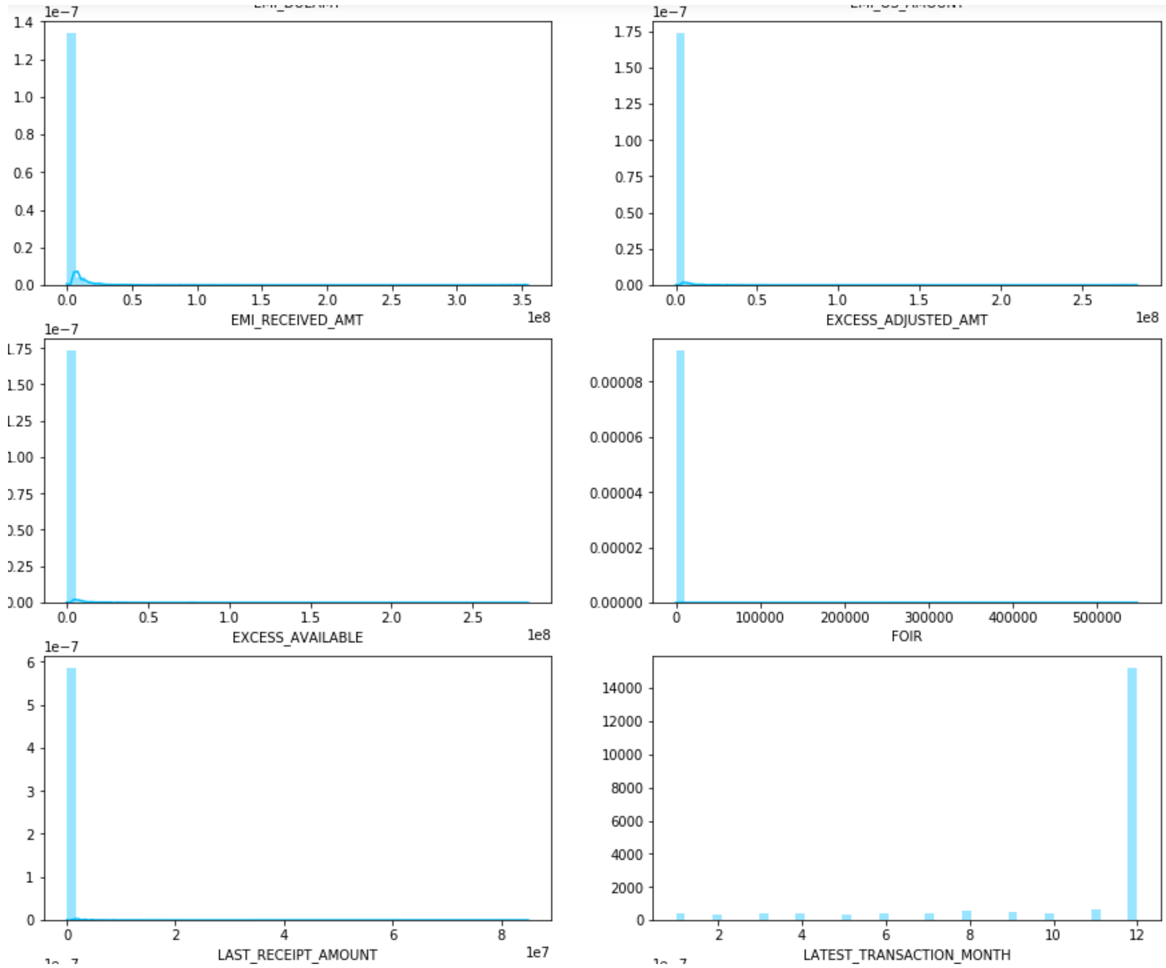
Below, a **distribution plot** has been used to understand about the distribution and skewness of the numeric variables visually and then statistically.

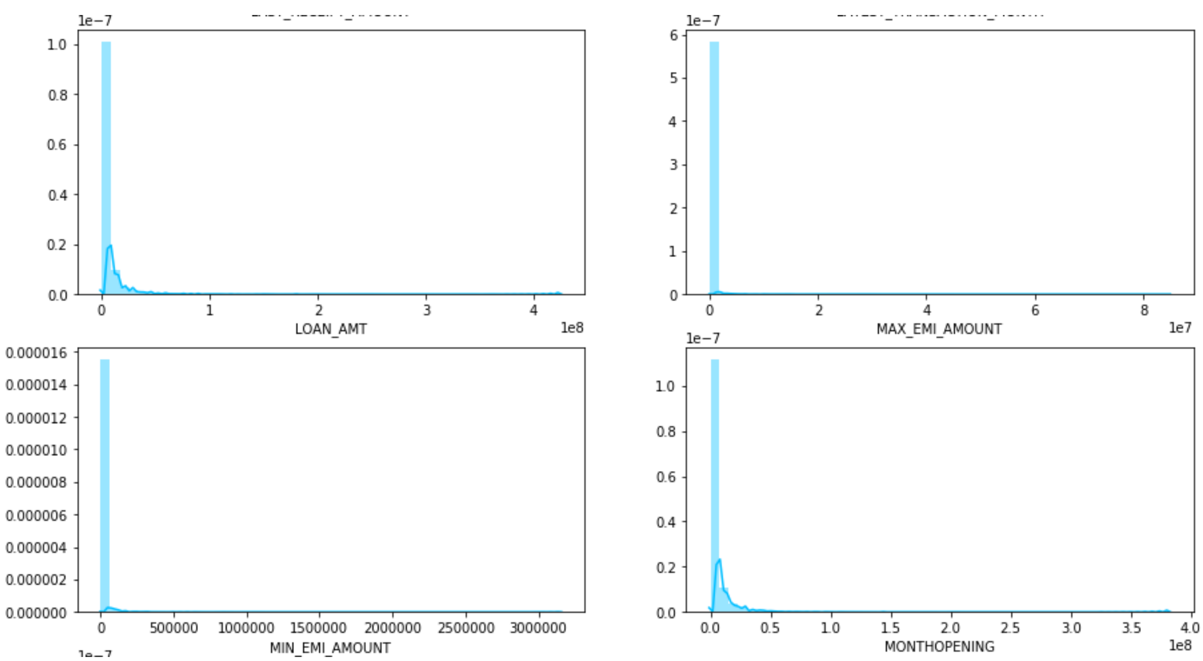
***Fig 1**: Distribution plot of the numeric variables*

