Vehicle Loan Default Prediction

Submitted towards partial fulfillment of the criteria for the award of

PGPDSE

by **Great Lakes Institute of Management**

Submitted by

Group No. 4

Batch: December 2019

Group Members:

Gokul Mahendran

Krishna Raj P

Nantha Kumar S

Rishi Kumar Raman

Vivekananth

Mentored by

Mr. Romil Gupta

CERTIFICATE OF COMPLETION

I hereby certify that the project titled "**Vehicle Loan Default Prediction**" was undertaken and completed under my supervision by **Gokul Mahendran, Krishna Raj P, Nantha Kumar S, Rishi Kumar Raman & Vivekananth** of Post Graduate Program in Data Science and Engineering (PGPDSE).

Romil Gupta
Date: 28/04/2020
Place: Chennai

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our mentor **Romil Gupta** for providing his invaluable guidance, comments and suggestions throughout the course of this project. We value the assistance of Great Learning, Chennai campus. Learning from their knowledge helped me to become passionate about my research topic.

We will be failing in our duty if each one of us don't express our gratitude for other team members, for the valuable contributions during the course of this project.

ABSTRACT

This research is conducted to determine the Vehicle loan defaulters for L&T. Loan takers of L&T have been studied over a period of time and their ability to pay installments were studied and analyzing this will help us to reduce loan rejection rates which helps us gain a customer base by providing the loans to correct persons.

“The need for a better credit risk scoring model was also raised by the institution”

CNS score, Loan to value of the asset, Amount of loan disbursed, Number of Primary and secondary accounts maintained.

TABLE OF CONTENTS

INTRODUCTION

Vehicle loans are the new front in asset quality problems for banks. To be more precise, Financial institutions incur significant losses due to the default of vehicle loans.

There is huge amount of uncertainty in commercial vehicle loan because it hugely depends upon on various parameters. This has led to the tightening up of vehicle loan underwriting and increased vehicle loan rejection rates.

This warrants a study to estimate the determinants of vehicle loan default. The problem here is to accurately predict the probability of loanee/borrower defaulting on a vehicle loan in the first EMI (Equated Monthly Instalments) on the due date.
Given various parameters such as
Loanee Information (Demographic data like age, Identity proof etc.)
Loan Information (Disbursal details, loan to value ratio etc.)
Bureau data & history (Bureau score, number of active accounts, the status of other loans, credit history etc.)
Doing so will ensure that clients capable of repayment are not rejected and important determinants can be identified which can be further used for minimizing the default rates

1. Problem statement:

L&T financial institution incurs significant losses due to default vehicle loans, so they have to accurately predict the probability of loanee/borrower defaulting on a vehicle loan in the first EMI (Equated Monthly Instalments) on the due date.
Doing so will ensure that clients capable of repayment are not rejected and important determinants can be identified which can be further used for minimizing the default rates.

2. Data-set:

The dataset represents Vehicle Loan Default of L&T financial institution. The dataset from L&T, includes over 41 features to analyze the loanee/borrower defaulting on a vehicle loan in the first EMI.

3. Shape:

233154 rows and 41 columns

LITERATURE:

Feature	Explanation	Data types
UniqueID	Identifier for customers (Unique ID for Customers)	Object
loan_default	Payment default in the first EMI on due date	Categorical
disbursed_amount	Amount of Loan disbursed	Continuous
asset_cost	Cost of the Asset	Continuous
ltv	Loan to Value of the asset	Continuous
branch_id	Branch where the loan was disbursed	Object
supplier_id	Vehicle Dealer where the loan was disbursed	Object
manufacturer_id	Vehicle manufacturer (Hero, Honda, TVS etc.)	Object
Current_pincode	Current pin code of the customer	Categorical
Date.of.Birth	Date of birth of the customer	object
Employment.Type	Employment Type of the customer (Salaried/Self Employed)	Categorical

DisbursalDate	Date of disbursement	object
State_ID	State of disbursement	Categorical
Employee_code_ID	Employee of the organization who logged the disbursement	object
MobileNo_Avl_Flag	if Mobile no. was shared by the customer then flagged as 1	Categorical
Aadhar_flag	if Aadhar was shared by the customer then flagged as 1	Categorical
PAN_flag	if pan was shared by the customer then flagged as 1	Categorical
VoterID_flag	if voter was shared by the customer then flagged as 1	Categorical
Driving_flag	if DL was shared by the customer then flagged as 1	Categorical
Passport_flag	if passport was shared by then 1	Categorical
PERFORM_CNS.SCORE	Bureau Score	Categorical
PERFORM_CNS.SCORE.DESCRIPTION	Bureau score description	Categorical
PRI.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement	Categorical

PRI.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement	Categorical
PRI.OVERDUE.ACCTS	count of default accounts at the time of disbursement	Categorical
PRI.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement	Continuous
PRI.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement	Continuous
PRI.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement	Continuous
SEC.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement	Categorical
SEC.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement	Categorical
SEC.OVERDUE.ACCTS	count of default accounts at the time of disbursement	Categorical
SEC.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement	Continuous

SEC.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement	Continuous
SEC.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement	Continuous
PRIMARY.INSTAL.AMT	EMI Amount of the primary loan	Continuous
SEC.INSTAL.AMT	EMI Amount of the secondary loan	Continuous
NEW.ACCTS.IN.LAST.SIX.MONTHS	New loans taken by the customer in last 6 months before the disbursement	Categorical
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS	Loans defaulted in the last 6 months	Categorical
AVERAGE.ACCT.AGE	Average loan tenure	Categorical
CREDIT.HISTORY.LENGTH	Time since first loan	Categorical
NO.OF_INQUIRIES	Enquires done by the customer for loans	Categorical

DATA CLEANING

The following steps have been done for better analysis, visualization and model building:

Checking Missing Values:

We can start analysis by looking at the percentage of missing values in each column. Missing values are fine when we do Exploratory Data Analysis, but they will have to be filled in for machine learning methods.

	Total	Percent
AVERAGE.ACCT.AGE	0	0.00
Aadhar_flag	0	0.00
CREDIT.HISTORY.LENGTH	0	0.00
Current_pincode_ID	0	0.00
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS	0	0.00
Date.of.Birth	0	0.00
DisbursalDate	0	0.00
Driving_flag	0	0.00
Employee_code_ID	0	0.00
Employment.Type	7661	3.29
MobileNo_Avl_Flag	0	0.00
NEW.ACCTS.IN.LAST.SIX.MONTHS	0	0.00
NO.OF_INQUIRIES	0	0.00
PAN_flag	0	0.00
PERFORM_CNS.SCORE	0	0.00
PERFORM_CNS.SCORE.DESCRIPTION	0	0.00
PRI.ACTIVE.ACCTS	0	0.00
PRI.CURRENT.BALANCE	0	0.00
PRI.DISBURSED.AMOUNT	0	0.00
PRI.NO.OF.ACCTS	0	0.00
PRI.OVERDUE.ACCTS	0	0.00
PRI.SANCTIONED.AMOUNT	0	0.00
PRIMARY.INSTAL.AMT	0	0.00
Passport_flag	0	0.00
SEC.ACTIVE.ACCTS	0	0.00
SEC.CURRENT.BALANCE	0	0.00
SEC.DISBURSED.AMOUNT	0	0.00
SEC.INSTAL.AMT	0	0.00
SEC.NO.OF.ACCTS	0	0.00
SEC.OVERDUE.ACCTS	0	0.00
SEC.SANCTIONED.AMOUNT	0	0.00
State_ID	0	0.00
UniqueID	0	0.00
VoterID_flag	0	0.00
asset_cost	0	0.00
branch_id	0	0.00
disbursed_amount	0	0.00
loan_default	0	0.00
ltv	0	0.00

