# Chapter 1

# Introduction

## Problem Statement

The objective of the capstone project is to predict whether the customers will be Payment default in the first EMI on Vehicle Loan on due date with respect to mainly the Disbursed amount, Loan to Value percentage of the asset. The dataset is taken from Loan Default Prediction Dataset from Kaggle.

## Data Sets

The dataset provided by Kaggle is under file ‘train.csv’. The dataset comprises of 233,146 rows and 41 columns.

## The

## Data Dictionary

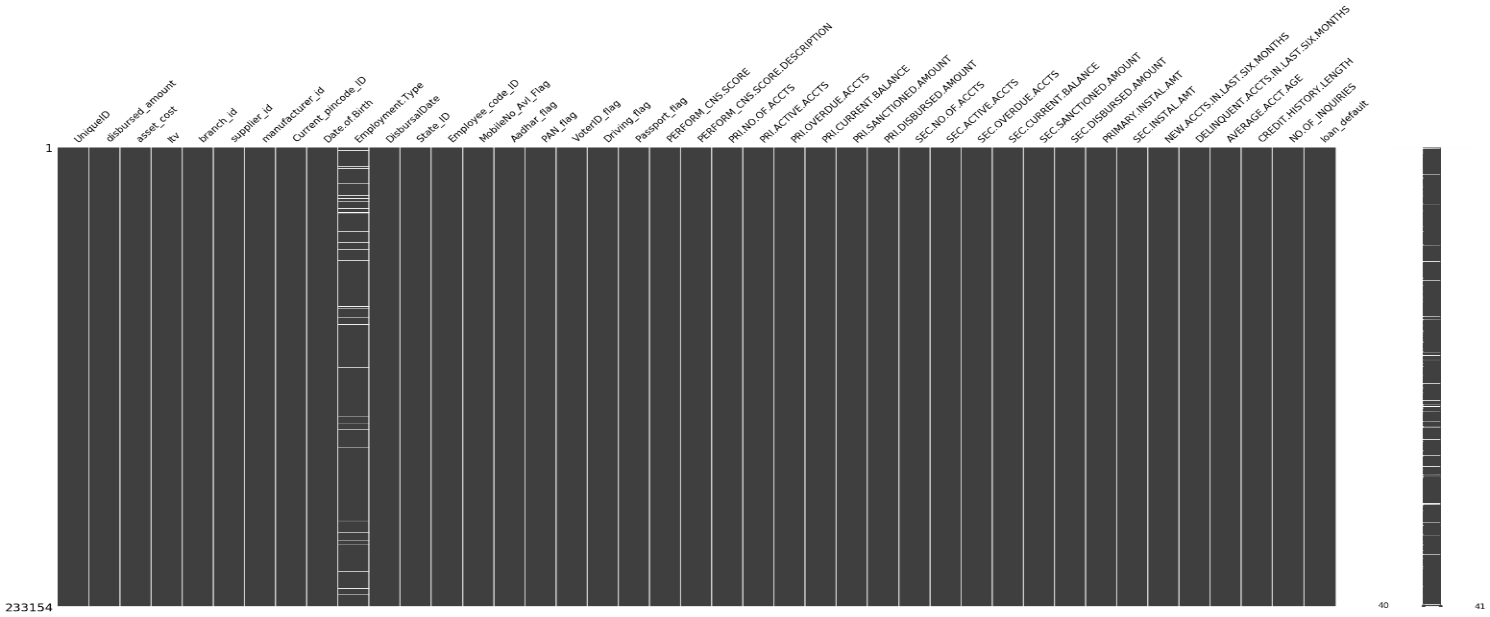
|  |  |
| --- | --- |
| **UniqueID** | Identifier for customers |
| **loan\_default** | Payment default in the first EMI on due date |
| **disbursed\_amount** | Amount of Loan disbursed |
| **asset\_cost** | Cost of the Asset |
| **ltv** | Loan to Value of the asset |
| **branch\_id** | Branch where the loan was disbursed |
| **Supplier\_id** | Vehicle Dealer where the loan was disbursed |
| **manufacturer\_id** | Vehicle manufacturer (Hero, Honda, TVS etc.) |
| **Current\_pincode** | Current pincode of the customer |
| **Date.of.Birth** | Date of birth of the customer |
| **Employment.Type** | Employment Type of the customer (Salaried/Self Employed) |
| **DisbursalDate** | Date of disbursement |
| **State\_ID** | State of disbursement |
| **Employee\_code\_ID** | Employee of the organization who logged the disbursement |
| **MobileNo\_Avl\_Flag** | if Mobile no. was shared by the customer then flagged as 1 |
| **Aadhar\_flag** | if aadhar was shared by the customer then flagged as 1 |
| **PAN\_flag** | if pan was shared by the customer then flagged as 1 |
| **VoterID\_flag** | if voter was shared by the customer then flagged as 1 |
| **Driving\_flag** | if DL was shared by the customer then flagged as 1 |
| **Passport\_flag** | if passport was shared by the customer then flagged as 1 |
| **PERFORM\_CNS.SCORE** | Bureau Score |
| **PERFORM\_CNS.SCORE.DESCRIPTION** | Bureau score description |
| **PRI.NO.OF.ACCTS** | count of total loans taken by the customer at the time of disbursement |
| **PRI.ACTIVE.ACCTS** | count of active loans taken by the customer at the time of disbursement |
| **PRI.OVERDUE.ACCTS** | count of default accounts at the time of disbursement |
| **PRI.CURRENT.BALANCE** | Principal outstanding amount of the active loans at the time of disbursement |
| **PRI.SANCTIONED.AMOUNT** | Amount that was sanctioned for all the loans at the time of disbursement |
| **PRI.DISBURSED.AMOUNT** | Amount that was disbursed for all the loans at the time of disbursement |
| **SEC.NO.OF.ACCTS** | count of total loans taken by the customer at the time of disbursement |
| **SEC.ACTIVE.ACCTS** | count of active loans taken by the customer at the time of disbursement |
| **SEC.OVERDUE.ACCTS** | count of default accounts at the time of disbursement |
| **SEC.CURRENT.BALANCE** | total Principal outstanding amount of the active loans at the time of disbursement |
| **SEC.SANCTIONED.AMOUNT** | total amount that was sanctioned for all the loans at the time of disbursement |
| **SEC.DISBURSED.AMOUNT** | total amount that was disbursed for all the loans at the time of disbursement |
| **PRIMARY.INSTAL.AMT** | EMI Amount of the primary loan |
| **SEC.INSTAL.AMT** | EMI Amount of the secondary loan |
| **NEW.ACCTS.IN.LAST.SIX.MONTHS** | New loans taken by the customer in last 6 months before the disbursment |
| **DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS** | Loans defaulted in the last 6 months |
| **AVERAGE.ACCT.AGE** | Average loan tenure |
| **CREDIT.HISTORY.LENGTH** | Time since first loan |
| **NO.OF\_INQUIRIES** | Enquries done by the customer for loans |
| **loan\_default** | Defaulters to be predicted. |