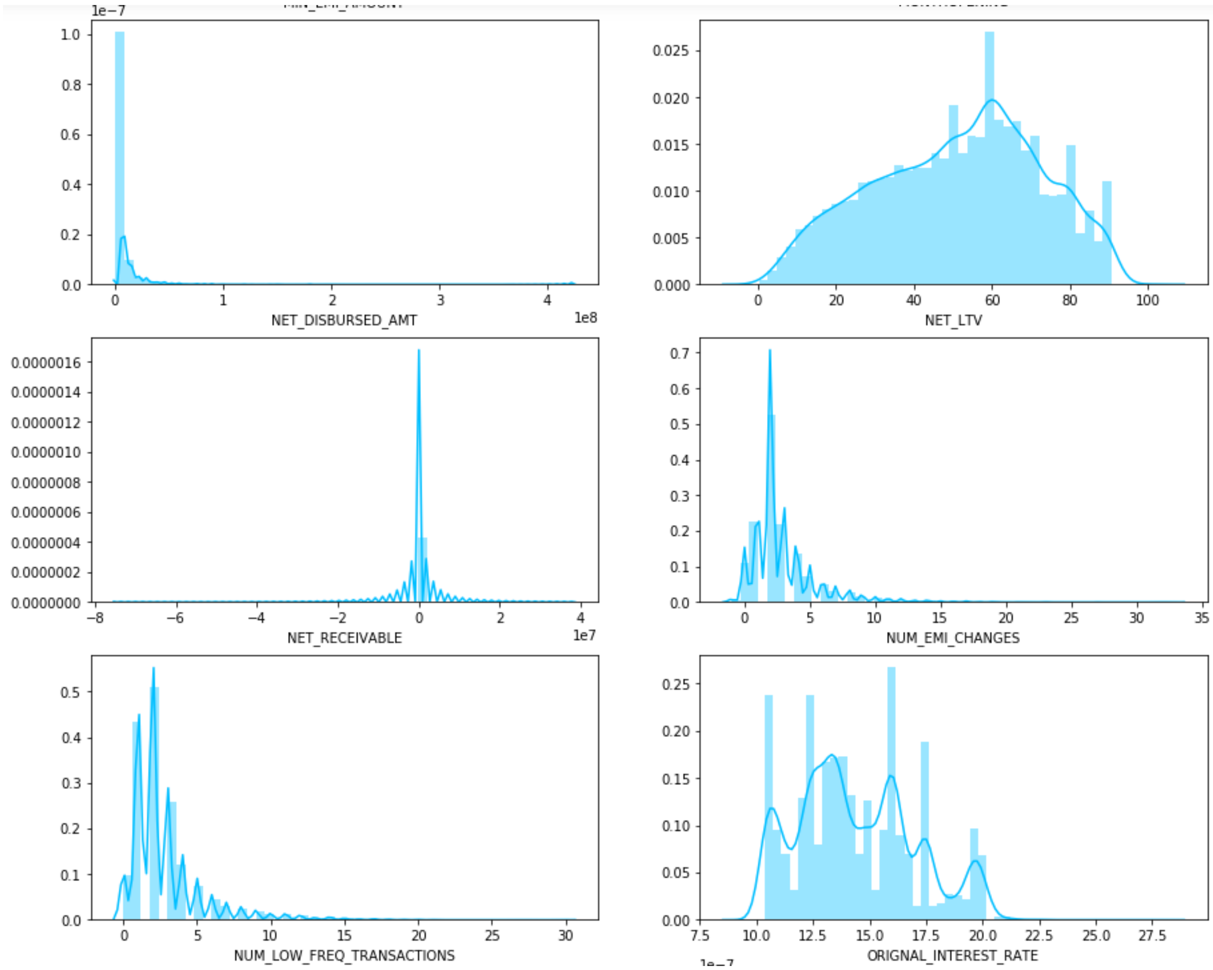


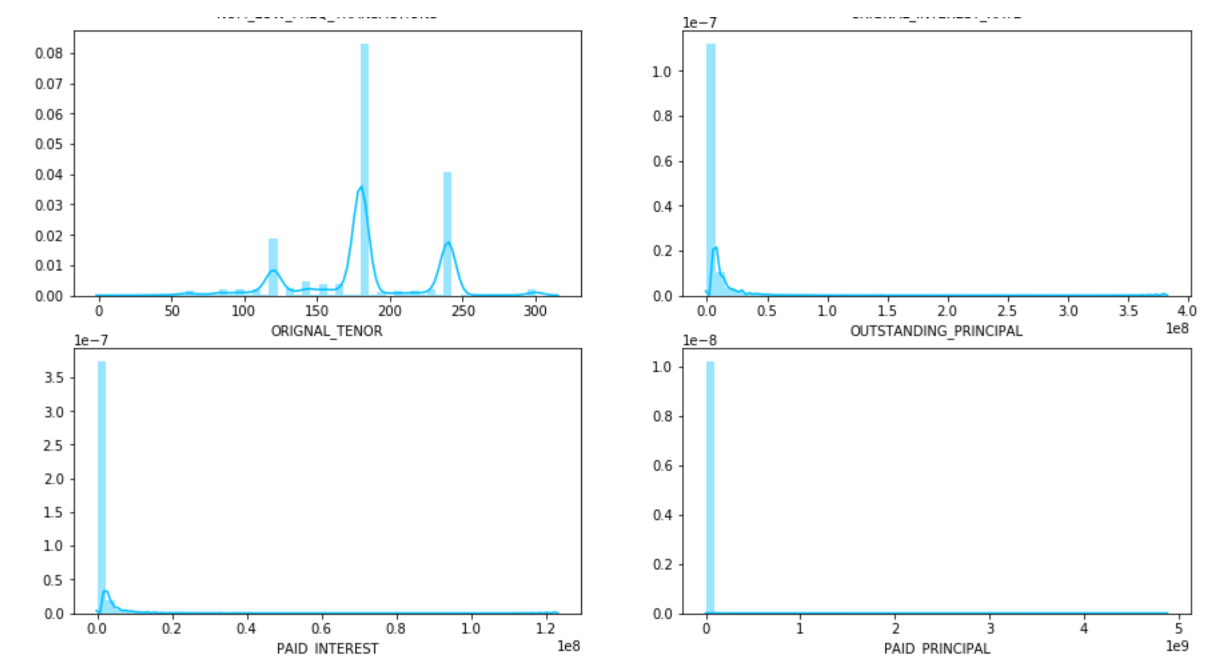

DIFF_AUTH_INT_DATE
DIFF_CURRENT_INTEREST_RATE_MAX_MIN
1e−7
DIFF_EMI_AMOUNT_MAX_MIN
1e7
DIFF_ORIGINAL_CURRENT_INTEREST_RATE
DIFF_ORIGINAL_CURRENT_TENOR
DPD

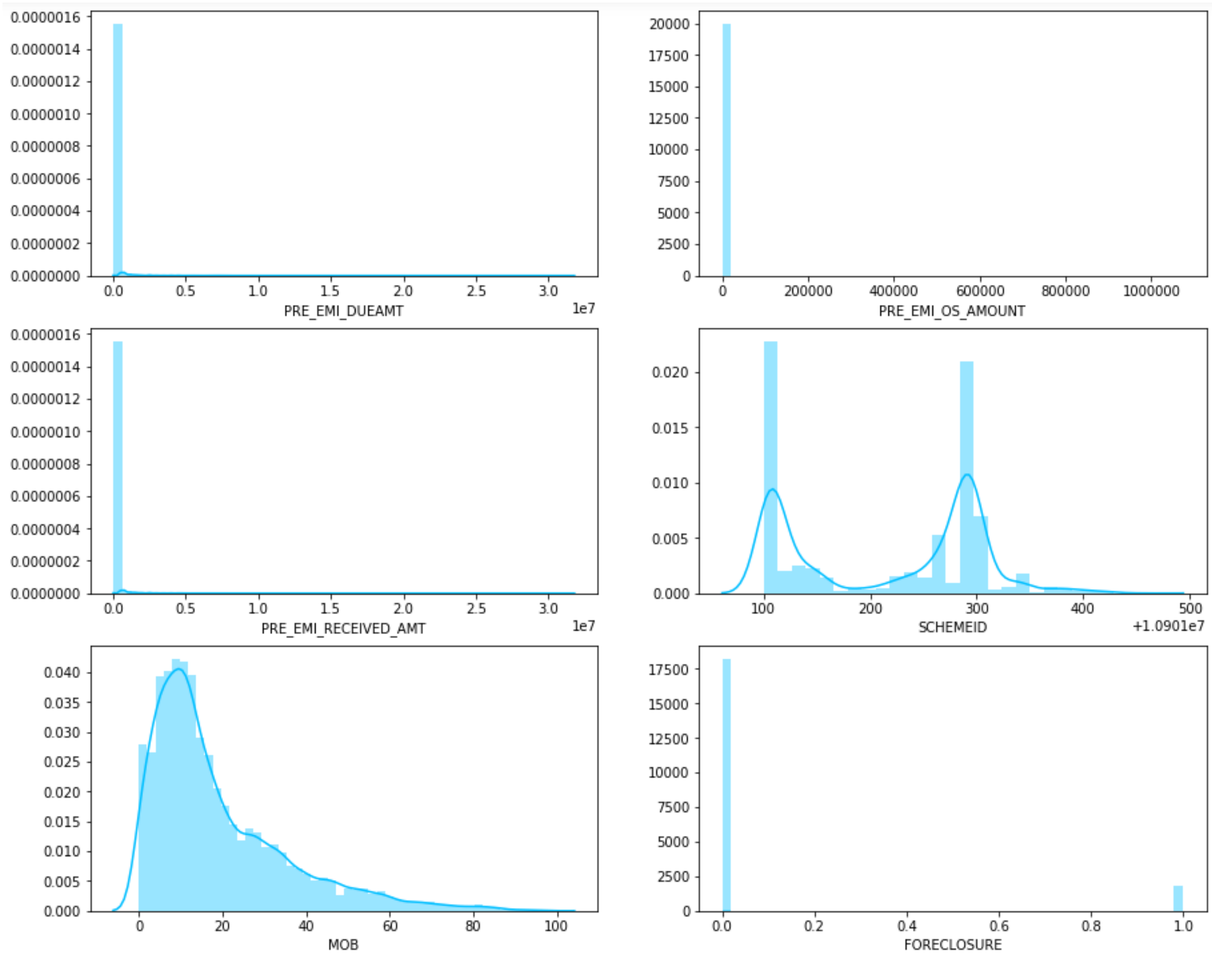

DUEDAY
EMI_AMOUNT
1e−7
EMI_DUEAMT
1e8
EMI_OS_AMOUNT
1e7


1e−7
EMI_RECEIVED_AMT
1e8
1e−7
EXCESS_ADJUSTED_AMT
1e8
1e−7
EXCESS_AVAILABLE
1e8
FOIR
1e−7
LAST_RECEIPT_AMOUNT
1e7
LATEST_TRANSACTION_MONTH


1e−7
LOAN_AMT
1e8
1e−7
MAX_EMI_AMOUNT
1e7
MIN_EMI_AMOUNT
1e−7
MONTHOPENING
1e8


1e−7
NET_DISBURSED_AMT
1e8
NET_LTV
NET_RECEIVABLE
1e7
NUM_EMI_CHANGES
NUM_LOW_FREQ_TRANSACTIONS
ORIGNAL_INTEREST_RATE


1e−7
ORIGNAL_TENOR
OUTSTANDING_PRINCIPAL
1e8
1e−7
PAID_INTEREST
1e8
1e−8
PAID_PRINCIPAL
1e9



<u>Observations:</u>

- None of the variables seem to be normally distributed
- Several variables like "PAID_PRINCIPAL", "EMI_AMOUNT", "FOIR" etc. seem to be severely right skewed.
- Several variables like "NET_RECEIVABLE" and "LATEST_TRANSACTION_MONTH" are moderately left skewed.
- "ORIGNAL_INTEREST_RATE", "CURRENT_INTEREST_RATE" "CURRENT_TENOR" etc. have a multimodal distribution
- Most of the variables with differences such as "DIFF_ORIGINAL_CURRENT_INTEREST_RATE", "DIFF_CURRENT_INTEREST_RATE_MAX_MIN", "DIFF_ORIGINAL_CURRENT_TENOR" etc. have the greatest mode at/around 0.