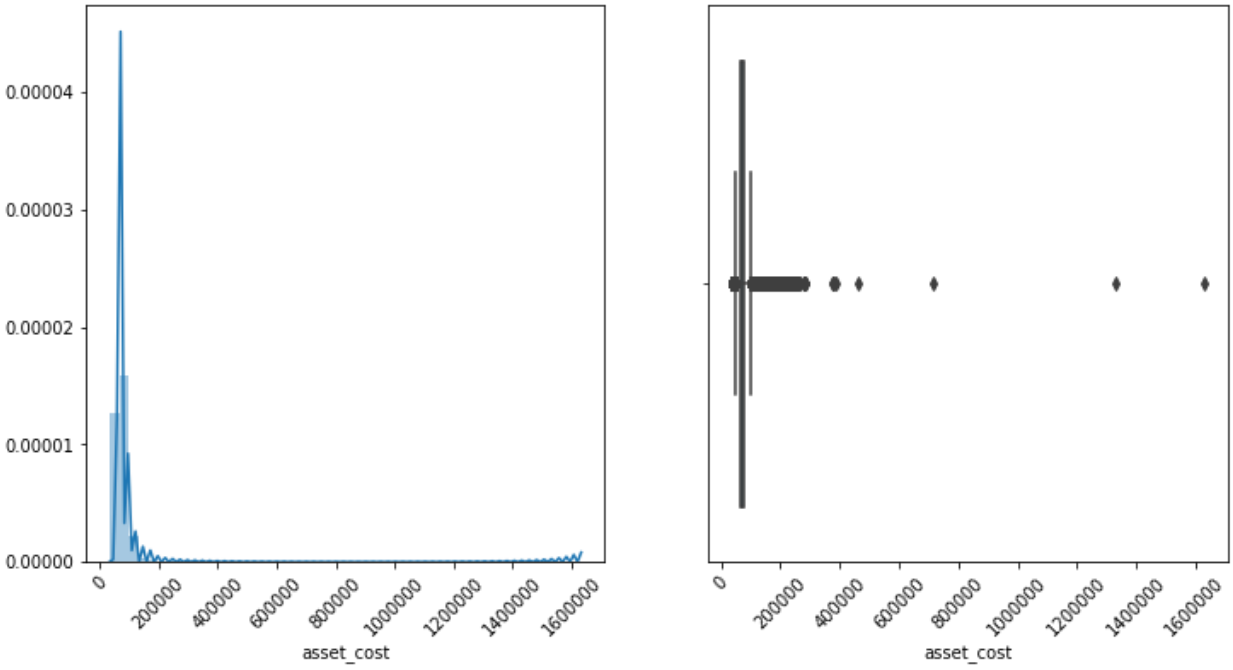
EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an open-ended process where we make plots and calculate statistics in order to explore our data. The purpose is to to find anomalies, patterns, trends, or relationships. These may be interesting by themselves (for example finding a correlation between two variables) or they can be used to inform modeling decisions such as which features to use. In short, the goal of EDA is to determine what our data can tell us! EDA generally starts with a high-level overview and then narrows into specific parts of the dataset once as we find interesting areas to examine.

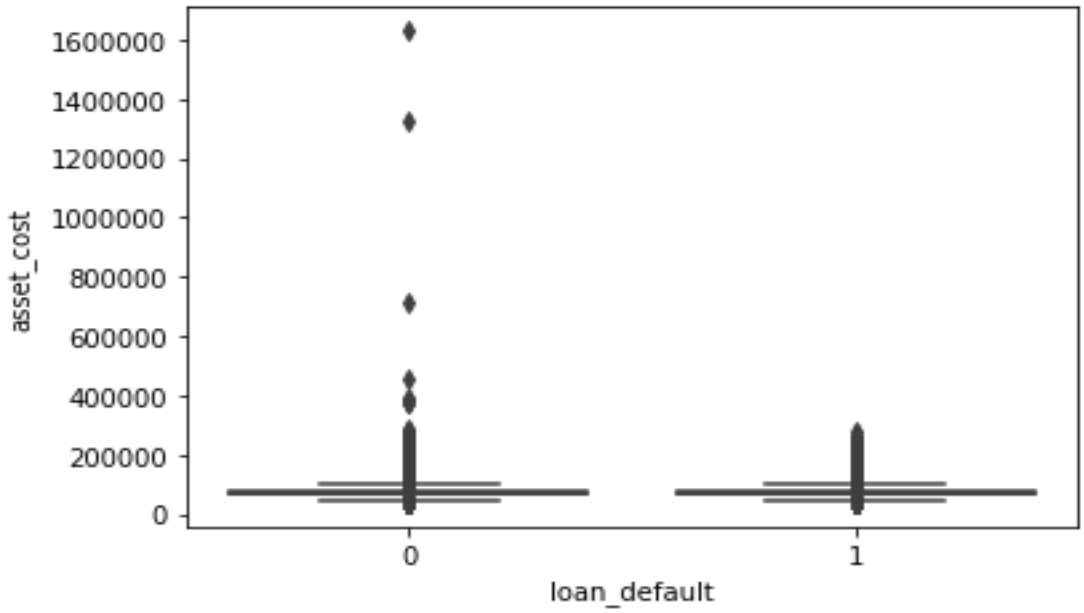
Numerical Columns

Asset Cost

To calculate how much the value of these assets has decreased due to depreciation, we first need to know their original costs. This is defined as the original price of the asset from which we can determine its depreciated value over the course of its useful life.



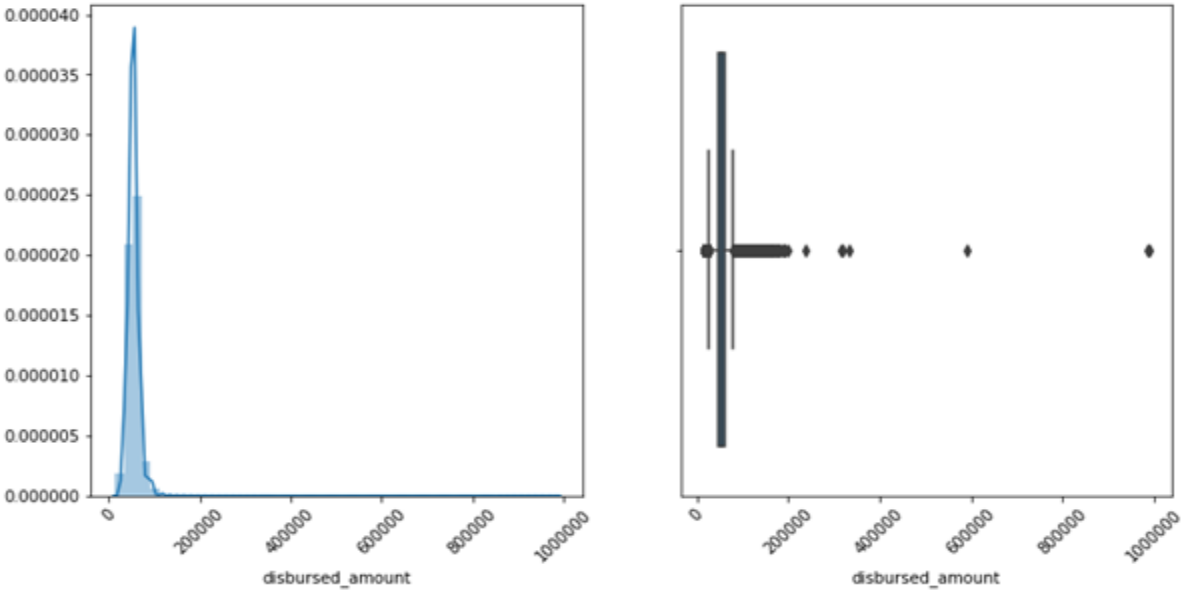
The data is right skewed and there are outlies present in the upper range.



The upper range of outliers are present migratorily among non-defaulter group which is very evident from the boxplot.

Disbursed Amount:

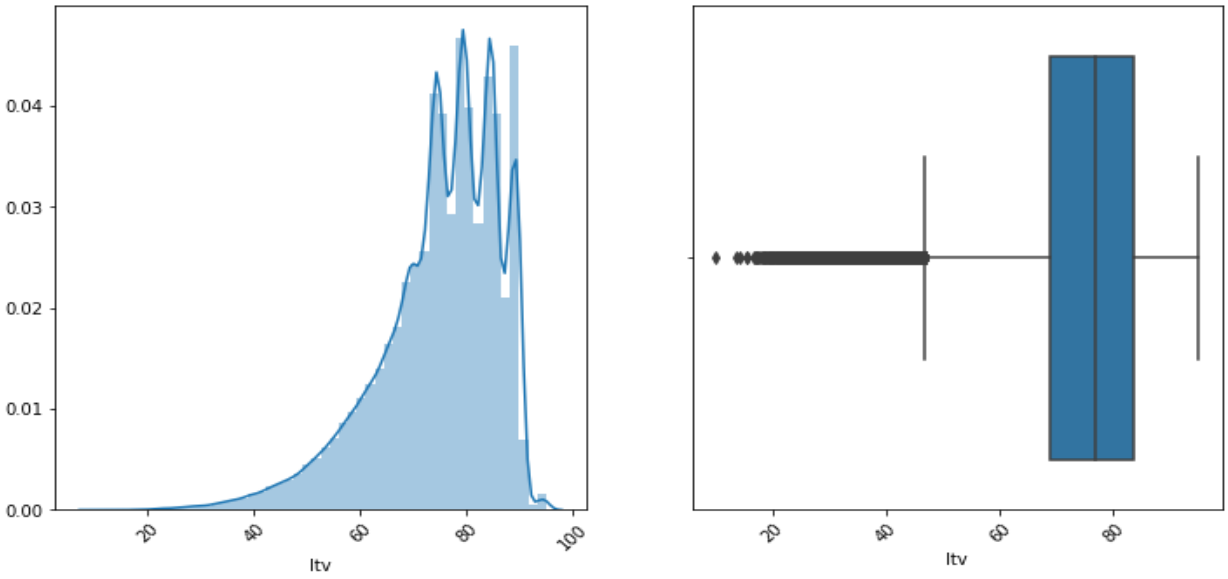
In accounting terms, a disbursement also called a cash disbursement or cash payment refers to a wide range of payment types made in a specific period, including interest payments on loans and operating expenses. Disbursement can also refer to a loan payment, such as a vehicle loan.



