

# Vehicle Loan Default Prediction

## Team Details

Krishnamurthy S

Krishnaraj Palanychamy

Prabhakaran S

Pravin Kumar S

SR Abilash

Vishwanath Kannan

# Introduction

- People avail vehicle loan from banks to buy their dream cars. Car loans have taken off in India witnessing an increase in growth of 18-20% which is a huge increase in 2019.
- Bank and vehicle finance companies are making this dream come true by providing the vehicle loan facility.
- Financing a vehicle involves a lot of technicalities like the kind of vehicle to be financed, the route on which the vehicle will be plying, the operating expenses of the customer, etc.
- It is also being influenced by processing fee, loan clearance time, requirement of documentation and methodology being followed in computation of interest.

## Problem Definition

- The objective of this project is to predict whether the customer will default based on these critical features.
  - To find out consumers' awareness about vehicle finance activities
  - To identify reasons for availing of vehicle finance
  - To find out problem faced by consumers in availing of vehicle finance
- Finding out the critical features that to help him/the company to evaluate the probability of default of the customer, as well as prevent losing out potential customers otherwise lost.
- Implementing a model that will help Banks to provide a better understanding about the customers and their status
- Reducing the losses suffered by the banks by availing loans to potential defaulters.

## Data Dictionary:

- The Dataset contains 233154 Rows and 41 Columns
- 14 Numerical columns, 2 Datatype columns and 25 Categorical columns
- The Only column with missing values is Employment Type. It has 7661 missing values which is about 3.3% of the data
- Current Balance has 1,50,000 Entries as zero, out of which 70,000 entries have applied for their first loan

There are few possible columns which can be requested to better predict the data , namely –

1. Interest Rate : The Interest rate which is charged for the current Loan.
2. Total Previous loan : Total Amount of loan which was availed by the customer before taking the current loan.
3. Age of the vehicle : zero for new vehicle and no of years if it's a used vehicle.

## Description of Columns:

### Numerical

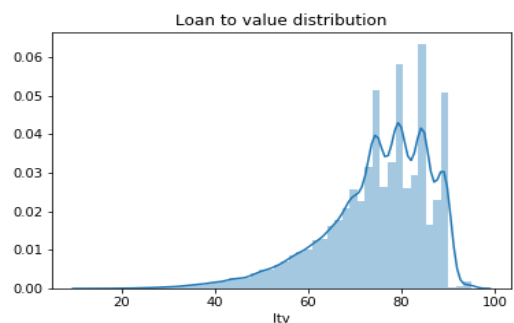
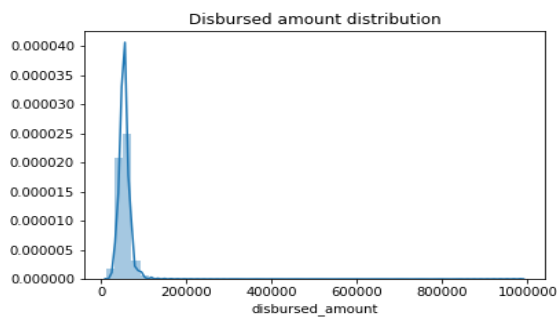
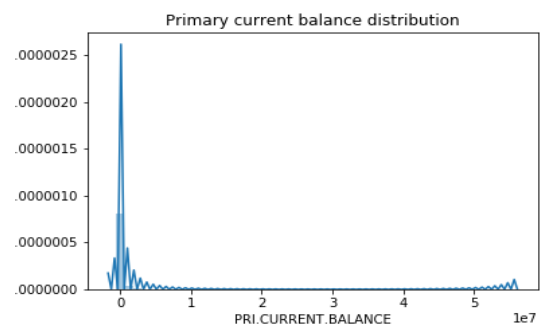
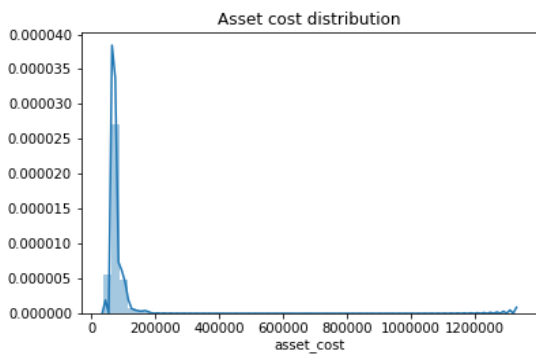
Column Name	Description
Disbursed Amount	Amount of Loan disbursed
Asset Cost	Cost of the Vehicle
LTV	Loan to Asset Value Ratio
Primary Current Balance	Total Outstanding Amount of Loan at Disbursement
Installment Amount	EMI Amount of the Loan
Age at disbursal	Age of the customer when loan was disbursed
Delinquent Accounts in last six months	Loan Defaulted in last six months