# 2. Data Preprocessing

14 Numerical columns,2 Date type columns and 25 Categorical columns

* UniqueID int64
* disbursed\_amount int64
* asset\_cost int64
* ltv float64
* branch\_id int64
* supplier\_id int64
* manufacturer\_id int64
* Current\_pincode\_ID int64
* Date.of.Birth object
* Employment.Type object
* DisbursalDate object
* State\_ID int64
* Employee\_code\_ID int64
* MobileNo\_Avl\_Flag int64
* Aadhar\_flag int64
* PAN\_flag int64
* VoterID\_flag int64
* Driving\_flag int64
* Passport\_flag int64
* PERFORM\_CNS.SCORE int64
* PERFORM\_CNS.SCORE.DESCRIPTION object
* PRI.NO.OF.ACCTS int64
* PRI.ACTIVE.ACCTS int64
* PRI.OVERDUE.ACCTS int64
* PRI.CURRENT.BALANCE int64
* PRI.SANCTIONED.AMOUNT int64
* PRI.DISBURSED.AMOUNT int64
* SEC.NO.OF.ACCTS int64
* SEC.ACTIVE.ACCTS int64
* SEC.OVERDUE.ACCTS int64
* SEC.CURRENT.BALANCE int64
* SEC.SANCTIONED.AMOUNT int64
* SEC.DISBURSED.AMOUNT int64
* PRIMARY.INSTAL.AMT int64
* SEC.INSTAL.AMT int64
* NEW.ACCTS.IN.LAST.SIX.MONTHS int64
* DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS int64
* AVERAGE.ACCT.AGE object
* CREDIT.HISTORY.LENGTH object
* NO.OF\_INQUIRIES int64
* loan\_default int64

# 2.1 Missing Values Imputation



From the Missingno matrix, The Employment Type feature is having 3.29% missing values from the dataset.

# 2.2 Features

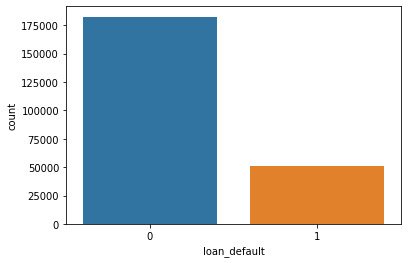
## 2.2.1 Loan Default

Loan Default column is to predict whether the customer has defaulted during the first EMI on Vehicle Loan on due date. 78.2 of the customers are not-defaulted and 21.7% of defalted.

0 182543

1 50611

Name: loan\_default, dtype: int64



## 2.2.2 Disbursed Amount

It is the amount of Loan disbursed to the customer. It is a continuous numerical column. Customers getting loan amounts below 200,000 the highest and low greater than 500,000.