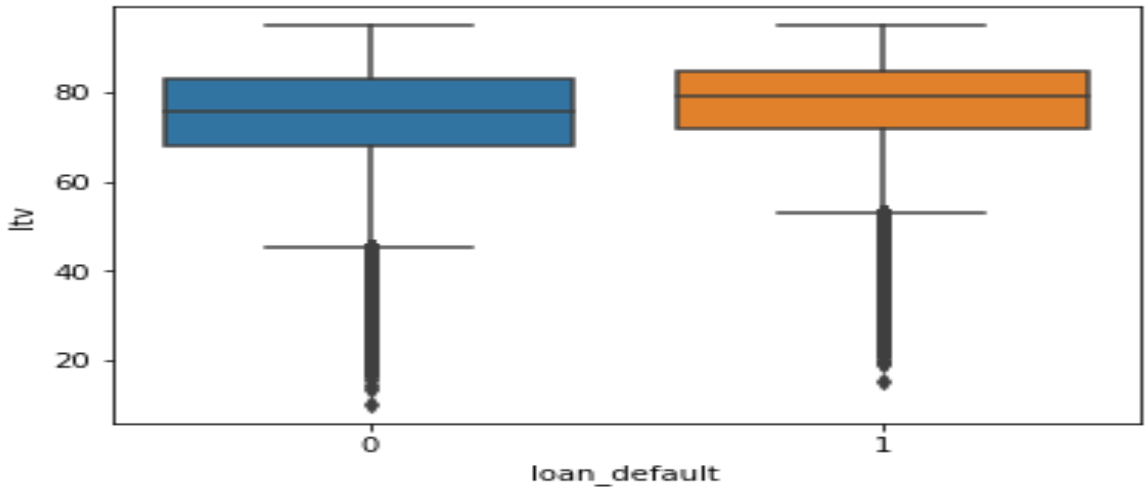
We can see data is right-skewed and outliers are found in the upper range. Extreme Outliers found in the above image are from the non-defaulter's category.

LTV:

The loan-to-value (LTV) the ratio is a financial term used by lenders to express the ratio of a loan to the value of an asset purchased. The term is commonly used by banks and building societies to represent the ratio of the first mortgage line as a percentage of the total appraised value of a real property.



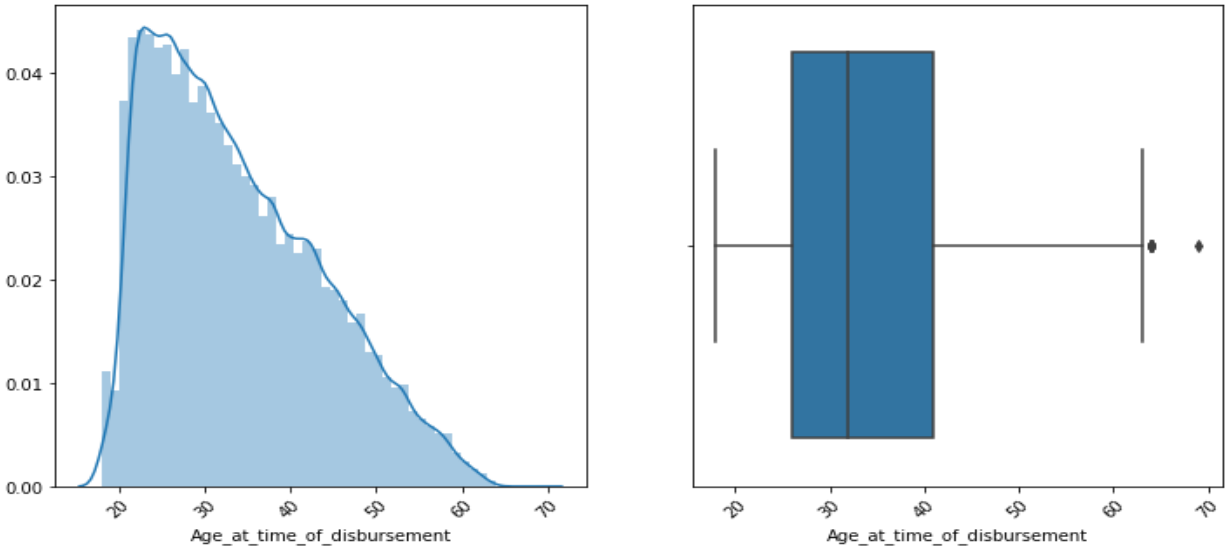
The data seems to be left skewed and we could see outliers in the lower range. Most customers seem to be taking loans for the most part of vehicle cost which seems correct as they don't want most of their capital into vehicles.



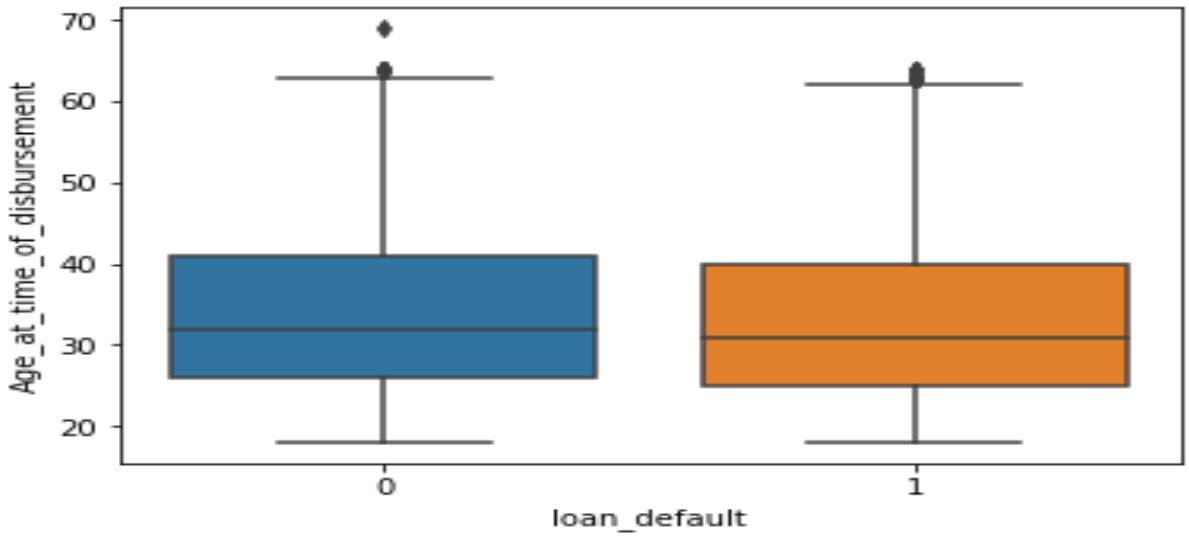
Non-defaulters lower range seems to be less than defaulters and the spread is larger than defaulters.

Age at time of disbursement:

This is a created feature from Disbursal Date and Date_of_Birth.



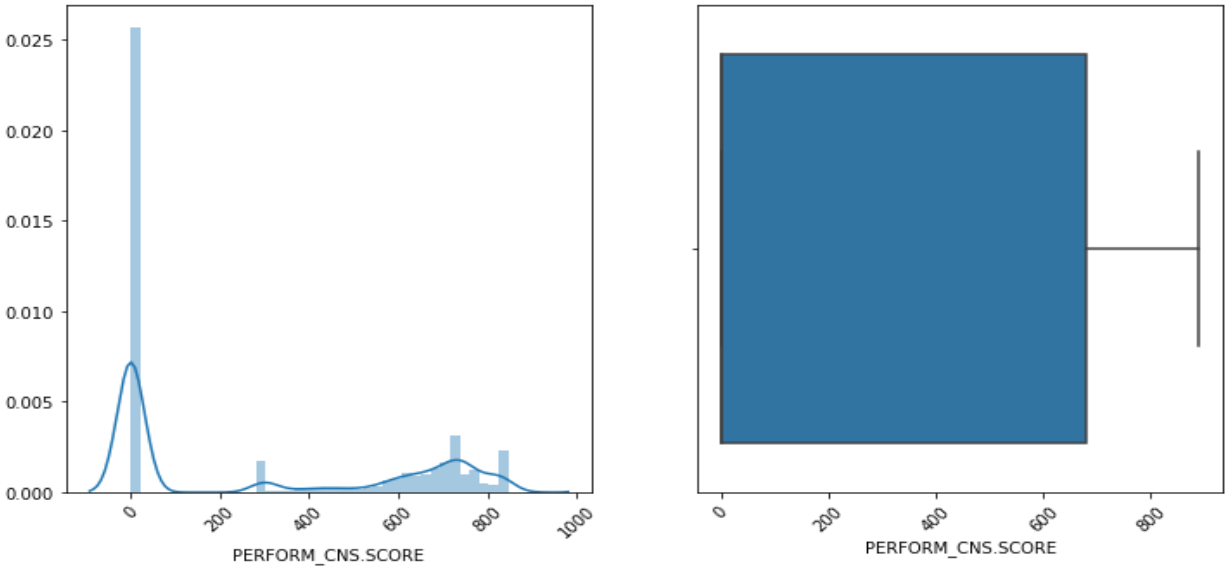
This variable seems little close to being a normal distribution with little skewness towards the right. We can spread from 20 to 70 with peakedness at 20-30.