***Table 4**: Skew measures in the variables*

Variable	Skew
FOIR	141.46
PAID_PRINCIPAL	139.43
DIFF_AUTH_INT_DATE	94.53
PRE_EMI_OS_AMOUNT	74.46
LAST_RECEIPT_AMOUNT	67.9
EMI_OS_AMOUNT	55.96
DIFF_EMI_AMOUNT_MAX_MIN	46.4
MAX_EMI_AMOUNT	45.97
PRE_EMI_RECEIVED_AMT	43.62
PRE_EMI_DUEAMT	43.42
EXCESS_ADJUSTED_AMT	41.57
EXCESS_AVAILABLE	36.12
BALANCE_EXCESS	34.31
MIN_EMI_AMOUNT	33.19
EMI_RECEIVED_AMT	16.18
EMI_DUEAMT	15.83
DPD	15.5
EMI_AMOUNT	13.34
PAID_INTEREST	13.21
OUTSTANDING_PRINCIPAL	11.67
MONTHOPENING	11.2
NET_DISBURSED_AMT	10.97
LOAN_AMT	10.86

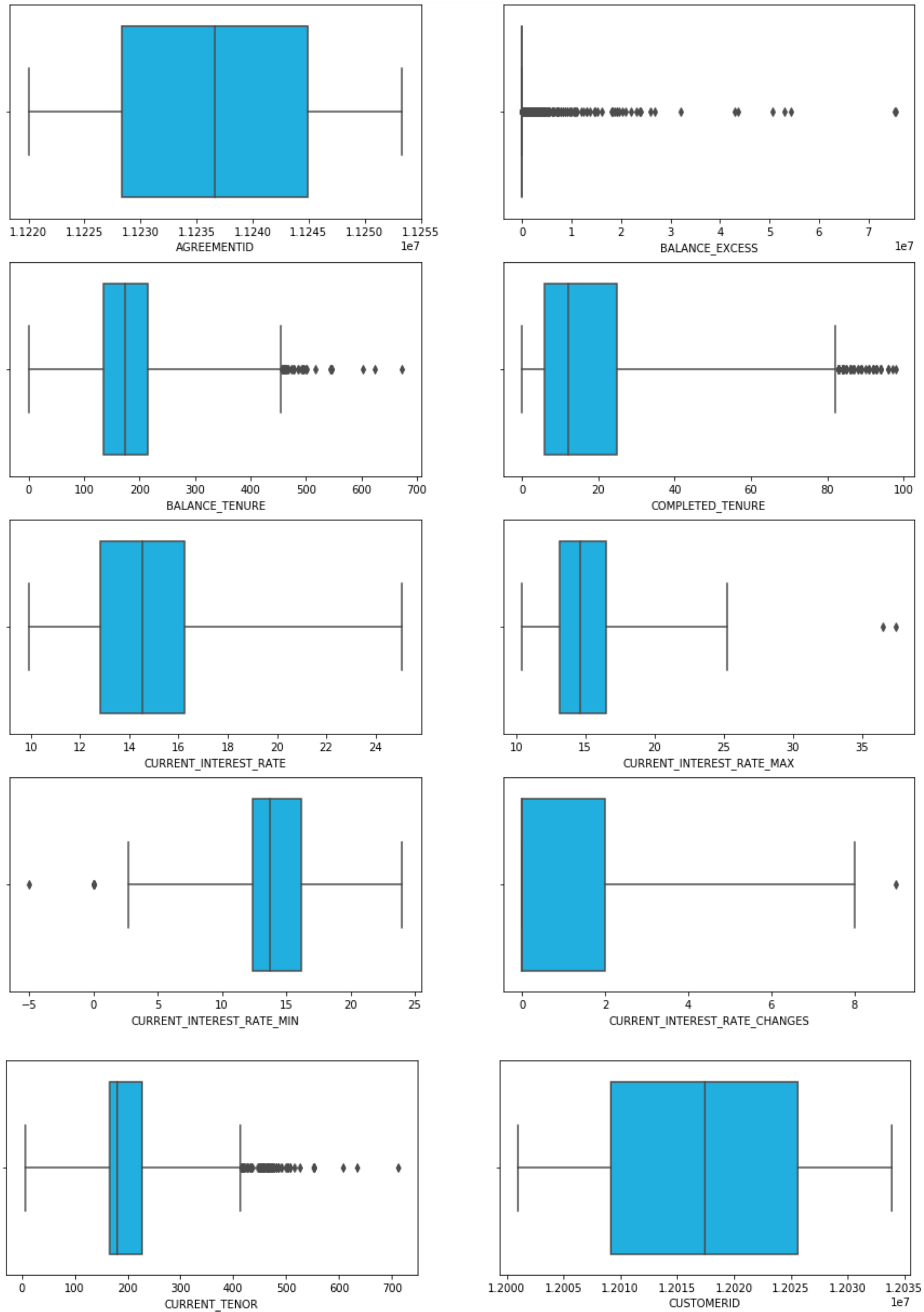
Variable	Skew
DIFF_CURRENT_INTEREST_RATE_MAX_MIN	4.18
DUEDAY	3.05
FORECLOSURE	2.87
NUM_EMI_CHANGES	2.63
NUM_LOW_FREQ_TRANSACTIONS	2.58
CURRENT_INTEREST_RATE_CHANGES	1.56
COMPLETED_TENURE	1.53
MOB	1.5
CURRENT_TENOR	0.49
ORIGNAL_INTEREST_RATE	0.41
CURRENT_INTEREST_RATE_MIN	0.39
BALANCE_TENURE	0.31
CURRENT_INTEREST_RATE	0.29
CURRENT_INTEREST_RATE_MAX	0.29
DIFF_ORIGINAL_CURRENT_INTEREST_RATE	0.28
AGREEMENTID	0.01
CUSTOMERID	0.01
SCHEMEID	-0.13
DIFF_ORIGINAL_CURRENT_TENOR	-0.15
ORIGNAL_TENOR	-0.21
NET_LTV	-0.21
LATEST_TRANSACTION_MONTH	-2.16
NET_RECEIVABLE	-28.77

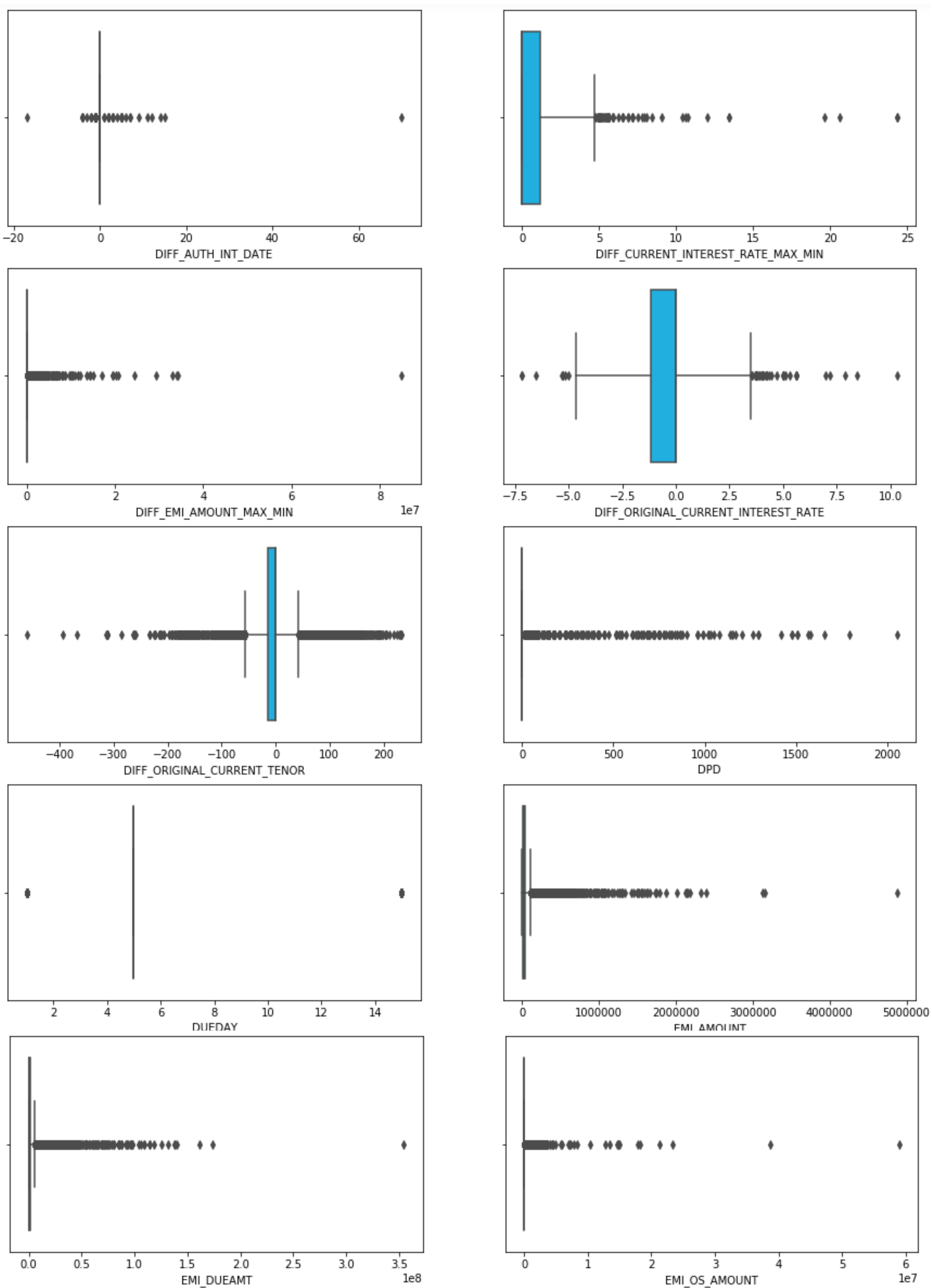
Next, **Boxplots** have been used to understand further about the distribution and skewness of the numeric data as well as detect the presence of outliers visually. Here, the whiskers are taken at 3x IQR indicating extreme upper and lower bounds.

