

BFSI: CREDIT RISK ASSIGNMENT

Prepared By: Viraj Bhosale

INDEX :

1. Objective

2. Background

3. Data Analysis

- Pre Processing of Data

- EDA

- Model Building

- Interpreting the Results

- Recommendations

OBJECTIVE

- The objective is to build a statistical model to estimate borrowers' **Loss Given Default (LGD)**

$$\text{LGD} = \frac{\text{Loan Amount} - (\text{Collateral value} + \text{Sum of Repayments})}{\text{Loan Amount}}$$

BACKGROUND

- ❖ Credit risk analytics in the context of the banking sector and model a common metric used for estimating the expected credit loss (ECL)
- ❖ ECL method is used for provisioning the capital buffer to protect banks against possible default of the customers.

Expected credit loss = Exposure at default x Probability of Default x Loss given default

- ❖ The **loss given default (LGD)** is a measure of the amount of loss that a bank is expected to incur in the event of a default by a borrower.

DATA SOURCES

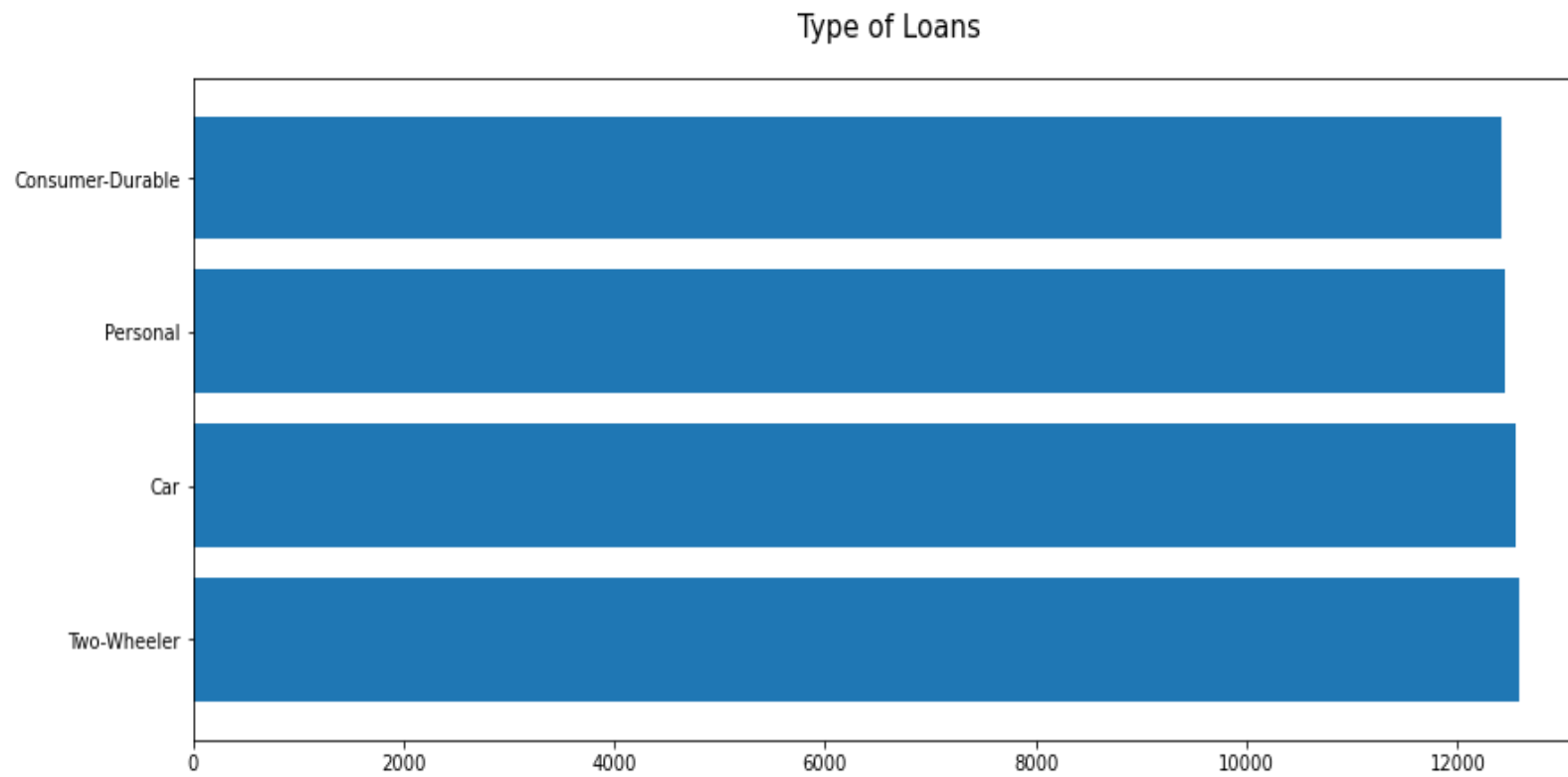
- ❖ Used 3 Data sets for model Building
 - ❖ The `main_loan_base` data set contains information about loan accounts and other relevant information for the corresponding borrowers.
 - ❖ The `repayment_base` data set contains information about the repayments received by the banks in the form of EMIs or through other collection efforts.
 - ❖ The `monthly_balance_base` contains the information pertaining to the monthly balance statements in the borrower's accounts.