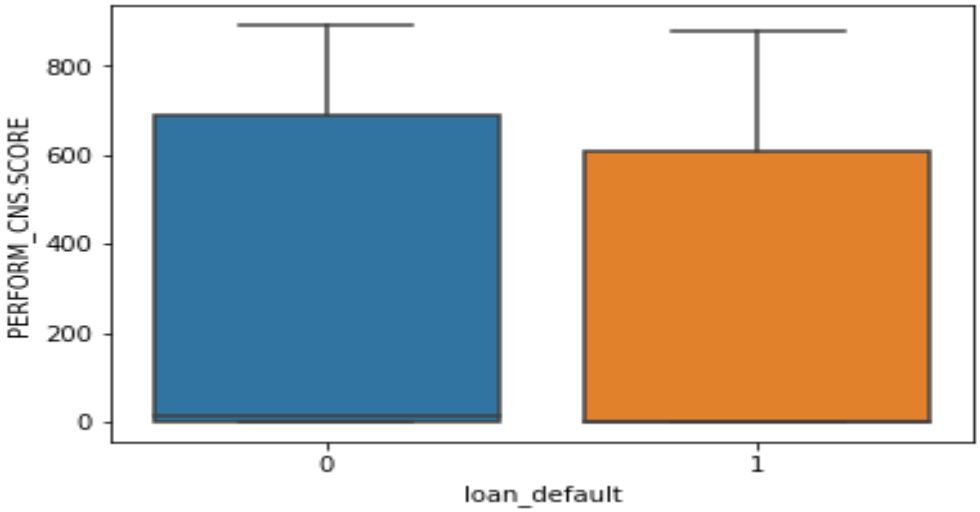
The mean, median, min and max seem to be same for both groups. It won't be a decisive variable in splitting the target variable. Age together with other variables could be decisive.

PERFORM CNS SCORE

A CIBIL score is a three-digit number between 300-900, 300 being the lowest, that represents an individual's creditworthiness. A higher CIBIL score suggests good credit history and responsible repayment behaviour.



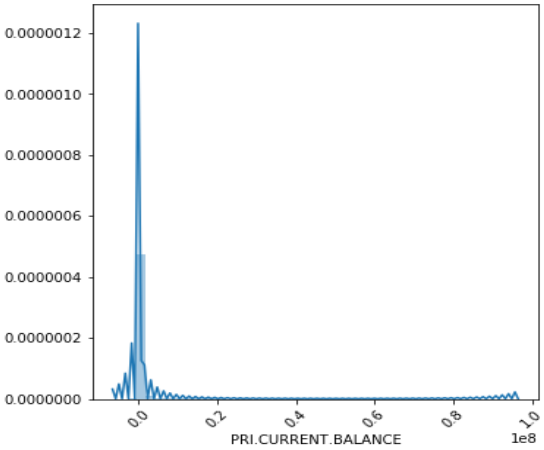
As we lot zeros in the graph which conveys, they are probably first-time loan takers. This score has a particular range between 300-900 which portrays the non-existence of outliers.

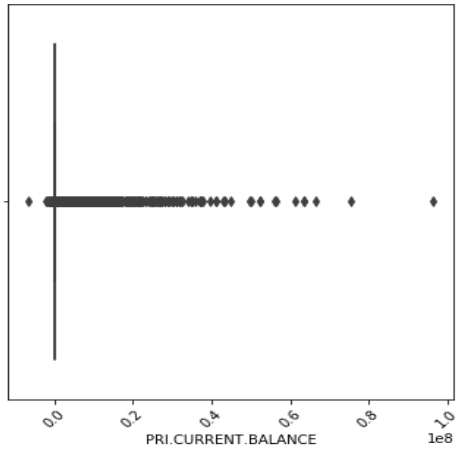


Non-defaulters seem to have higher 75% quantile than defaulters as defaulters tend to have lesser cibil score.

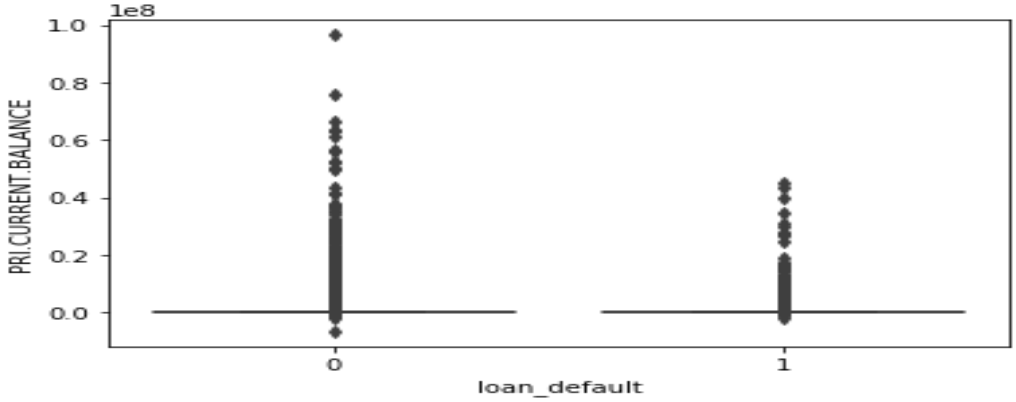
PRI CURRENT BALANCE

When referring to a loan, such as an auto loan or a mortgage, your current balance is the amount you currently still owe on the loan according to the date of your statement.





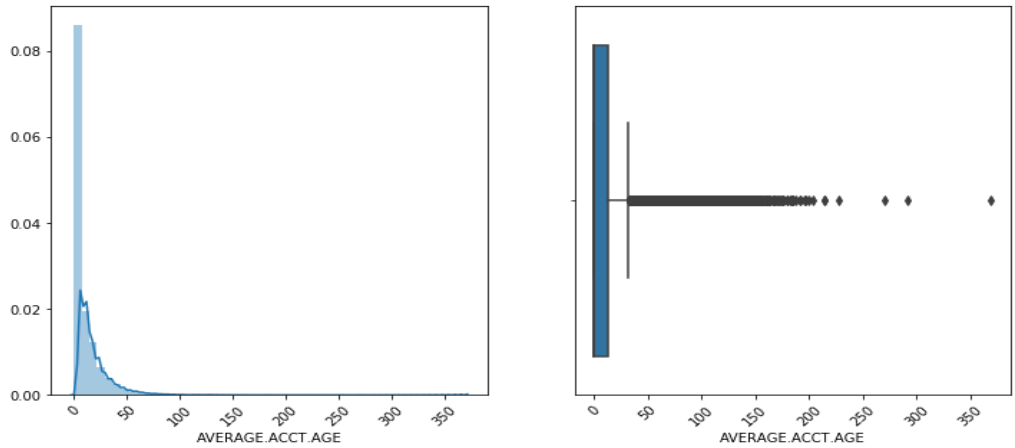
The data is right skewed and nearly 75% of values seems to zero which shows everyone have paid the previous loan installments. Negative values in the data shows bank owes amount back to borrower.



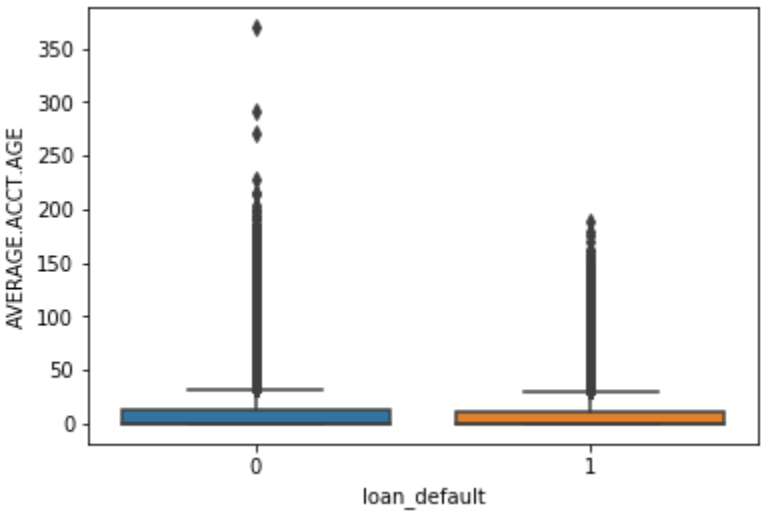
Non-defaulters have higher current balance which may due to previous not yet completed with some installments remaining which could be seen with some bivariate analysis

AVERAGE ACCT AGE

AVERAGE.ACCT.AGE gives the information regarding the average duration of loan tenure of all accounts owned by borrowers.



