# 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

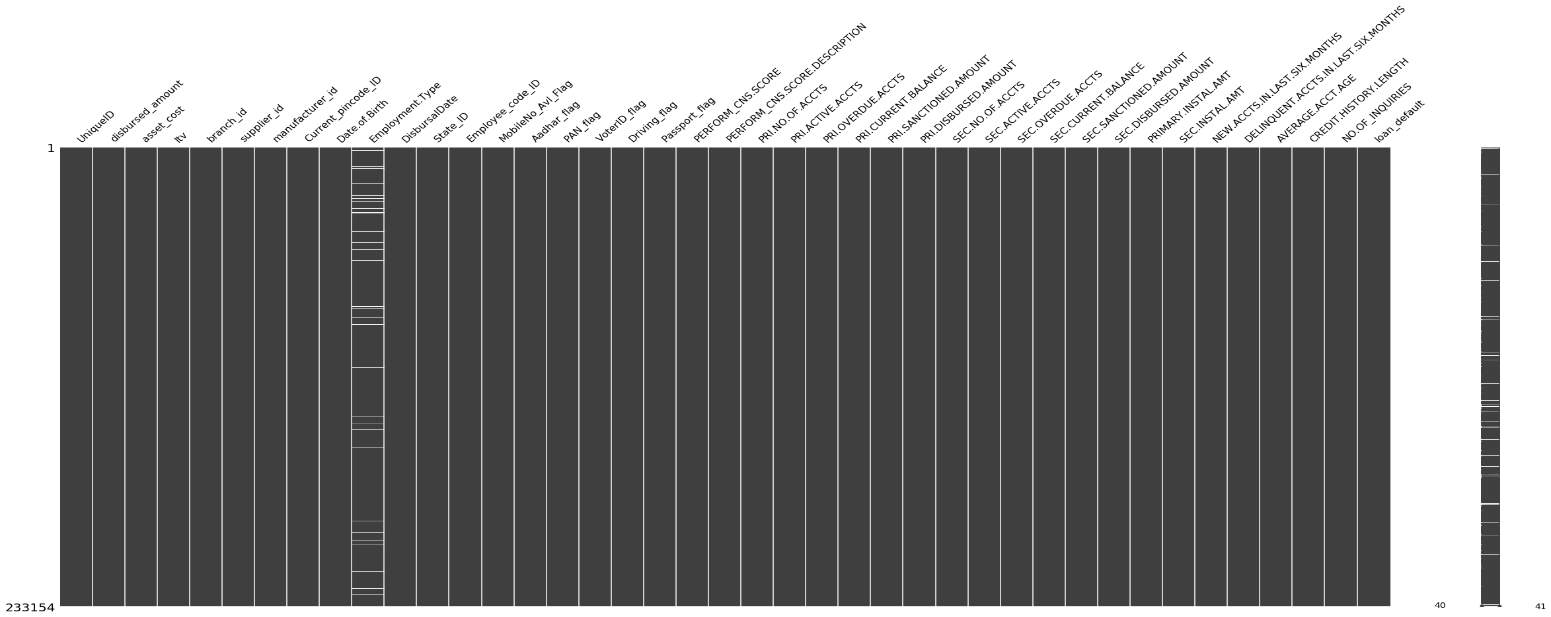
## Data Dictionary

|  |  |
| --- | --- |
| **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 Datetype columns and 25 Categorical columns

# Missing Values Imputation



From the Missingno matrix , The Employment Type column is having