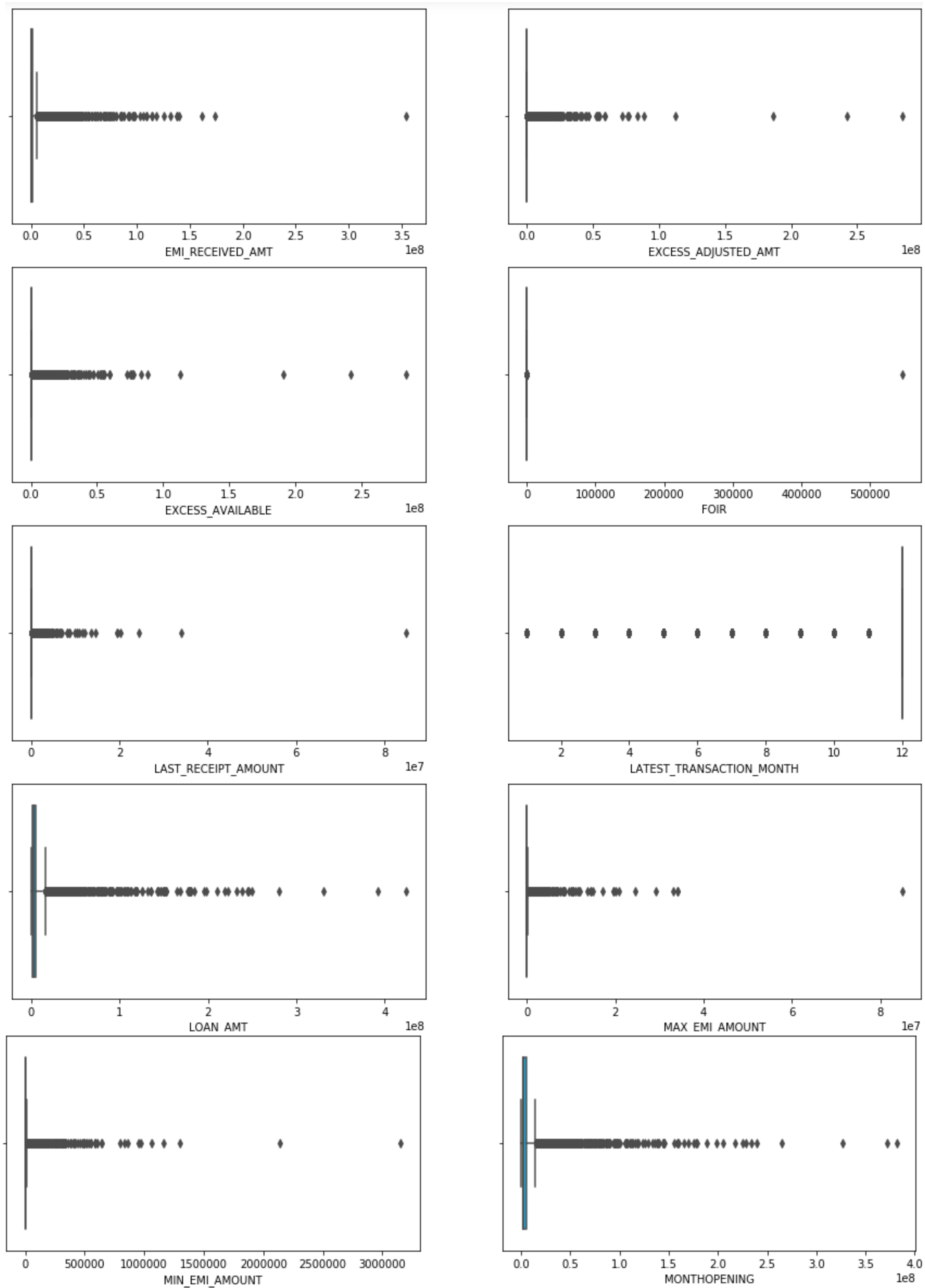
Observations:

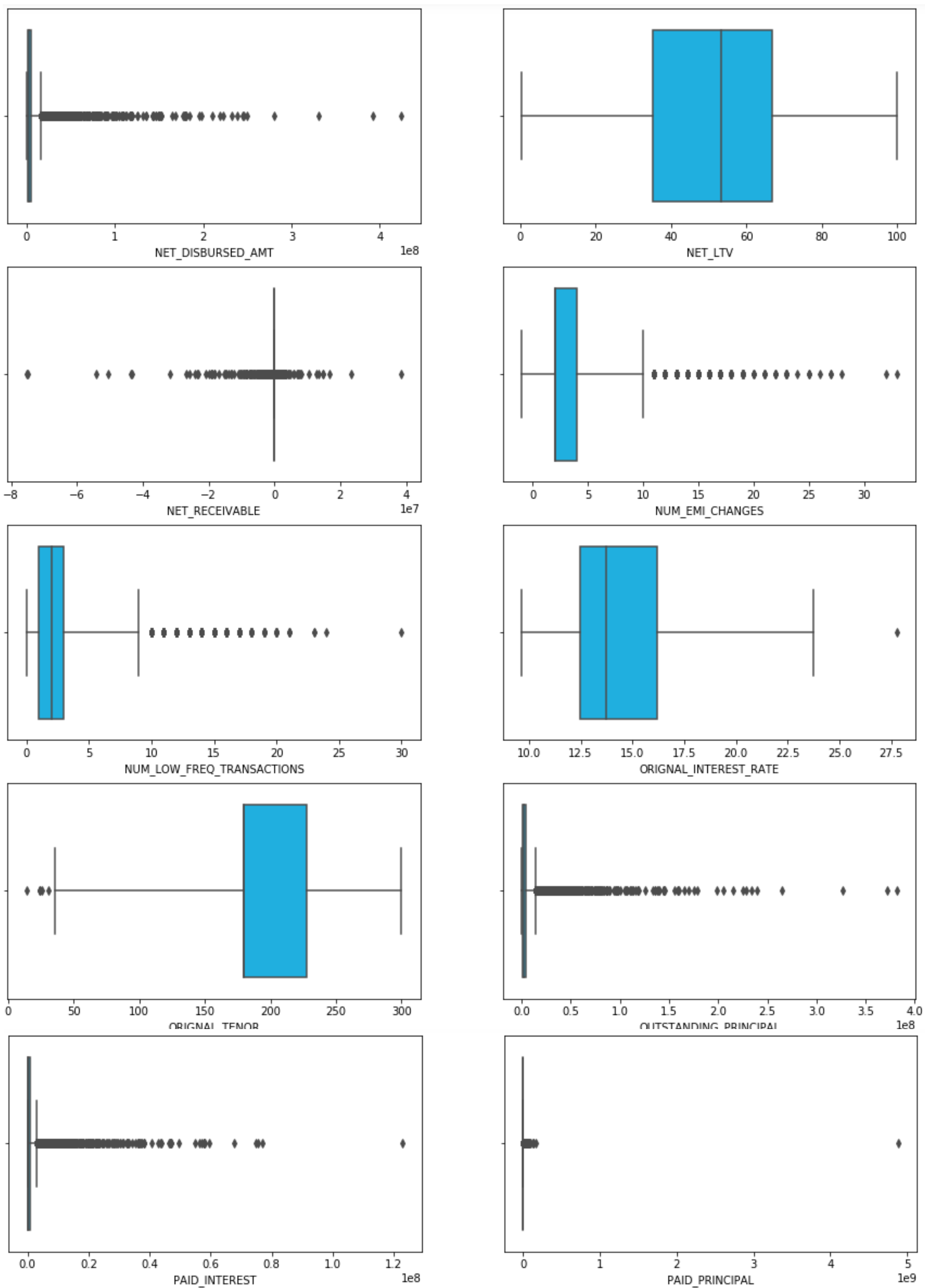
- From the boxplots below, we can see that several variables contain outliers.
- Over 10% of the data in variables such as net receivables, balance excess, max EMI amount etc. are outliers
- Several variables such as balance excess, max EMI amount, current interest rate max, etc. have data points that are much farther away from their natural clusters. Such variables will have to be studied further.