We could see outliers in upper range and data see to be right-skewed with a lot of zeros.

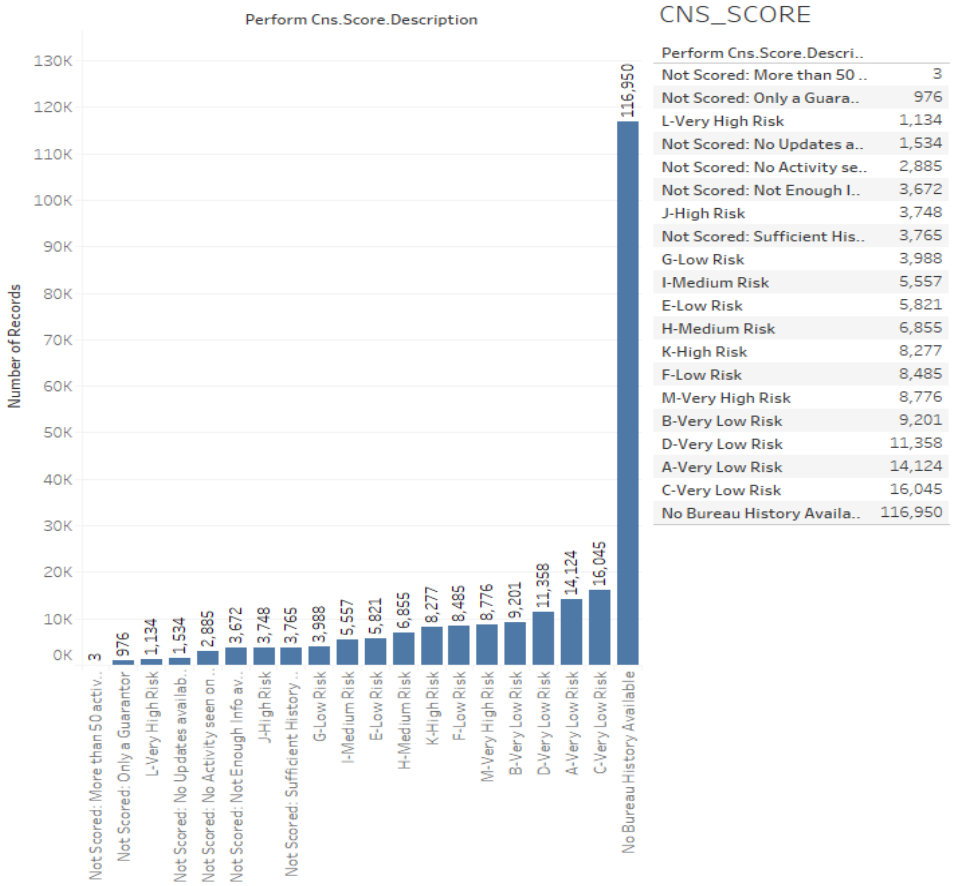


We could see longer tenure duration in non-defaulters which converts into lower monthly installments which also reduces the risk of getting defaulted.

Categorical Columns

PERFORM CNS SCORE DESCRIPTION:

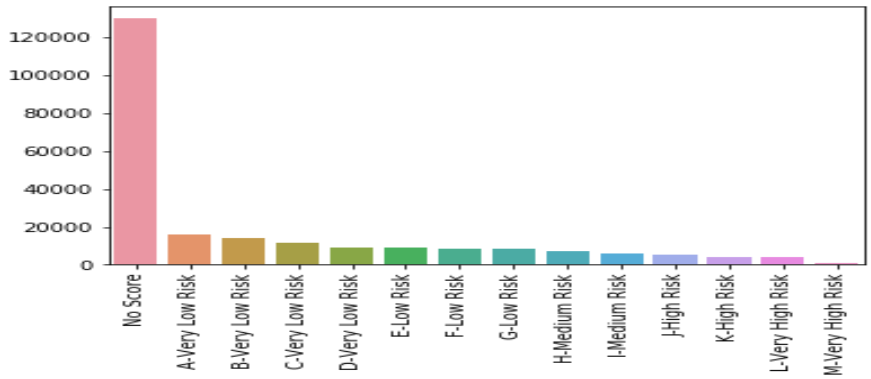
A single numerical score, based on information in an individual's credit report, that measures that individual's creditworthiness. Credit scores are based on statistical studies of the relationship between the different items in a credit report and the likelihood of default.



CNS_SCORE

Perform Cns.Score.Descri..	
Not Scored: More than 50 ..	3
Not Scored: Only a Guara..	976
L-Very High Risk	1,134
Not Scored: No Updates a..	1,534
Not Scored: No Activity se..	2,885
Not Scored: Not Enough I..	3,672
J-High Risk	3,748
Not Scored: Sufficient His..	3,765
G-Low Risk	3,988
I-Medium Risk	5,557
E-Low Risk	5,821
H-Medium Risk	6,855
K-High Risk	8,277
F-Low Risk	8,485
M-Very High Risk	8,776
B-Very Low Risk	9,201
D-Very Low Risk	11,358
A-Very Low Risk	14,124
C-Very Low Risk	16,045
No Bureau History Availa..	116,950

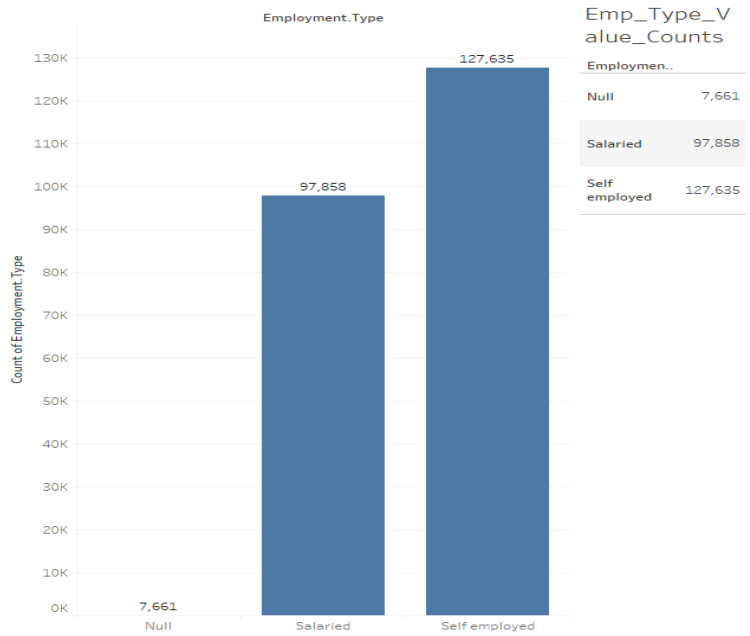
Feature extraction was performed to the CNS Score Description wherein the No scorers were binned. The result of binning is shown in the bar plot below.



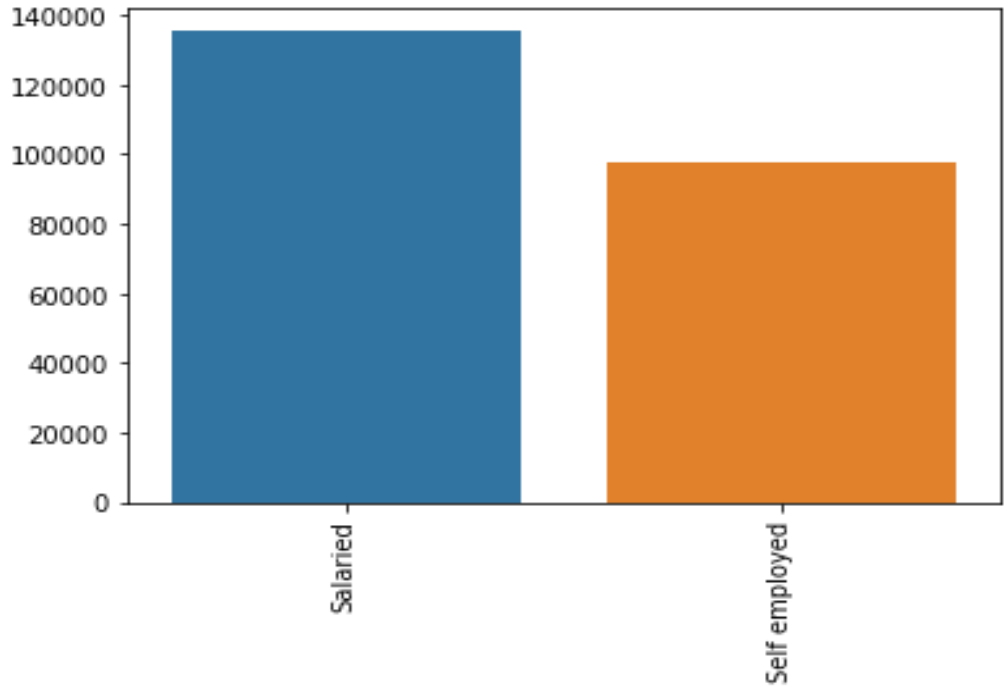
It is very evident from the bar plot that as the credit score increases the number of people having high score decreases. An import analysis made out from this is there are a greater number of people in Low scoring category.