# Categorical

Column Name	Description
Number of accounts	Total number of accounts of the customer
Active Accounts	Loan accounts at time of disbursements
Overdue Accounts	Count of overdue accounts at time of disbursement
New accounts in last 6 months	New loan taken by customer in last 6 months
No of inquiries	Enquiries done by customer for the loan
Average Loan tenure	Average period of loan in months
Time_since_1st_loan	Time since first loan in months
Aadhar Flag	If Aadhar was shared-Flag 1
Passport Flag	If passport was shared-Flag 1
PAN Flag	If PAN was shared-Flag 1

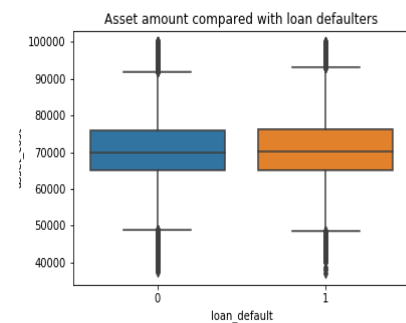
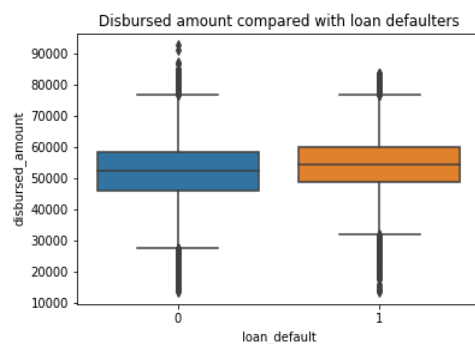
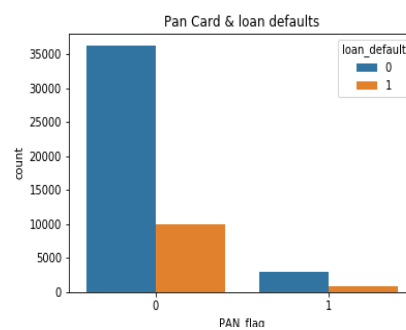
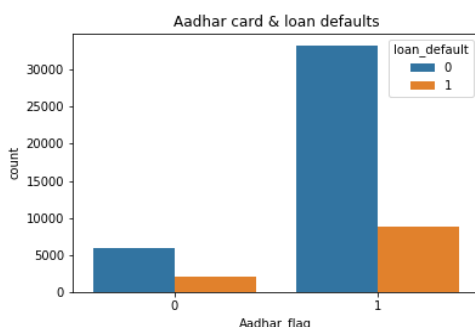
## Distribution plot on numerical features:

Here numerical features are used for distribution plot. Most of the numerical variable like Asset cost distribution, primary CB, disbursed amount follows right skewed distribution whereas LTV follows left skewed distribution. Because in case if the borrower defaults on the loan the lender can reposes the collateral and collect the money by selling it off. They can recover the losses for defaulted loan.



## Plots on various features:

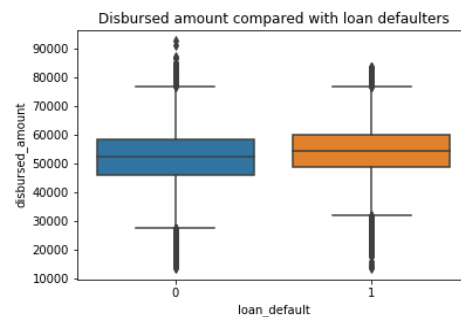
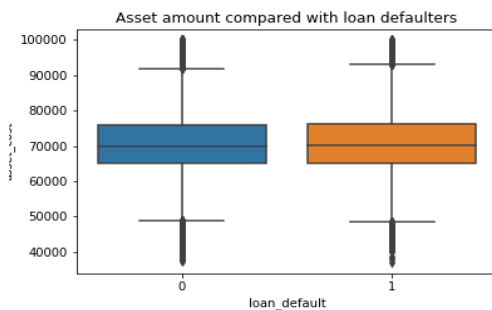
Count plot for employment type is compared with the loan defaulters here where most of the self employed are a non-loan defaulters. Aadhar and pan details compared with loan defaulters. Person with Aadhar are a loan defaulter whereas its just opposite to Pan details. Significant features like disbursed amount and asset amount plotted with loan defaulters, here the distribution is almost same shows no correlation with the target variable.