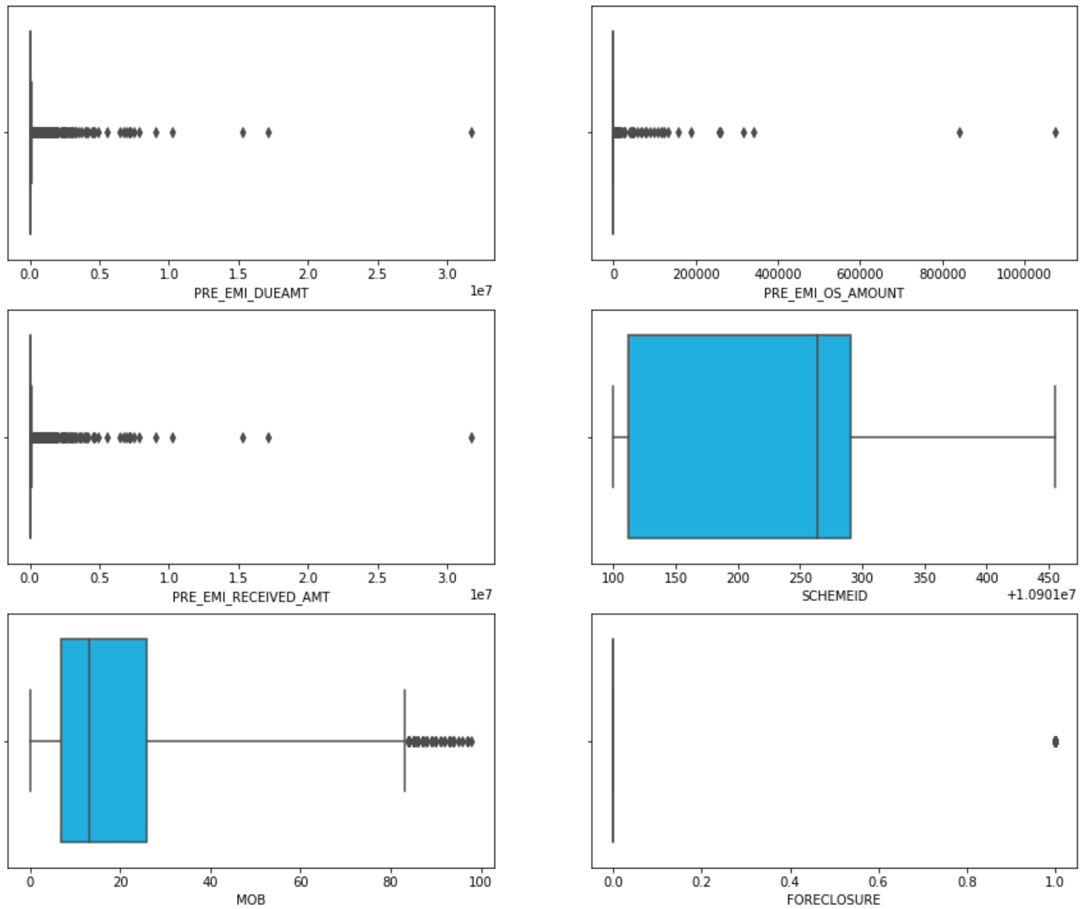
***Fig 2**: Boxplot of the numeric variables*




DIFF_AUTH_INT_DATE
−20 0 20 40 60
DIFF_CURRENT_INTEREST_RATE_MAX_MIN
0 5 10 15 20 25
DIFF_EMI_AMOUNT_MAX_MIN
0 2 4 6 8
1e7
DIFF_ORIGINAL_CURRENT_INTEREST_RATE
−7.5 −5.0 −2.5 0.0 2.5 5.0 7.5 10.0
DIFF_ORIGINAL_CURRENT_TENOR
−400 −300 −200 −100 0 100 200
DPD
0 500 1000 1500 2000
DUEDAY
2 4 6 8 10 12 14
EMI_AMOUNT
0 1000000 2000000 3000000 4000000 5000000
EMI_DUEAMT
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
1e8
EMI_OS_AMOUNT
0 1 2 3 4 5 6
1e7

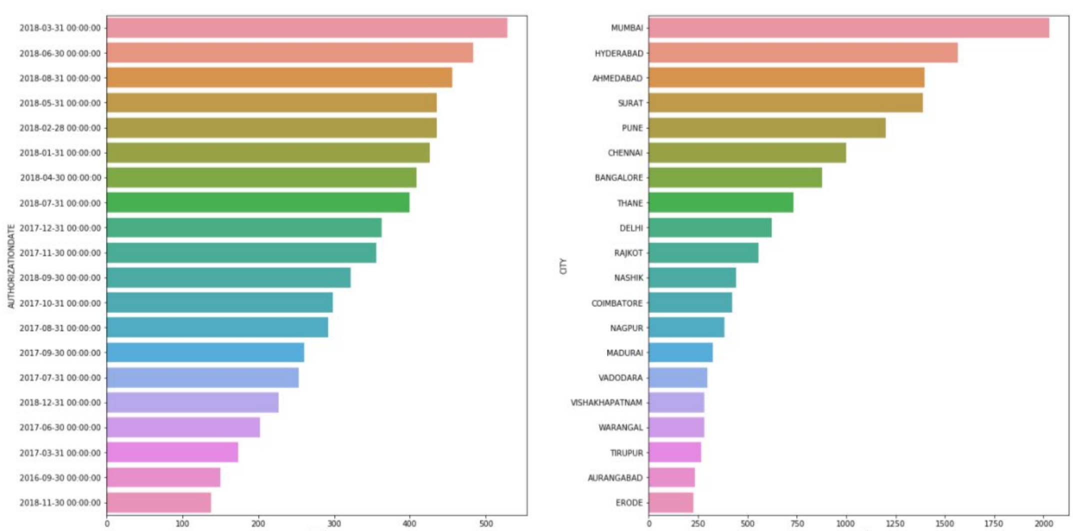

EMI_RECEIVED_AMT
1e8
EXCESS_ADJUSTED_AMT
1e8
EXCESS_AVAILABLE
1e8
FOIR
LAST_RECEIPT_AMOUNT
1e7
LATEST_TRANSACTION_MONTH
LOAN_AMT
1e8
MAX_EMI_AMOUNT
1e7
MIN_EMI_AMOUNT
MONTHOPENING
1e8

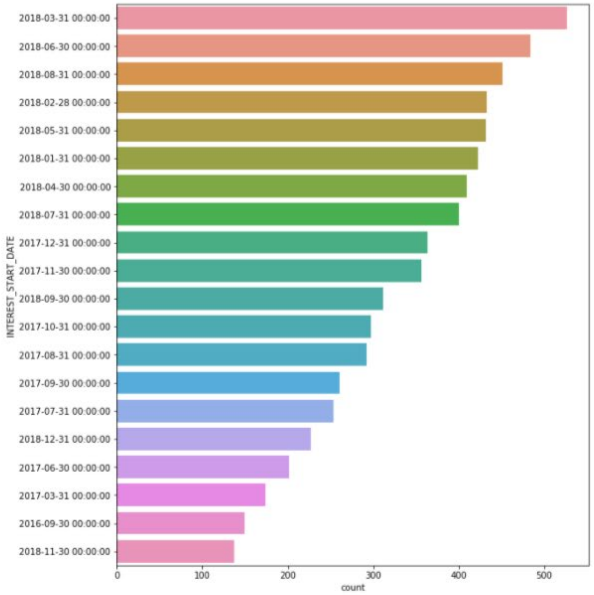

NET_DISBURSED_AMT
1e8
NET_LTV
NET_RECEIVABLE
1e7
NUM_EMI_CHANGES
NUM_LOW_FREQ_TRANSACTIONS
ORIGNAL_INTEREST_RATE
ORIGNAL_TENOR
OUTSTANDING_PRINCIPAL
1e8
PAID_INTEREST
1e8
PAID_PRINCIPAL
1e9

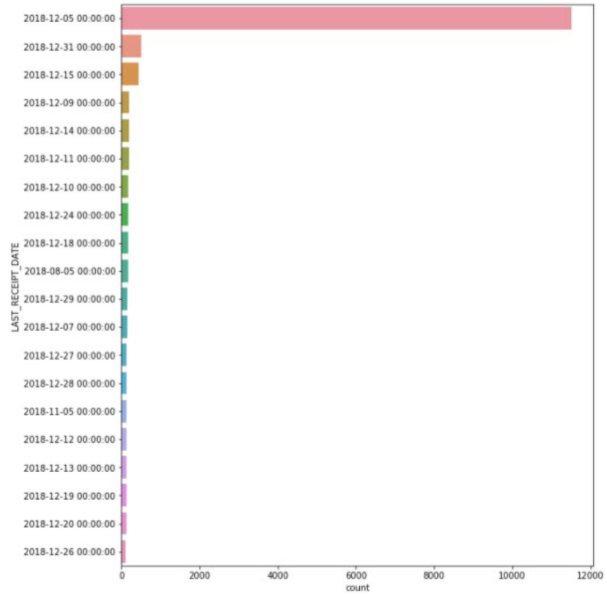


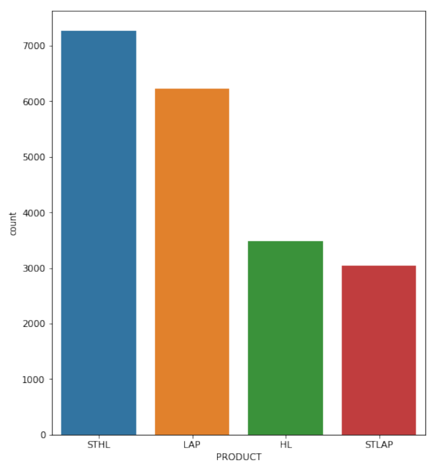
Lastly in univariate analysis, **count plots** (below) were used to understand about the distribution of the categorical variables.