Employment Type:

There are two types of employees present salaried and self-employed.



The nulls present in the employee column accounted for 3.29% of the data which was filled with the most frequent employee type present. After imputations, the data resembles the below structure.



STATISTICAL DATA ANALYSIS (Numerical Columns):

OneWay-Anova Test:

Since our target variable is dichotomous to find whether the feature is significant or not. We will do an oneway-anova test to find out the significance of the feature.

	F_stat	P_value
disbursed_amount	4.094325e+06	0.000000e+00
asset_cost	3.738909e+06	0.000000e+00
ltv	9.854228e+06	0.000000e+00
Age_at_time_of_disbursement	2.746409e+06	0.000000e+00
PERFORM_CNS.SCORE	1.703651e+05	0.000000e+00
PRI.CURRENT.BALANCE	7.227380e+03	0.000000e+00
PRI.SANCTIONED.AMOUNT	1.973823e+03	0.000000e+00
PRI.DISBURSED.AMOUNT	1.961044e+03	0.000000e+00
PRIMARY.INSTAL.AMT	1.747700e+03	0.000000e+00
AVERAGE.ACCT.AGE	7.725110e+04	0.000000e+00
CREDIT.HISTORY.LENGTH	7.337463e+04	0.000000e+00
SEC.SANCTIONED.AMOUNT	3.699434e+02	2.072709e-82
SEC.DISBURSED.AMOUNT	3.604951e+02	2.356066e-80
SEC.CURRENT.BALANCE	2.369991e+02	1.829059e-53
SEC.INSTAL.AMT	1.005817e+02	1.142408e-23

As we have done our test, all our variables have rejected null hypothesis in favour of alternate hypothesis which states that the mean of two groups are not equal which will help in splitting our target variable.

These test does explain significance but not the strength of the variable associated with the target variable. So, we performed a point biserial test to explain the strength.

Point Biserial R Test for Numerical Variables

The point biserial correlation coefficient is a special case of Pearson's correlation coefficient. It measures the relationship between two variables:

- One continuous variable (must be ratio scale or interval scale).
- One naturally binary variable.

	Strength
ltv	0.0982078
disbursed_amount	0.0776749
PERFORM_CNS.SCORE	0.0579291
CREDIT.HISTORY.LENGTH	0.0421261
Age_at_time_of_disbursement	0.0365491
PRI.CURRENT.BALANCE	0.0273857
AVERAGE.ACCT.AGE	0.0247809
asset_cost	0.0142613
PRI.SANCTIONED.AMOUNT	0.0113045
PRI.DISBURSED.AMOUNT	0.0111555
PRIMARY.INSTAL.AMT	0.0106158
SEC.SANCTIONED.AMOUNT	0.00635432
SEC.DISBURSED.AMOUNT	0.0062483
SEC.CURRENT.BALANCE	0.00553145
SEC.INSTAL.AMT	0.00154848

We found that ltv, disbursed amount and CNS.SCORE are most important variable, while Secondary loan associated variable seems to be less important.

STATISTICAL DATA ANALYSIS (Categorical Columns) (Chi-Square Test):

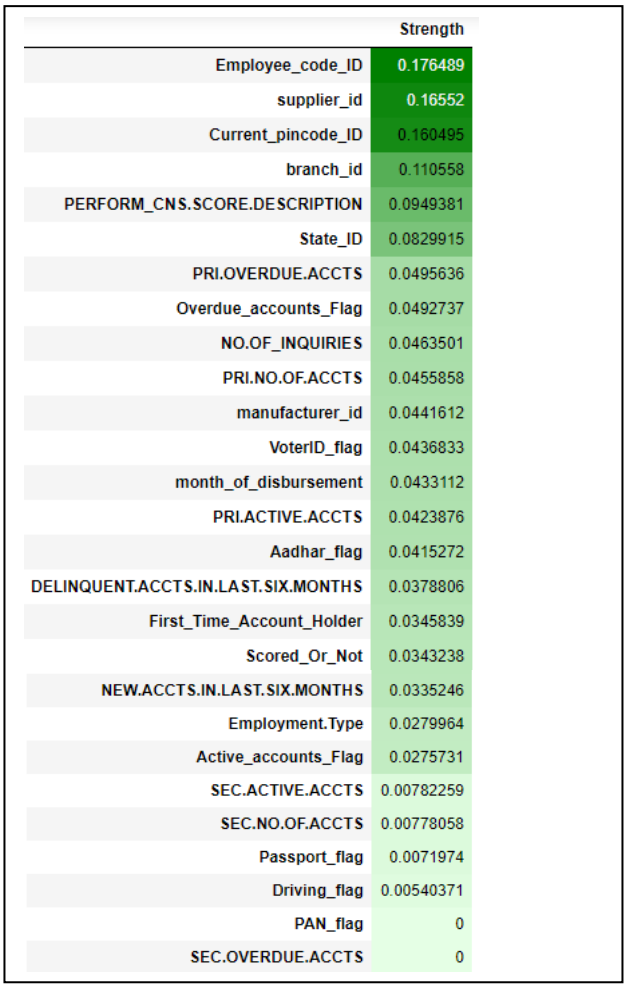
For Categorical columns, we have to perform Chi Contingency Test to explain significance of the variable. The Chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables. The frequency of each category for one nominal variable is compared across the categories of the second nominal variable. The data can be displayed in a contingency table where each row represents a category for one variable and each column represents a category for the other variable.

	Chi_Stat	P_value
Current_pincode_ID	12702.769933	0.000000e+00
branch_id	2930.842784	0.000000e+00
State_ID	1626.861781	0.000000e+00
PERFORM_CNS.SCORE.DESCRIPTION	2114.465408	0.000000e+00
supplier_id	9339.709282	0.000000e+00
Employee_code_ID	10531.308447	0.000000e+00
Overdue_accounts_Flag	567.070856	2.434934e-125
PRI.OVERDUE.ACCTS	593.752865	3.291632e-112
VoterID_flag	445.908843	5.603683e-99
month_of_disbursement	439.361517	3.925174e-96
NO.OF_INQUIRIES	524.890463	1.117569e-95
manufacturer_id	464.697640	1.527738e-93
Aadhar_flag	403.074838	1.179200e-89
PRI.ACTIVE.ACCTS	457.907532	6.539244e-73
PRI.NO.OF.ACCTS	591.507498	3.947704e-68
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS	347.560233	2.528829e-66
First_Time_Account_Holder	279.861154	8.051059e-63
Scored_Or_Not	275.683095	6.551645e-62
NEW.ACCTS.IN.LAST.SIX.MONTHS	287.040943	2.366614e-46
Employment.Type	183.744606	7.376772e-42
Active_accounts_Flag	178.260773	1.161915e-40
Passport_flag	13.077913	2.987982e-04
Driving_flag	7.808088	5.201291e-03
SEC.ACTIVE.ACCTS	36.267393	2.842133e-02
SEC.NO.OF.ACCTS	50.114650	5.918409e-02
PAN_flag	0.957671	3.277743e-01
UniqueID	233154.000000	4.990263e-01

From the test, we can see Pan flag, Unique_id, mobileflag and Sec.Overdue.Accts have failed the test and so they are insignificant variables to the target variable.