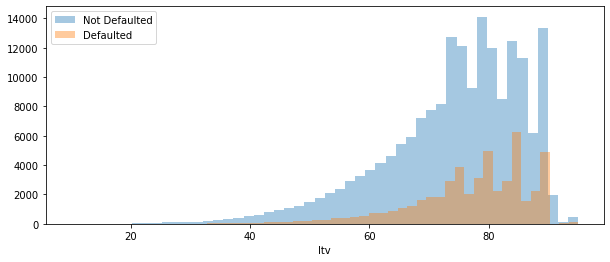
## 2.2.3 Asset Cost

Cost of the vehicle. It is a numerical continuous column to be featured.



## 2.2.4 Loan to Value

Loan to Value of the asset/vehicle.

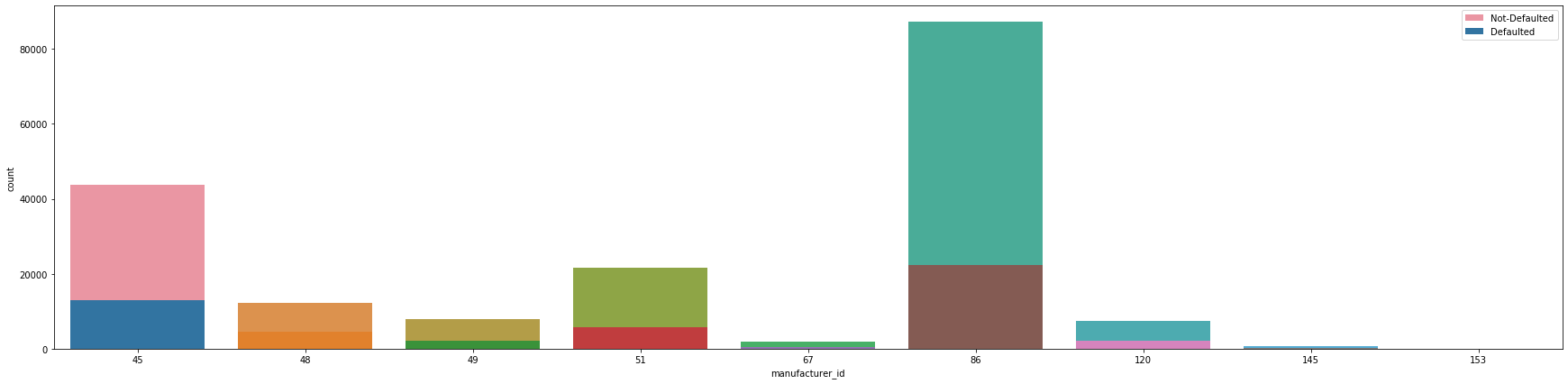


## 2.2.5 Supplier ID

Vehicle Dealer where the loan was disbursed. There are 2953 Suppliers in the current report.

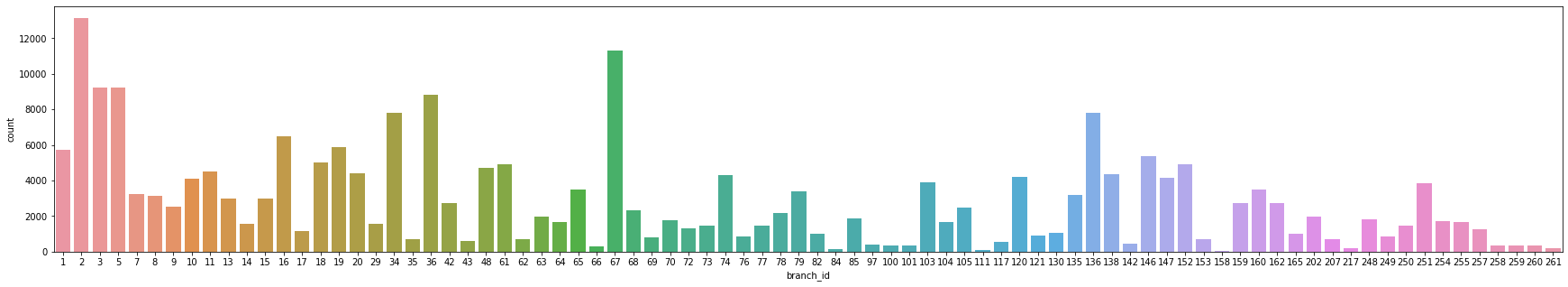
## 2.2.6 Manufacture ID

Vehicle Manufacturer i-e. TVS, Honda, Hero etc. There are 11 brands given in the label encoded form

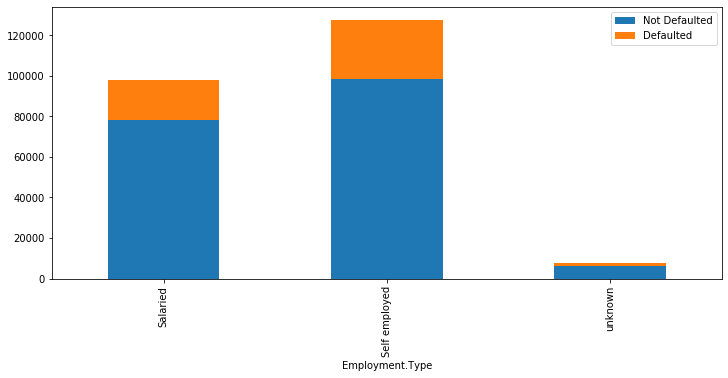


## 2.2.7 Branch ID

Branch where the loan was disbursed. 82 branches are given in the current report.