## Box plot on significant feature:

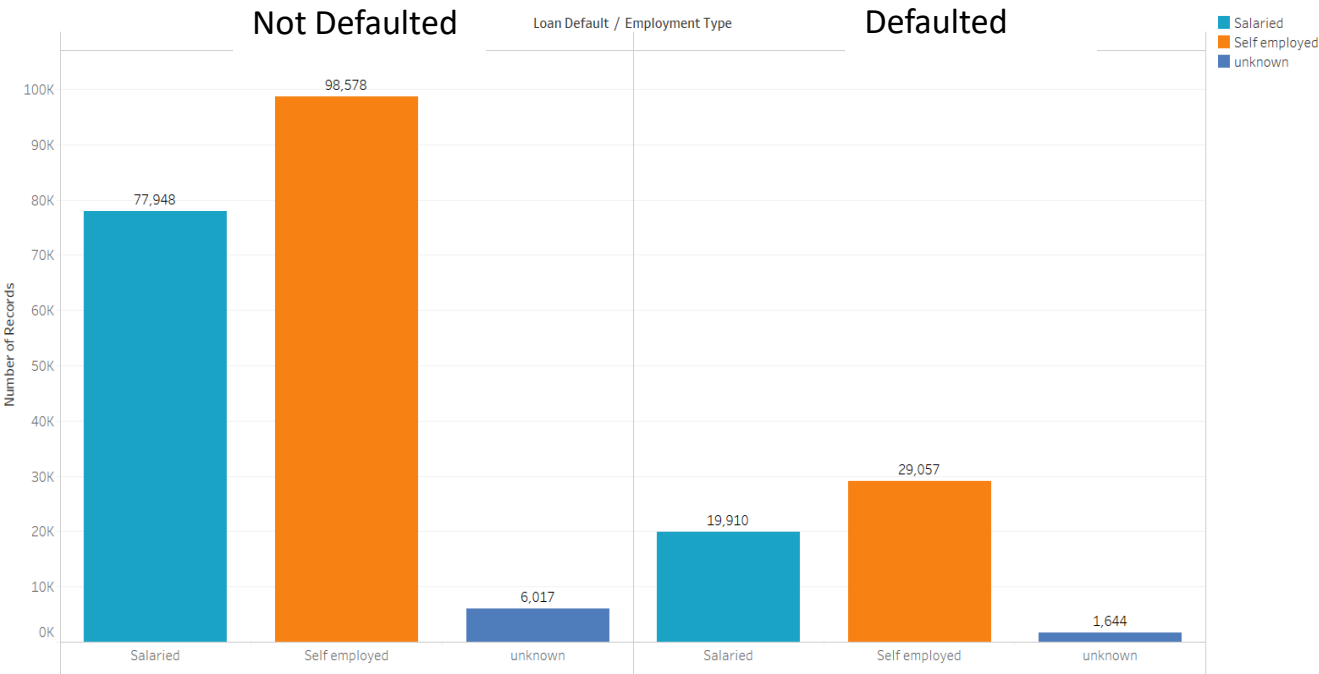
Significant features like disbursed amount and asset amount plotted with loan defaulters, here the distribution is almost same shows no correlation with the target variable.



## Multi collinearity check:

- Multi collinearity check is performed.
- Features like primary disbursed amount, secondary sanctioned amount and secondary disbursed amount is highly correlated with target variable.

# Loan Default from Employment Type



## Statistical significance of variables

Numerical	Categorical
Disbursed amount	No. of accounts
Asset cost	Active accounts
loan-to-value ratio	Overdue accounts
current-balance	New accounts in last 6 months
installment-amount	No of inquiries
Primary current balance	Avg Loan tenure
Age_at_disbursal	Time_since_1st_loan
	Aadhar Flag

# Feature Engineering

## 1. Transformations :

Yes . Since the data is not normally distributed (Highly right skewed), We tried Log transform, Sqrt transform.

## 2. Scaling the data :

Yes, the data has been scaled.

## 3. Dimensionality reduction :

Since there is no multicollinearity in our data presently(removed after checking VIF), We haven't done PCA.

# Assumptions

- **General Assumptions:**
  1. No redundant data.
  2. Absence of Multicollinearity
- **Naive Bayes** - Attributes are conditionally independent.
- Decision Tree, Random Forest, KNN – No assumptions

## Reference

<https://github.com/xavierigneous/Vehicle-Loan-Default-Prediction>

Thank You