Cramer's test for Categorical Variables

Cramer's V is a number between 0 and 1 that indicates how strongly two categorical variables are associated

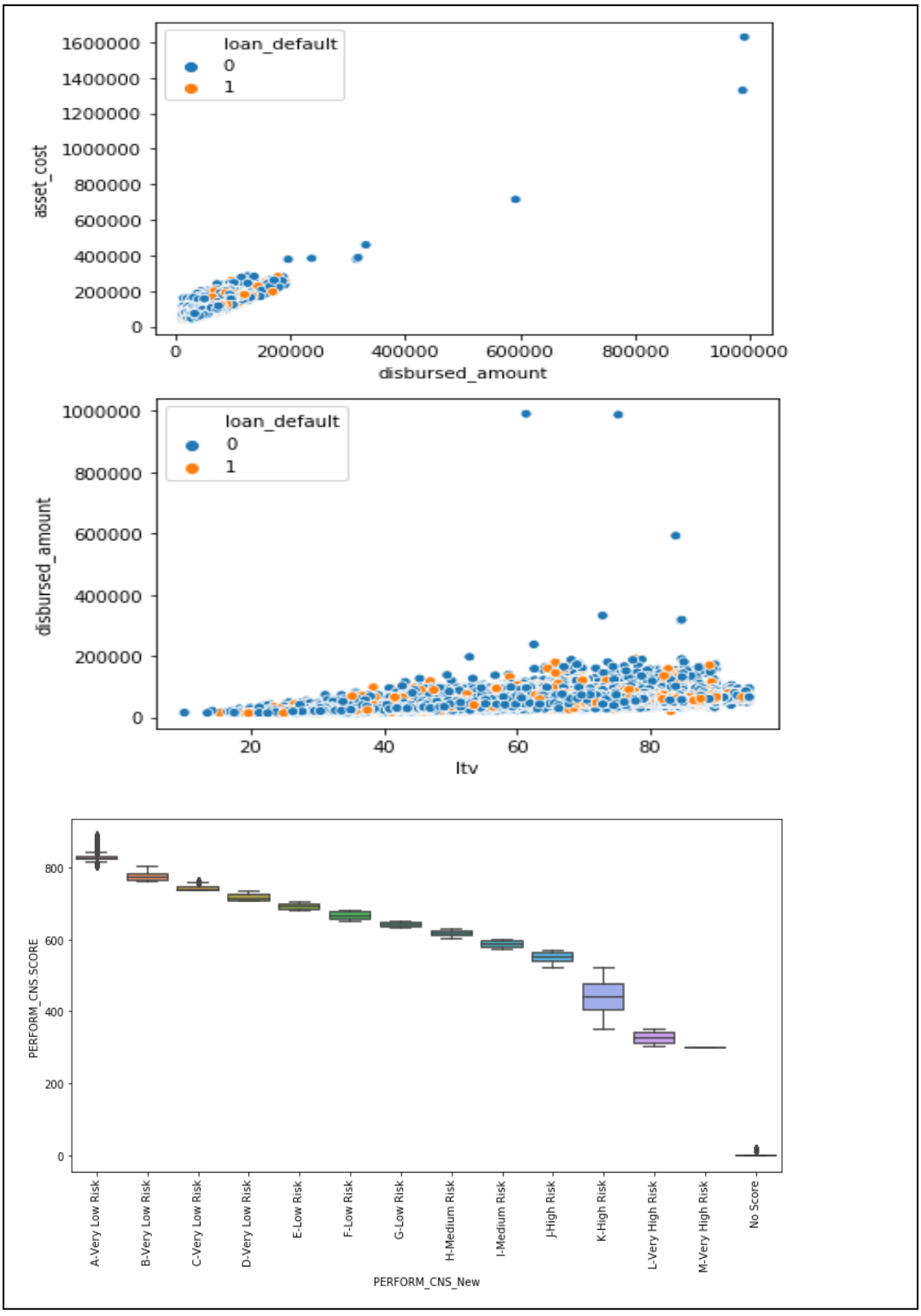


	Strength
Employee_code_ID	0.176489
supplier_id	0.16552
Current_pincode_ID	0.160495
branch_id	0.110558
PERFORM_CNS.SCORE.DESCRIPTION	0.0949381
State_ID	0.0829915
PRI.OVERDUE.ACCTS	0.0495636
Overdue_accounts_Flag	0.0492737
NO.OF_INQUIRIES	0.0463501
PRI.NO.OF.ACCTS	0.0455858
manufacturer_id	0.0441612
VoterID_flag	0.0436833
month_of_disbursement	0.0433112
PRI.ACTIVE.ACCTS	0.0423876
Aadhar_flag	0.0415272
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS	0.0378806
First_Time_Account_Holder	0.0345839
Scored_Or_Not	0.0343238
NEW.ACCTS.IN.LAST.SIX.MONTHS	0.0335246
Employment.Type	0.0279964
Active_accounts_Flag	0.0275731
SEC.ACTIVE.ACCTS	0.00782259
SEC.NO.OF.ACCTS	0.00778058
Passport_flag	0.0071974
Driving_flag	0.00540371
PAN_flag	0
SEC.OVERDUE.ACCTS	0

The strength of categorical variables can be explained with Cramer's test. We can see that employee_code_id, supplier_id, pincode and branch seems to be strong with the target. But these variables have huge unique values in them. Using them effectively in the model is highly difficult as it can increase no of columns which can cause the curse of dimensionality. Hence, PERFORM.CNS.SCORE.DESC is the most important variable in our case.

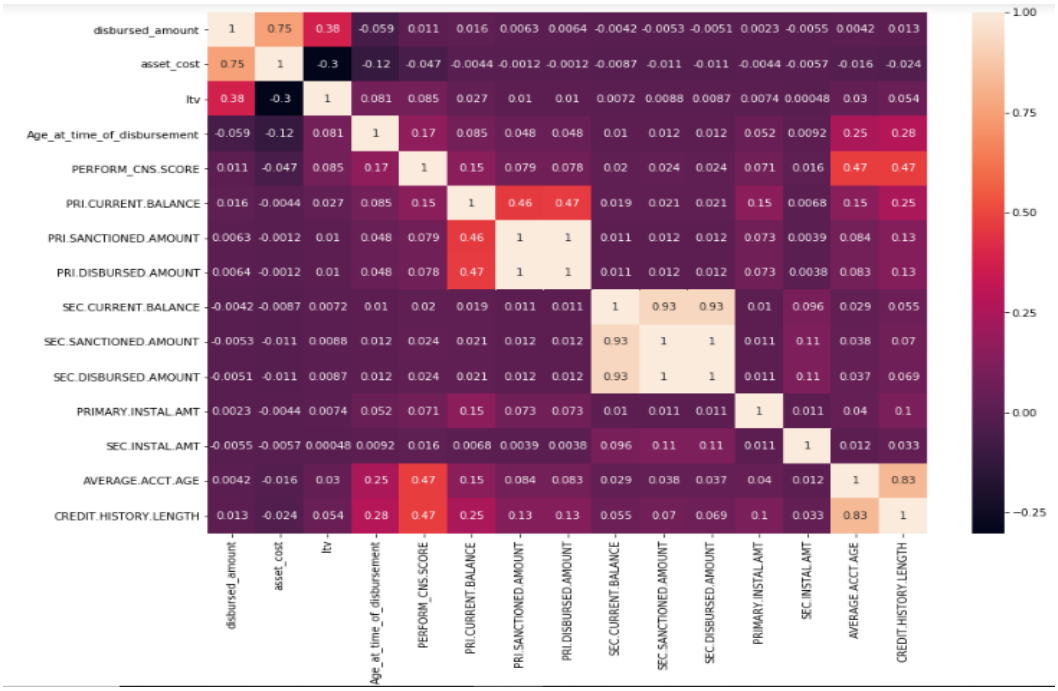
Bivariate analysis and Multivariate analysis

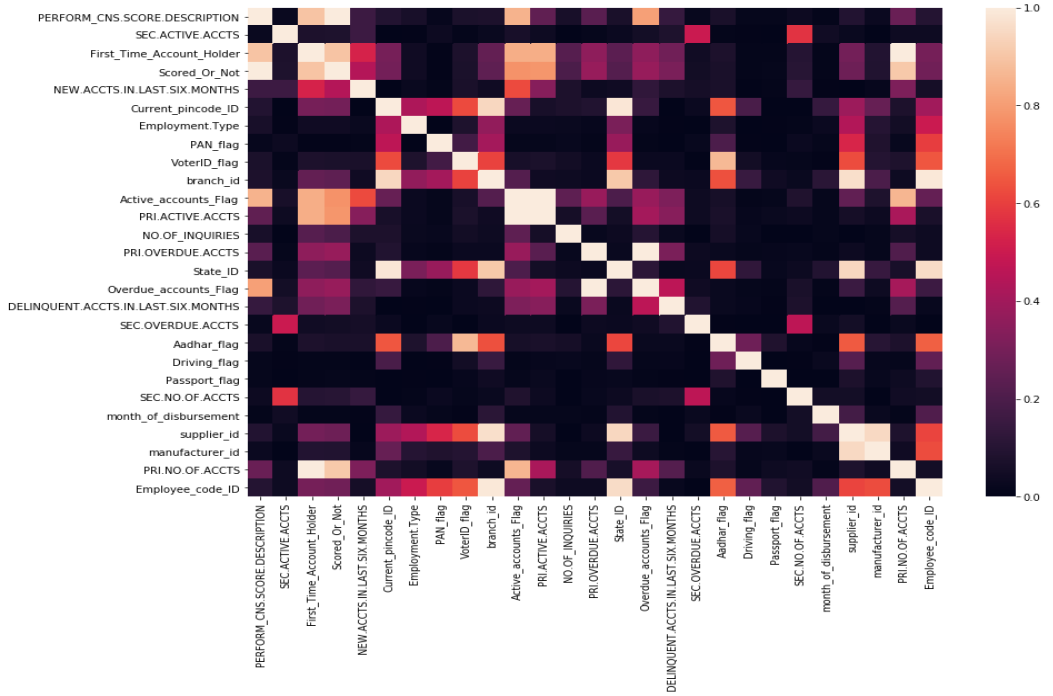
Bivariate and multivariate analyses are statistical methods to investigate relationships between data samples. The Bivariate analysis looks at two paired data sets, studying whether a relationship exists between them. The Multivariate analysis uses two or more variables and analyzes which, if any, are correlated with a specific outcome. The goal in the latter case is to determine which variables influence or cause the outcome.