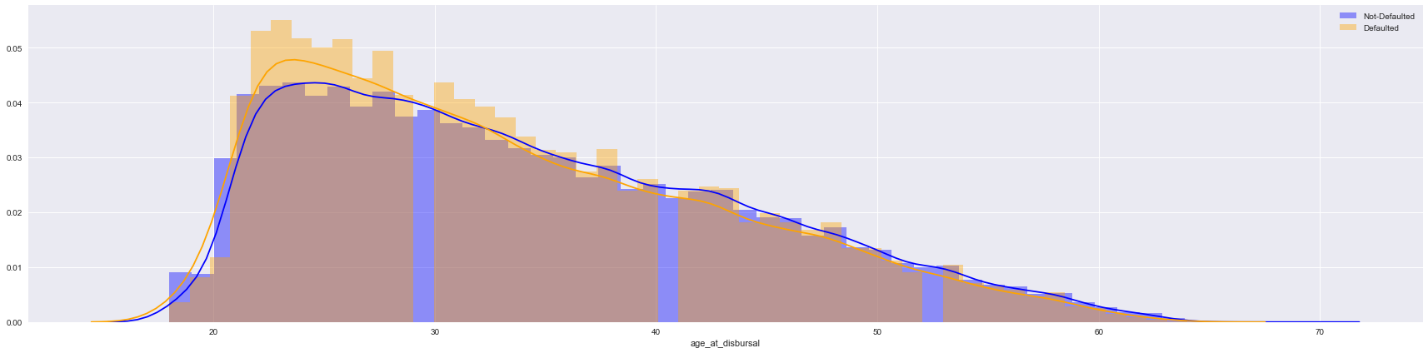
2.2.8 Employment Type

Employment Type of the customer (Salaried/Self Employed) Since There are null values in the Employment Type column. Decided to replace NaN values to ‘unknown’. 

## 2.2.10 Age of Disbursal

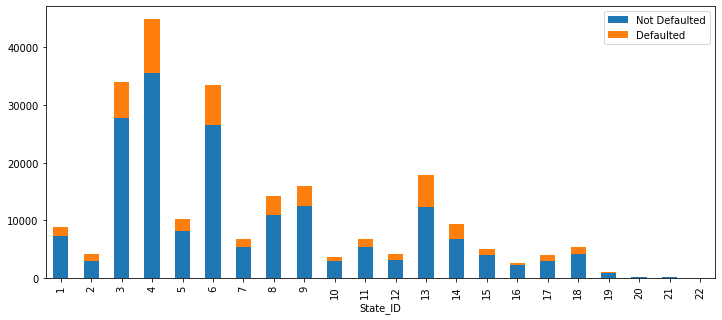
Implemented to combine Date of Birth and Disbursed Date Column to get the age of the customer when he has taken the loan.





## 2.2.11 State ID

State at which the loan had been distributed.

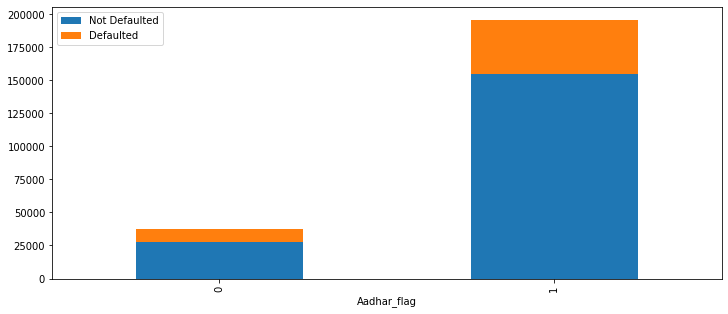


## 2.2.12 Employee ID

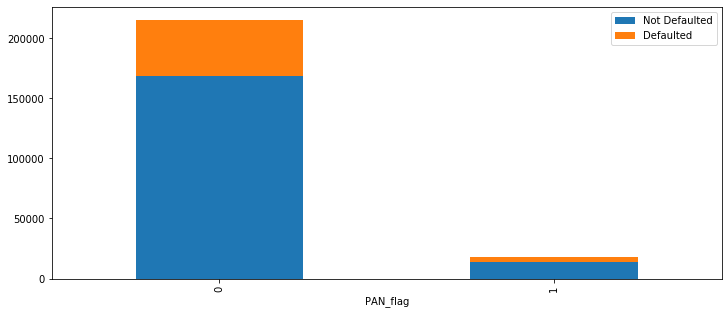
Employee who had sanctioned the transaction. It is the people who has sanctioned the loan 3270 employees