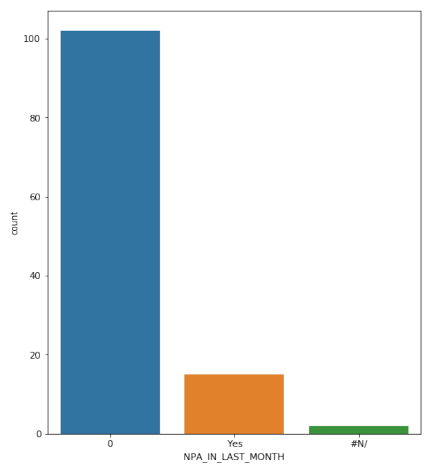
***Fig 3**: Count plot for the datetime and categorical variables*

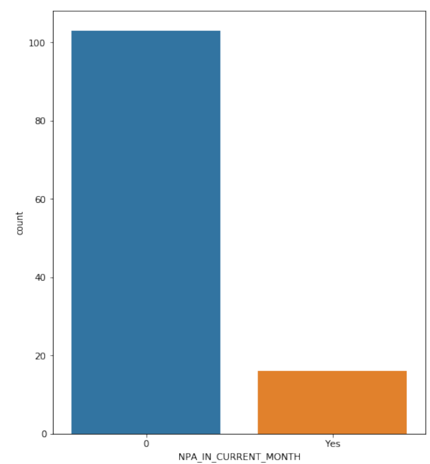












Observations:

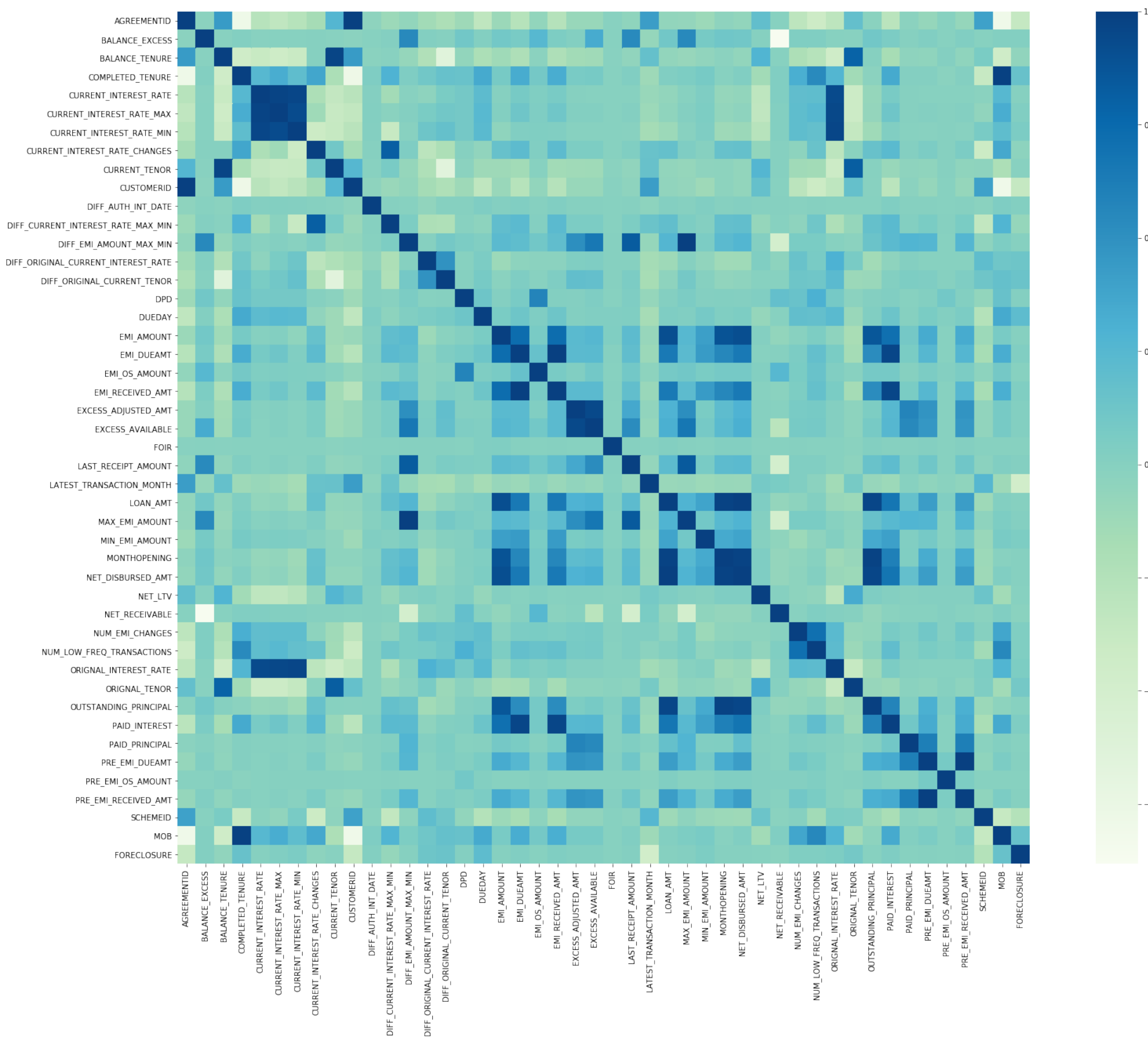
For the dates and the city variables, only the top 20 elements have been visualized here to show the greatest amount of information.

- The end dates in the months seem to have the highest number of loans authorized
- The highest number of loans were authorized on March 31, 2018 followed by June 30, 2018.
- Most of the loans originate from Mumbai, followed by Hyderabad and Ahmedabad
- Interest start date is very closely associated with the authorization date and tends to follow a similar pattern stated above
- December 5, 2018 has seen an unusual number of payments received
- STHL is the most popular loan product with over 7000 loans under this label, followed by LAP. STLAP is the least popular in the lot
- 0 NPA in last and current month with a count of little over 100. However, there are a lot of missing values in these two variables

Bivariate Analysis

First, we will analyze a **correlation heatmap** to interpret the direction and strength of relationship between any two numeric variables.

***Fig 4**: Correlation Heatmap*



Observations:

- Since a lot of the variables are derived, a strong relationship is seen between the original and the derived variables, such as Net Receivable with Balance Excess, Diff_original_current_tenor with balance tenor etc.
- Completed tenure has a very strong correlation with MOB

Additionally, further analysis was done between the target and independent variables to uncover hidden information that would be help the business in understanding more about the anomalies and the facts present in the data.

Insights:

- Loan product LAP has only 2 foreclosures out of 6000+ records
- Product STLAP has 0 foreclosures
- Just 2 scheme IDs (10901104 & 10901112) out of 236 accounts for 69% of foreclosures
- Customers whose due day is on 5th of the month foreclose the most out of all
- Foreclosures are high for customers whose original interest rate is around 17 and 13%

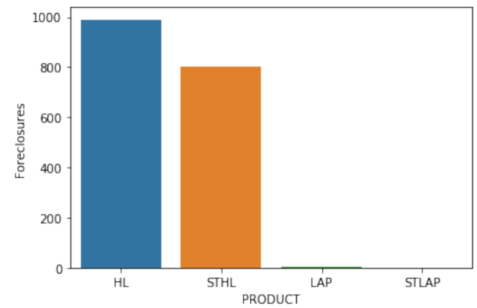


***Fig 5**: Foreclosures in Product*

3. Data Cleaning and Pre-processing

Data Cleaning

7 same city names were identified to have been misspelt or were spelt in colloquial way. E.g., Thiruvallur and Tiruvallur; Vijaywada and Vijayawada etc. Such inconsistencies were fixed by keeping one of the spellings

Outlier treatment

Every variable was given individual attention prior deciding to treat outliers in them. In this case, a lot of variables are dependent on each other and using capping/flooring method or mean/median/mode imputation etc. of a single datapoint might hamper the integrity and the functional relationship between variables. Thus, through a detailed study of the visuals from the boxplots and records in the dataset and exercising judgement, only the extreme data points/records were eliminated/dropped. Here, extreme datapoints refer to datapoints which are much farther away from the general cluster of points in a variable. These are unnatural and may hamper the generalization capability of the algorithms. In total, 50 records were dropped which is just about 0.25% of the original records.