Correlation for Numerical and Categorical Columns

Primary Disbursed Amount, Primary Sanctioned Amount and Primary Current Balance are highly correlated variables.





Base Model of Different Estimators

Here we are trying Linear, distance and tree-based models in the conviction which splits the target variables at its best. Since the metric of interest for the problem statement is AUC, from the below output, we can conclude that tree-based generally outperforms linear based models, hence we would be using the tree-based model for our further analysis.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa
0	Light Gradient Boosting Machine	0.7834	0.6609	0.0147	0.541	0.0286	0.0174
1	CatBoost Classifier	0.7833	0.6643	0.0278	0.5137	0.0527	0.0312
2	Gradient Boosting Classifier	0.7832	0.6564	0.0047	0.5767	0.0094	0.0059
3	Extreme Gradient Boosting	0.7831	0.6559	0.003	0.5891	0.0059	0.0037
4	Logistic Regression	0.7829	0.5518	0	0.1667	0.0001	-0
5	Ridge Classifier	0.7828	0	0.0016	0.4048	0.0031	0.0015
6	Ada Boost Classifier	0.7828	0.6469	0.0152	0.4897	0.0294	0.0166
7	Linear Discriminant Analysis	0.7824	0.6392	0.0068	0.4265	0.0133	0.0066
8	Extra Trees Classifier	0.7773	0.612	0.0498	0.3976	0.0884	0.0421
9	Random Forest Classifier	0.771	0.5881	0.0656	0.3522	0.1106	0.0455
10	K Neighbors Classifier	0.7443	0.5406	0.1103	0.2768	0.1578	0.0389
11	SVM - Linear Kernel	0.7087	0	0.154	0.2138	0.1578	0.0139
12	Decision Tree Classifier	0.6711	0.5307	0.2827	0.2616	0.2718	0.0597
13	Naive Bayes	0.6009	0.5767	0.3325	0.2074	0.1344	0.0041
14	Quadratic Discriminant Analysis	0.2277	0.5394	0.9896	0.2181	0.3575	0.0027

MODEL BUILDING

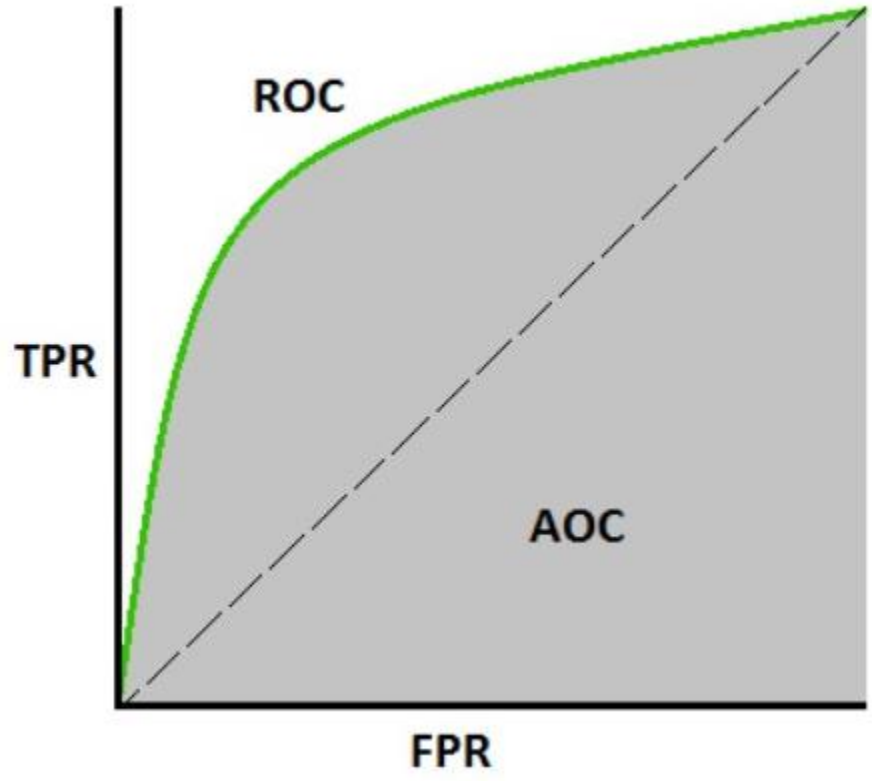
Models that we choose for this dataset are:

1. Random Forest Classifier (RFC)
2. Logistic Regression
3. Light GBM
4. KNN Classifier

Evaluation Metrics for the model:

The evaluation metric we would like to choose is ROC_AUC.

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represent degree or measure of separability. It tells how much model is capable of distinguishing between classes.



Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

Accuracy means how many data points/observations are predicted correctly out of all number of observations.

In a Confusion Matrix point of View Accuracy is calculated by:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Where:

TP	True Positives
TN	True Negatives
FP	False Positive
FN	False Negative

True Positives and True Negatives are the ones which the classifier correctly predicts the class of an observation.

So, adding TP and TN gives us all the correct values predicted by the classifier and dividing by the sum all possible states is the same as dividing by the total number of observations. This gives us an output value ranging from 0 to 1 where 0 means the classifier has 0% accuracy meaning it has classified all classes wrong and 1 means the classifier has 100% accuracy classifying all classes correctly.

Another metric called is Precision. It is the number of positive prediction values divided by the total number of positive class values predicted. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP + FP}}$$

Recall is another metric that evaluates the total number of positive predictions divided by the number of positive class values. It's the intuitively the ability of the classifier to find all the positive samples.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 score is the balance between Precision and Recall. It is the weighted average of the precision and recall.

$$F1 = 2 \times \left(\frac{(precision \times recall)}{(precision + recall)} \right)$$

So, we will be using ROC_AUC as my evaluation metrics for this project.

Classification Results:

Logistic Regression (LR), K Nearest Neighbors (KNN) and Random Forest (RF) classifiers using fivefold cross validations are used to classify. The ROC_AUC scores have been mentioned below for each classifier before and after interim:

Algorithm	ROC Before Interim	ROC After Interim
LR	63.33%	65.55%
KNN	55.98%	55.55%
RF	60.89%	66.54%
LGBM	62.21%	66.96%

Final Model:

By comparing ROC and Accuracy score results of models and then we choose the best model having the best evaluation scores.

But it's also not necessary that a model with the highest accuracy is the best model as it might be not doing well predicting positive classes. A model with accuracy 80% might be better than a model with an accuracy score of 90%. This is called the Accuracy Paradox.

Out of these 5 models, LightGBM will give us better results than any. LightGBM can perform very well handling high dimensional data.

All models have its pros and cons, and our final solution model will be LightGBM.

```
accuracy: 0.7839650020729981
              precision    recall  f1-score   support

           0       0.79      0.99      0.88     54800
           1       0.52      0.04      0.07     15147

    accuracy                           0.78     69947
   macro avg       0.65      0.51      0.48     69947
weighted avg       0.73      0.78      0.70     69947

roc_auc: 0.6696002026852177
```

CONCLUSION AND SUMMARY

- In this project Vehicle loan defaulters in the first EMI for L&T have been determined. The best performing the models were ensemble-based models.
- The data seems to exactly mimic the real-life scenario which is very evident since there many zero values present which corresponds to first time customers.
- As the scope of Feature engineering was low, we tried out multiplying numerical columns, frequency encoding of categorical columns and binning of numerical columns.

REFERENCES AND BIBLIOGRAPHY

[1] Kaggle: https://www.kaggle.com/c/forest-covertype-prediction/forums/t/10693/features-engineering-benchmark

[2]https://www.chicagofed.org/~/media/publications/economic-perspectives/2008/ep-3qtr2008-part2-agarwal-etal-pdf.pdf

[3]https://www.paisabazaar.com/car-loan/6-factors-that-affect-car-loan-interest-rates/