## 2.2.13 Aadhar Flag

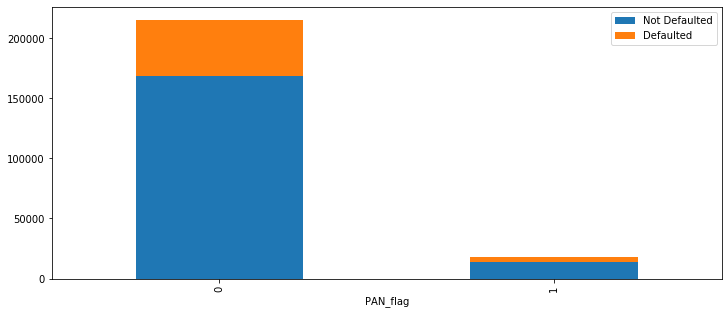
People who have been flagged for the aadar



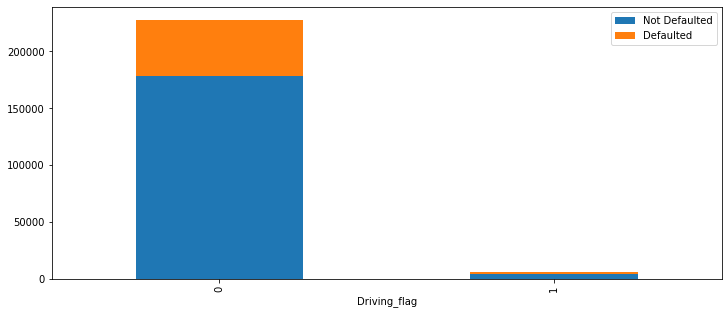
## 2.2.14 PAN Flag



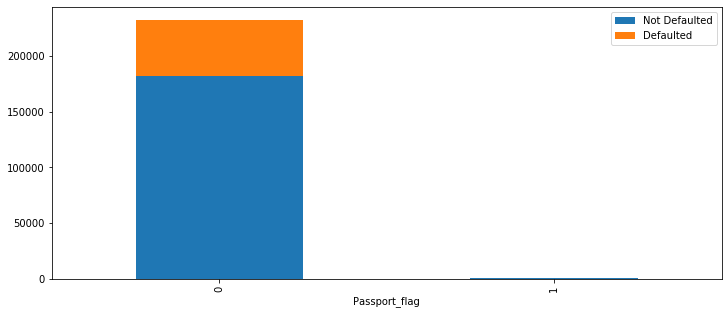
## 2.2.15 Voter ID Flag



## 2.2.16 Driving License Flag

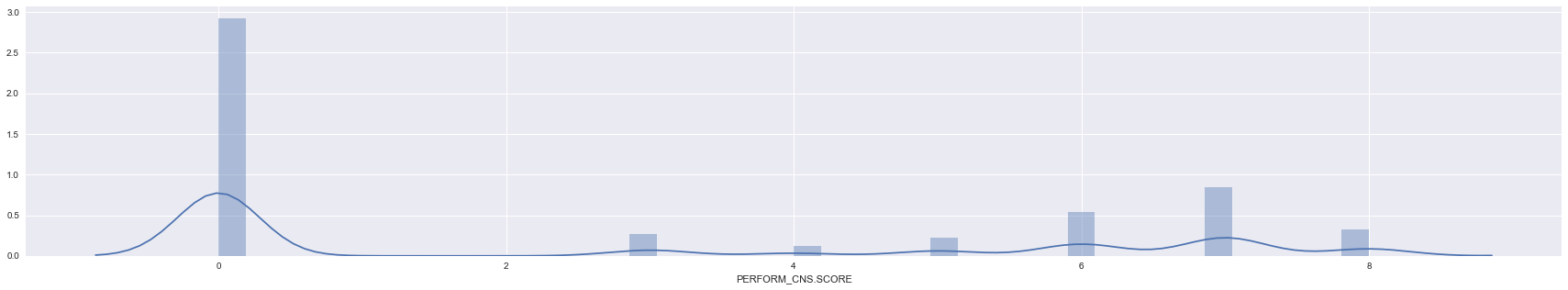


## 2.2.17 Passport Flag

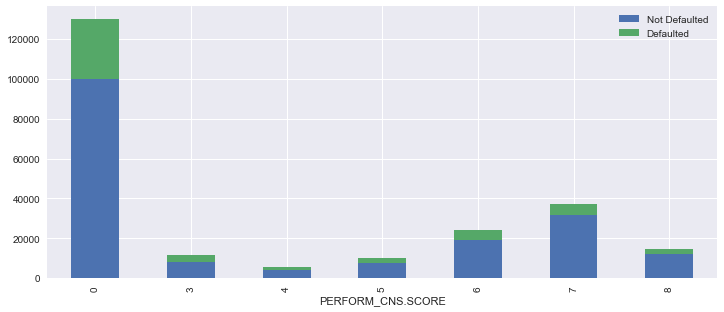


## 2.2.18 CNS Score

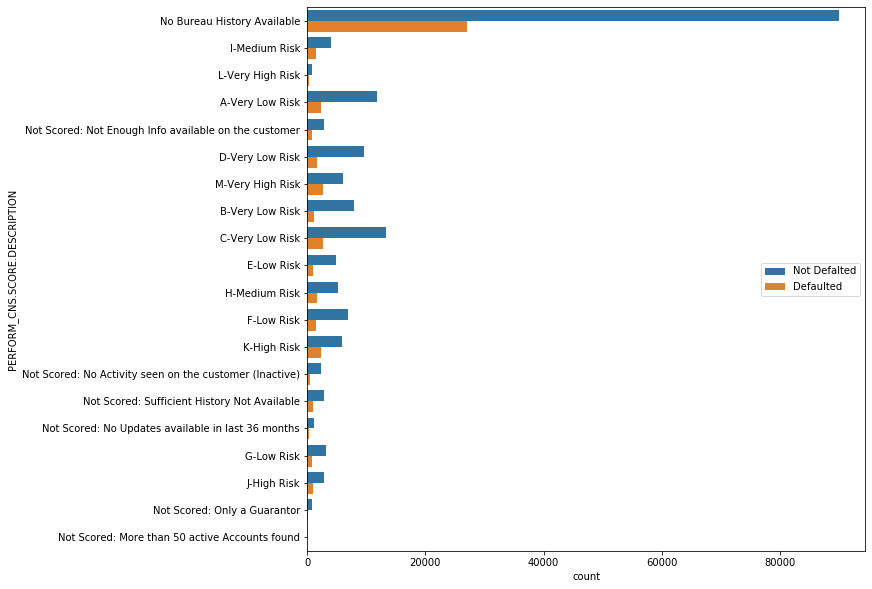
### 2.2.18.1 PERFORM CNS Score



Converted the CNS Scores in steps of 100 since the score defines his rating



### 2.2.18.2 PERFORM CNS Score Description



## 2.2.19 Primary Attributes

### 2.2.19.1 Primary Number of Accounts:

Count of total loans taken by the customer at the time of disbursement.

### 2.2.19.2 Primary Active Accounts

Count of active loans taken by the customer at the time of disbursement.

### 2.2.19.3 Primary Overdue Accounts

Count of default accounts at the time of disbursement

### 2.2.19.4 Primary Current Balance

total Principal outstanding amount of the active loans at the time of disbursement

### 2.2.19.5 Primary Sanctioned Amount

total amount that was sanctioned for all the loans at the time of disbursement

### 2.2.19.6 Primary Disbursed Amount

total amount that was disbursed for all the loans at the time of disbursement

### 2.2.19.7 Primary Installment Amount

EMI Amount of the primary loan

## 2.2.20 Secondary Attributes

### 2.2.20.1 Secondary Number of Accounts:

Count of total loans taken by the customer at the time of disbursement.