***Table 5**: Outlier count prior treatment*

Variable	No. of Outliers	No. of Outliers / Total Records
NET_RECEIVABLE	5860	29.28%
LATEST_TRANSACTION_MONTH	4734	23.66%
BALANCE_EXCESS	4462	22.30%
EXCESS_ADJUSTED_AMT	4419	22.08%
EXCESS_AVAILABLE	4392	21.95%
DIFF_EMI_AMOUNT_MAX_MIN	2081	10.40%
MAX_EMI_AMOUNT	2056	10.27%
MIN_EMI_AMOUNT	1809	9.04%
FORECLOSURE	1795	8.97%
PRE_EMI_RECEIVED_AMT	1760	8.79%

Variable	No. of Outliers	No. of Outliers / Total Records
PRE_EMI_DUEAMT	1760	8.79%
PAID_PRINCIPAL	1705	8.52%
DUEDAY	1669	8.34%
LAST_RECEIPT_AMOUNT	1601	8.00%
OUTSTANDING_PRINCIPAL	1453	7.26%
EMI_RECEIVED_AMT	1444	7.22%
EMI_DUEAMT	1439	7.19%
DIFF_ORIGINAL_CURRENT_TENOR	1430	7.15%
EMI_AMOUNT	1425	7.12%
MONTHOPENING	1413	7.06%
PAID_INTEREST	1349	6.74%
NET_DISBURSED_AMT	1315	6.57%
LOAN_AMT	1314	6.57%
EMI_OS_AMOUNT	1278	6.39%
DPD	1242	6.21%
NUM_LOW_FREQ_TRANSACTIONS	657	3.28%
NUM_EMI_CHANGES	460	2.30%
FOIR	306	1.53%
CURRENT_TENOR	104	0.52%
DIFF_AUTH_INT_DATE	86	0.43%
MOB	79	0.39%
DIFF_CURRENT_INTEREST_RATE_MAX_MIN	77	0.38%
COMPLETED_TENURE	75	0.37%
PRE_EMI_OS_AMOUNT	53	0.26%
DIFF_ORIGINAL_CURRENT_INTEREST_RATE	52	0.26%
BALANCE_TENURE	43	0.21%
CURRENT_INTEREST_RATE_MIN	6	0.03%
ORIGNAL_TENOR	6	0.03%
CURRENT_INTEREST_RATE_MAX	2	0.01%
ORIGNAL_INTEREST_RATE	1	0.01%
CURRENT_INTEREST_RATE_CHANGES	1	0.01%

Missing Value treatment

Missing value treatment was done after outlier treatment in order to keep the skewness in central tendency measures in check in case of imputation using any of these methods, particularly in case of mean imputation.

- Over 99% of records in the NPA variables contain null values as we have seen earlier. These two variables have been dropped.
- Mode imputation was performed for Scheme ID on the basis of the Product code. For each product, the count of scheme IDs were determined and mode imputation was performed on that basis

- Missing values in max and min EMI amount were imputed with the EMI Amount with the assumption that the customer has paid the actual EMI amount with no change
- Missing values in last receipt amount was fixed imputing the max EMI amount. This was done because last receipt amount had the highest correlation with max EMI amount
- For missing values in last receipt date, an average difference in days was determined between the last receipt date and the interest date and that number 557(days) was added to the interest state date and imputed
- Missing values in latest transaction month were fixed extracting the month from the last receipt date using the datetime month function

***Table 6**: Missing value count*

Variable	No. of NAs	No. of NAs / Total Records
NPA_IN_CURRENT_MONTH	19893	99.41%
NPA_IN_LAST_MONTH	19893	99.41%
CUSTOMERID	281	1.40%
SCHEMEID	281	1.40%
LAST_RECEIPT_AMOUNT	247	1.23%
MAX_EMI_AMOUNT	89	0.44%
MIN_EMI_AMOUNT	89	0.44%
DIFF_EMI_AMOUNT_MAX_MIN	89	0.44%
LAST_RECEIPT_DATE	75	0.37%
LATEST_TRANSACTION_MONTH	75	0.37%

Variable transformation

Categorical variables such as city, product, authorization, last receipt and interest start dates were encoded to convert to numeric prior to feeding to the ML algorithms as the algorithms in use here only take in numeric inputs.

Variables removed

The following variables were removed from the dataset as the two NPA variables have more than 99% of null values in them and the other two are unique IDs that won't add value in the prediction process:

- 'NPA_IN_CURRENT_MONTH',
- 'NPA_IN_LAST_MONTH',
- 'AGREEMENTID',
- 'CUSTOMERID'

4. Model Building

A train-test split was done on the cleaned data with a test size of 30% and a total of 12 base classifiers were tried out – 6 non-ensemble and 6 ensemble. The various classifiers were –

Non-Ensemble

1. Logistic Regression
2. Linear Discriminant Analysis
3. Decision Tree
4. Support Vector Machines
5. K-Nearest Neighbors
6. Naïve Bayes

Ensemble

1. Random Forest
2. Bagging
3. Adaboost
4. Gradient Boost
5. XGBoost
6. Light GBM

The performance metrics for the base **non-ensemble** classifiers on the test set are as follows:

***Table 7**: Non-ensemble Base Classifier Performance Comparison on Test set*

Metrics	DT_Test	SVM_Test	KNN_Test	Logit_Test	LDA_Test	NB_Test
Accuracy	0.993	0.966	0.968	0.939	0.916	0.845
Precision	0.972	0.926	0.918	0.728	0.523	0.340
Recall	0.953	0.672	0.706	0.518	0.670	0.778
F1 Score	0.962	0.779	0.798	0.605	0.587	0.473
AUC Score	0.975	0.500	0.500	0.500	0.500	0.500

Among these base non-ensemble classifiers, the Decision Tree Classifier was able to produce the best results on the test set in terms of all the performance metrics considered in this case. Hence, this non-ensemble classifier was taken for further model tuning and interpretation.

Here, F1-score has been taken as one of the major criterions for evaluating model performance along with the accuracy metric since the data is imbalanced.

The performance metrics for the base **ensemble** classifiers on the test set are as follows:

***Table 8**: Ensemble Base Classifier Performance Comparison on Test set*

Metrics	XGB_Test	LGBM_Test	RF_Test	Bag_Test	Adab_Test	GB_Test
Accuracy	0.996	0.996	0.989	0.995	0.992	0.994
Precision	0.985	0.983	0.984	0.987	0.962	0.975
Recall	0.967	0.976	0.888	0.953	0.948	0.961
F1 Score	0.976	0.979	0.934	0.970	0.955	0.968
AUC Score	1.000	1.000	0.998	0.997	0.998	0.999

Among these base ensemble classifiers, the XGBoost and LGBM Classifier was able to produce the best results on the test set in terms of the various performance metrics considered in this case (*except for Precision*). The performance metrics were very close for the two models; hence, these two ensemble classifiers were taken for further model tuning and interpretation.

The following techniques were applied to the Decision Tree, LGBM and XGB classifier in order to improve model performance:

- **Parameter Tuning** – The classifiers were tuned further to improve predictive performance and combat overfitting by applying regularization/pruning techniques.

- **Feature Selection and Elimination** – After tuning the model parameters, each model's feature importance scores were studied and features that were not adding value in the modeling process were dropped iteratively. Also, some more variables that have the possibility of overfitting during production and have less business interpretability such as latest transaction month, last receipt date etc. have been dropped from the modeling process.

- **Hyperparameter Optimization** – GridSearchCV technique was used to find the best set of model parameters for the given data with a 3-fold cross validation approach.

Finally, the features that were used in various models were studied and fine-tuned models' performance comparison was done.

***Table 9**: Fine-tuned Classifier Performance Comparison*

Metrics	Best_DT_Train	Best_DT_Test	Best_lgbm_Train	Best_lgbm_Test	Best_XGB_Train	Best_XGB_Test
Accuracy	0.9631	0.9614	0.9964	0.9823	0.9933	0.9853
Precision	0.8697	0.8714	0.9983	0.9736	0.9915	0.9807
Recall	0.6927	0.6685	0.9617	0.825	0.933	0.8529
F1 Score	0.7712	0.7566	0.9797	0.8931	0.9613	0.9124
AUC Score	0.9744	0.9519	1	0.9923	0.999	0.9939

Final Model – The XGBoost Classifier has achieved the best test set performance with an accuracy of 0.985 – indicating that the model is able to predict foreclosures with a very high accuracy. F1-score is also great at 0.91 and AUC score is 0.994. Precision 0.98 and recall metrics were also good at 0.98 and 0.85 respectively. Out of 5,989 test records, only 88 records were misclassified. Given the good performance, XGB model was chosen as the final model.

A 5-fold cross validation was also done with the final XGB model parameters and no signs of overfitting were noticed.