### 2.2.20.2 Secondary Active Accounts

Count of active loans taken by the customer at the time of disbursement.

### 2.2.20.3 Secondary Overdue Accounts

Count of default accounts at the time of disbursement

### 2.2.20.4 Secondary Current Balance

total Principal outstanding amount of the active loans at the time of disbursement

### 2.2.20.5 Secondary Sanctioned Amount

total amount that was sanctioned for all the loans at the time of disbursement

### 2.2.20.6 Secondary Disbursed Amount

total amount that was disbursed for all the loans at the time of disbursement

### 2.2.20.7 Secondary Installment Amount

EMI Amount of the primary loan

## 2.2.21 New Accounts in Last Six Months

New loans taken by the customer in last 6 months before the disbursement



## 2.2.22 Delinquent Accounts in Last Six Months

Loans defaulted in the last 6 months



## 2.2.23 Average Account Age

Average loan tenure

df['AVERAGE.ACCT.AGE']

0 0yrs 0mon

1 1yrs 11mon

2 0yrs 0mon

3 0yrs 8mon

4 0yrs 0mon

...

233149 1yrs 9mon

233150 0yrs 6mon

233151 0yrs 0mon

233152 0yrs 0mon

233153 0yrs 0mon

Name: AVERAGE.ACCT.AGE, Length: 233154, dtype: object

df['AVERAGE.ACCT.AGE']=df['AVERAGE.ACCT.AGE'].apply(lambda x:(re.sub('[a-z]','',x)).split())

df['AVERAGE.ACCT.AGE']=df['AVERAGE.ACCT.AGE'].apply(lambda x:int(x[0])\*12+int(x[1]))

df['AVERAGE.ACCT.AGE']

0 0

1 23

2 0

3 8

4 0

..

233149 21

233150 6

233151 0

233152 0

233153 0

Name: AVERAGE.ACCT.AGE, Length: 233154, dtype: int64



## 2.2.24 Credit History Length

Duration of the loan

df['CREDIT.HISTORY.LENGTH']

0 0yrs 0mon

1 1yrs 11mon

2 0yrs 0mon

3 1yrs 3mon

4 0yrs 0mon

...

233149 3yrs 3mon

233150 0yrs 6mon

233151 0yrs 0mon

233152 0yrs 0mon

233153 0yrs 0mon

Name: CREDIT.HISTORY.LENGTH, Length: 233154, dtype: object

Changing years and month format to months

df['CREDIT.HISTORY.LENGTH']=df['CREDIT.HISTORY.LENGTH'].apply(lambda x:

(re.sub('[a-z]','',x)).split())

df['CREDIT.HISTORY.LENGTH']=df['CREDIT.HISTORY.LENGTH'].apply(lambda x:

int(x[0])\*12+int(x[1]))

df['CREDIT.HISTORY.LENGTH']

0 0

1 23

2 0

3 15

4 0

..

233149 39

233150 6

233151 0

233152 0

233153 0

Name: CREDIT.HISTORY.LENGTH, Length: 233154, dtype: int64



## 2.2.25 Number of Inquires

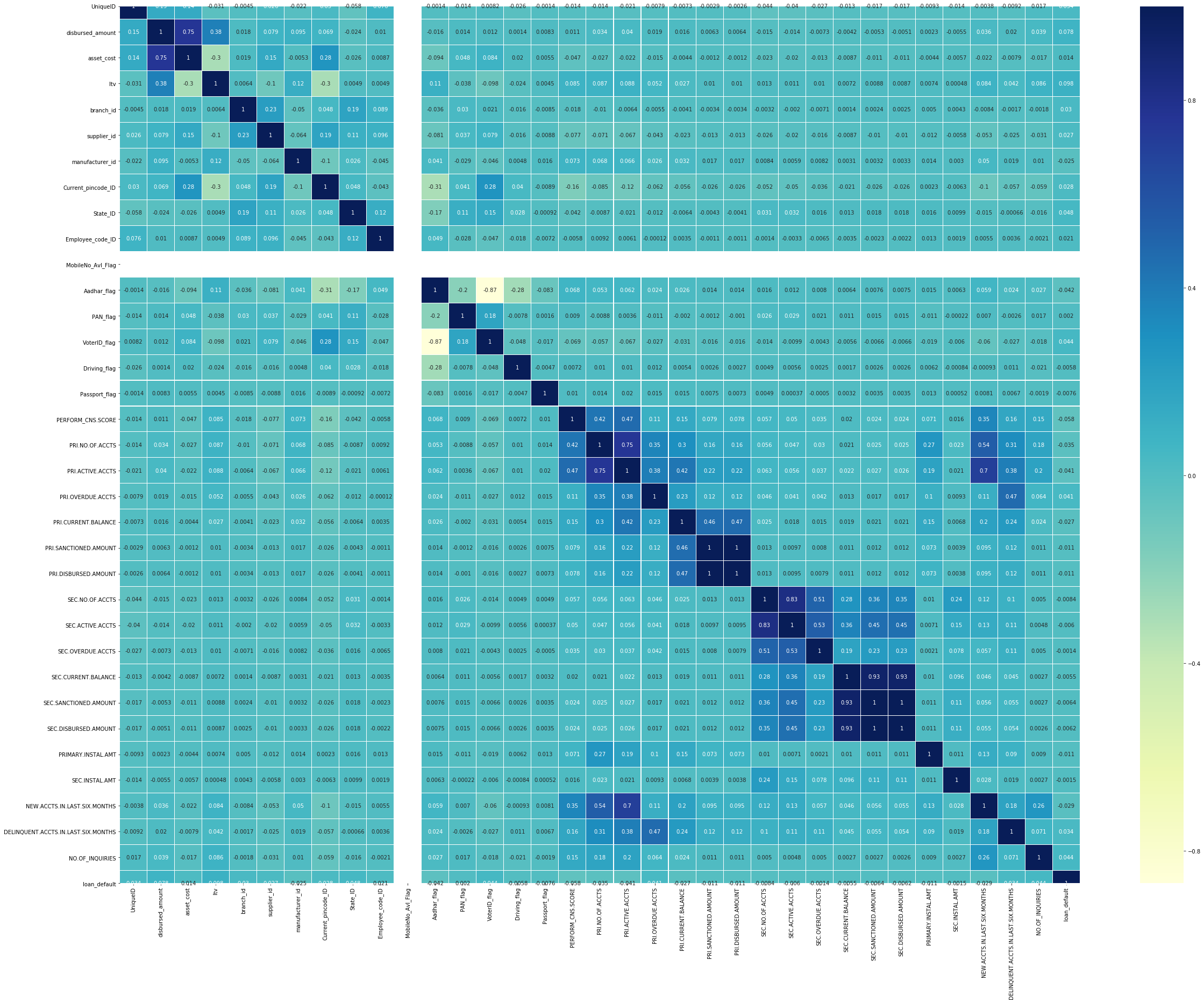
Enquiries

done by the customer for loans



# 3. Feature Engineering

## 3.1 Correlation Plot

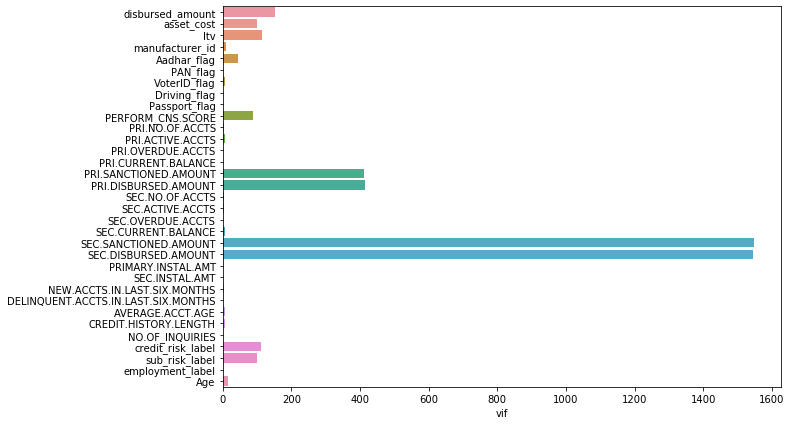


Since there is no Correlation in Mobile Avl Flag and Unique in the dataset. Removing Mobile Avl Flag and Unique ID from the dataset.

## 3.2 Multicollinearity Check

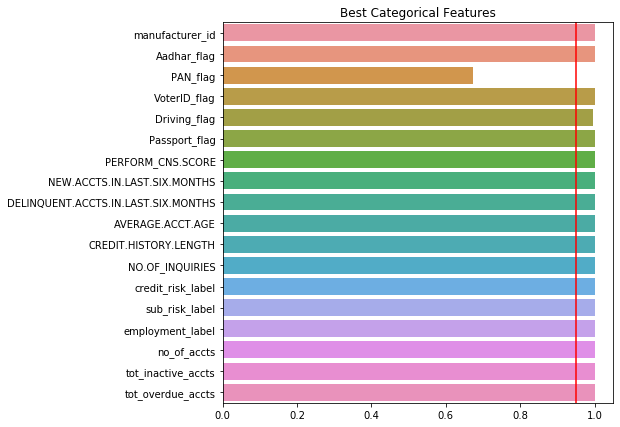
Checked multicollinearity with the help of variance influence factor(vif).

disbursed\_amount 8.74561717585509e-309  
asset\_cost 5.716223071536896e-12  
PRI.SANCTIONED.AMOUNT 4.798158421546997e-08  
SEC.NO.OF.ACCTS 5.1490255376949666e-05  
PRI.NO.OF.ACCTS 9.576575137572993e-66  
PRI.DISBURSED.AMOUNT 7.176942237800462e-08  
PRI.ACTIVE.ACCTS 3.448627479875517e-89  
PRI.OVERDUE.ACCTS 9.138488408377107e-87  
SEC.CURRENT.BALANCE 0.0075643427363124875  
SEC.SANCTIONED.AMOUNT 0.002153062273491789  
SEC.OVERDUE.ACCTS 0.5081054926877384  
SEC.DISBURSED.AMOUNT 0.0025523226185338705  
PRIMARY.INSTAL.AMT 2.958254960232989e-07  