The performance metrics of the final fine-tuned XGB model on the train set is as follows:

```
              precision    recall  f1-score   support

           0     0.9934    0.9992    0.9963     12720
           1     0.9915    0.9330    0.9613      1253

    accuracy                         0.9933     13973
   macro avg     0.9925    0.9661    0.9788     13973
weighted avg     0.9933    0.9933    0.9932     13973
```

```
Accuracy :  0.9933
Precision:  0.9915
Recall   :  0.933
F1 Score :  0.9613
........................................

Confusion Matrix:
 [[12710    10]
 [   84  1169]]

True Negative: 12710
False Positive: 10
False Negative: 84
True Positive: 1169

Records correctly classified: 13879
Records incorrectly classified: 94
........................................
AUC Score: 0.999
........................................
```

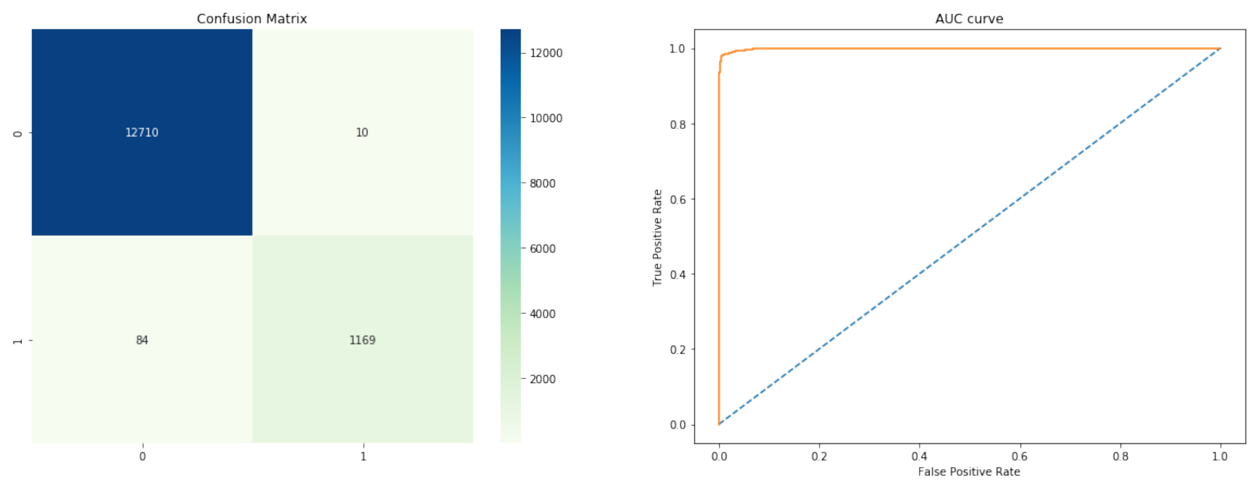


***Fig 6**: Final model – XGBoost Classifier Confusion Matrix and AUC Curve (train set)*

The performance metrics of the final fine-tuned XGB model on the test set is as follows:

```
              precision    recall  f1-score   support

           0     0.9857    0.9983    0.9920      5452
           1     0.9807    0.8529    0.9124       537

    accuracy                         0.9853      5989
```

```
   macro avg      0.9832     0.9256     0.9522       5989
weighted avg      0.9852     0.9853     0.9848       5989


Accuracy :  0.9853
Precision:  0.9807
Recall   :  0.8529
F1 Score :  0.9124
........................................

Confusion Matrix:
 [[5443    9]
 [  79  458]]

True Negative: 5443
False Positive: 9
False Negative: 79
True Positive: 458

Records correctly classified: 5901
Records incorrectly classified: 88
........................................

AUC Score: 0.9939

........................................
```

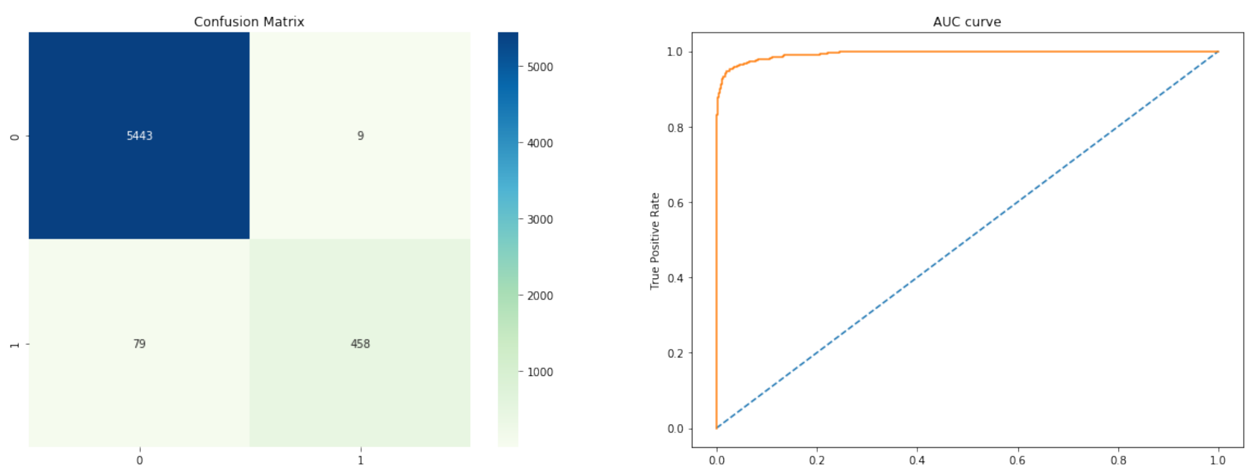


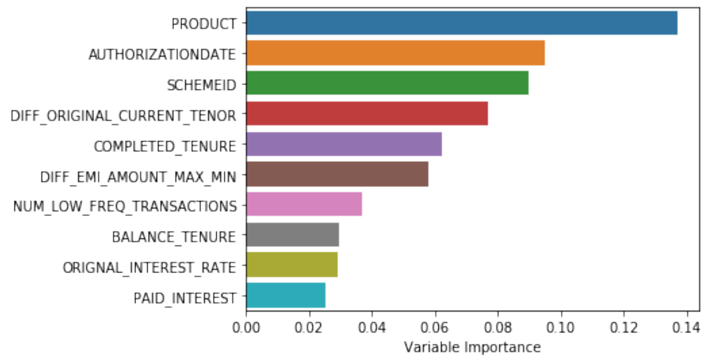
***Fig 7**: Final model – XGBoost Classifier Confusion Matrix and AUC Curve (test set)*

The variable importance for the final fine-tuned XGB model is as follows:

***Table 10**: Final XGBoost Model Feature Importance (top 10)*

Top 10 important variables	Feature Importance
PRODUCT	0.137
AUTHORIZATIONDATE	0.095
SCHEMEID	0.090
DIFF_ORIGINAL_CURRENT_TENOR	0.077
COMPLETED_TENURE	0.062
DIFF_EMI_AMOUNT_MAX_MIN	0.058
NUM_LOW_FREQ_TRANSACTIONS	0.037
BALANCE_TENURE	0.030
ORIGNAL_INTEREST_RATE	0.029
PAID_INTEREST	0.025

***Fig 8**: XGBoost Model Feature Importance Plot (top 10)*



- A total of 38 variables were used in this final XGB modeling process
- Product is the most important variable in this model in predicting foreclosures
- Authorization date and scheme ID are the next two important variables in classifying foreclosures
- City, FOIR, net disbursed amount, DPD etc. are some of the least important variables that have very less power in predicting foreclosures in this case compared to others used in this model.

5. Model Validation Approach

In this business case, the available data was imbalanced, where the records with foreclosures are about 9% of the entire dataset, which is not unusual in a real-world scenario. A higher number of foreclosures would mean heavy losses for the business where it will no longer be able to operate. Also, the class of prime importance in predictive modeling is usually the minority class or the foreclosures here, which we are trying to understand better and make predictions on. Here, if we are to build a classification model, the model when created could be biased towards the majority class and/or might overfit. Performance of such models will fail in production as it will not be able to generalize well. Thus, F1-score, i.e., the harmonic mean of recall and precision was another important measure that was considered for model validation. As high precision relates to a low false positive rate, the major objective in this case, i.e., to retain the good customers was achieved by not falsely classifying them as customers who will default on a loan payment leading to foreclosure. Apart from this, the model recall and AUC score were other important performance metrics considered in this problem.

6. Final interpretation / recommendation

- As was noticed during EDA, there are a total of 4 product categories, out of which, just 2 categories – HL and STHL contain almost all the foreclosures. Special attention must be given to loans under these products

- Scheme ID is another important feature in determining whether a loan may lead to foreclosure or not. Historically, loans under scheme ID 10901104 & 10901112 have seen high foreclosures, thus loans under these IDs will need moderation

- The end dates of the months seem to have a high number of loans authorized. This could potentially indicate that loans are given out by employees at the end of months in order to hit their sale targets. In such cases, borrower documents might not have been properly scrutinized and loans were given out without much of background check. Thus, loans authorized at the end of months must be monitored closely and the authorization process must be refined in case such operational loopholes are found.