SEC.INSTAL.AMT 0.4546434321302706  
NEW.ACCTS.IN.LAST.SIX.MONTHS 9.30229371021266e-46  
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS 3.2892517686894386e-62  
AVERAGE.ACCT.AGE 5.261091482095756e-33  
CREDIT.HISTORY.LENGTH 4.6500173864982836e-92  
NO.OF\_INQUIRIES 7.912566786376203e-99



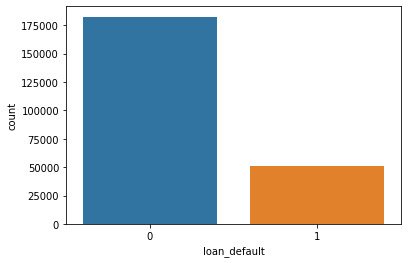
## 3.3 Statistical Test for Numerical Columns

## 3.4 Statistical Test for Categorical Columns



## 3.5 Data Imbalance

Target Variable



Not-Defaulters 78.2%

Defaulters 21.3%

In the 233,546 rows in the dataset, which is highly imbalance. The model cannot will make a wrong prediction because of the data.

## 3.5.1 Under-sampling

not\_default = df[df.loan\_default==0]

default = df[df.loan\_default==1]

not\_default\_downsampled = resample(not\_default,

replace = True, # sample without replacement

n\_samples = len(default), # match minority n

random\_state = 0)

downsampled = pd.concat([not\_default\_downsampled, default])

downsampled['loan\_default'].value\_counts()

1 50611

0 50611

Name: loan\_default, dtype: int64

From the 233,456 rows in the dataset, we are losing over 50% of the data, so decided not to implement Downsampling

## 3.5.2 Synthetic Minority Over-sampling Technique (SMOTE)

from sklearn.utils import resample

from imblearn.over\_sampling import SMOTE

print("X shape",X.shape)

print('y shape',y.shape)

X shape (233154, 39)

y shape (233154,)

sm = SMOTE(random\_state=0)

X\_smote,y\_smote = sm.fit\_sample(X,y)

print("X shape",X\_smote.shape)

print('y shape',y\_smote.shape)

X shape (365086, 39)

y shape (365086,)