PRE PROCESSING OF DATA

- ❖ For each data set converted Data types if necessary
- ❖ Null values are handled using deletion and imputation techniques. As well duplicate values are removed from data sets.
- ❖ Merging the data sets and created target variable(LGD)
- ❖ Exploratory Data Analysis has been performed
- ❖ Variable Transformation
- ❖ Dummy Encoding
- ❖ Scaling using Standard Scaler

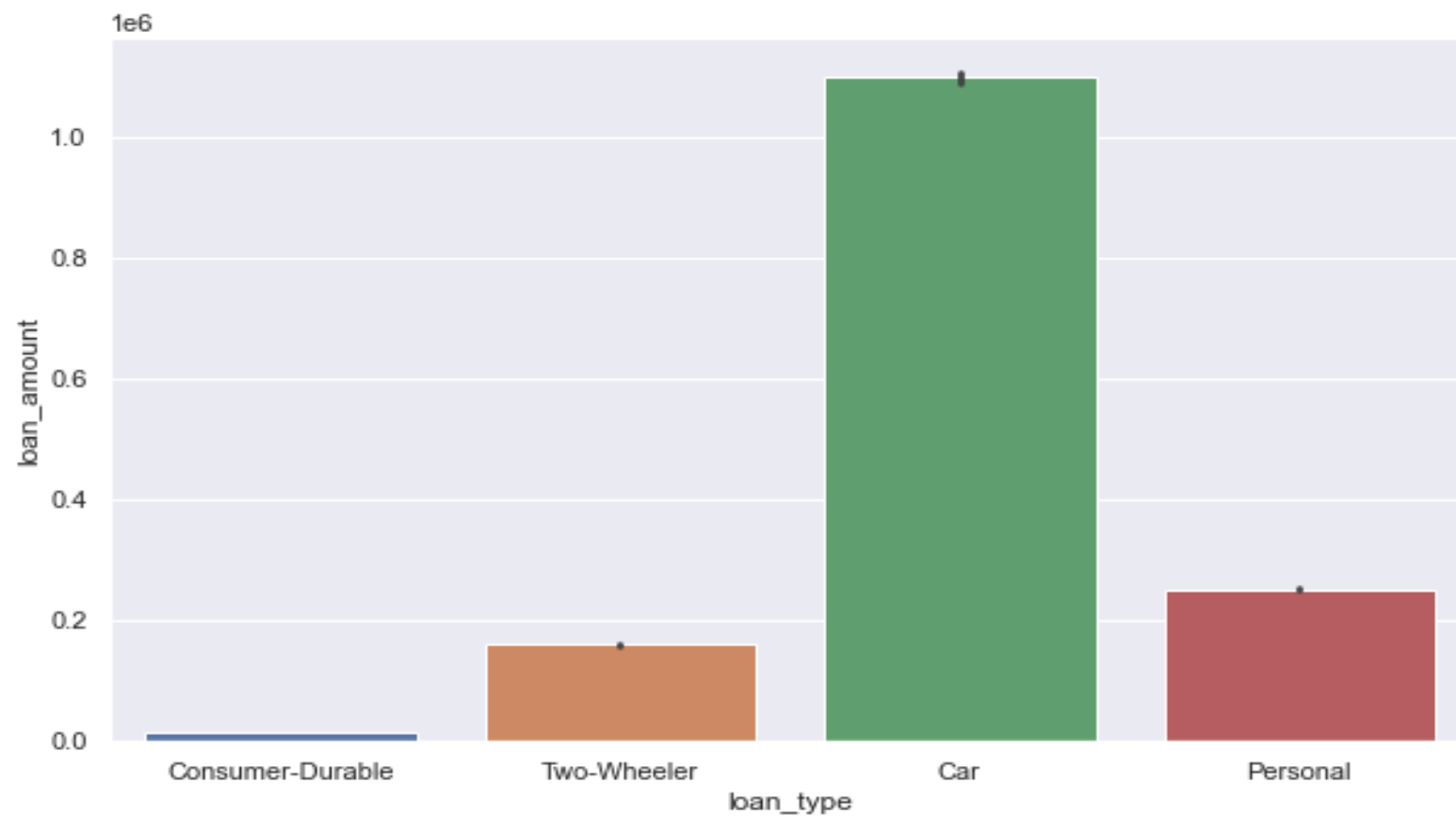
EDA

Number of loans in Two- wheeler is higher than all others.



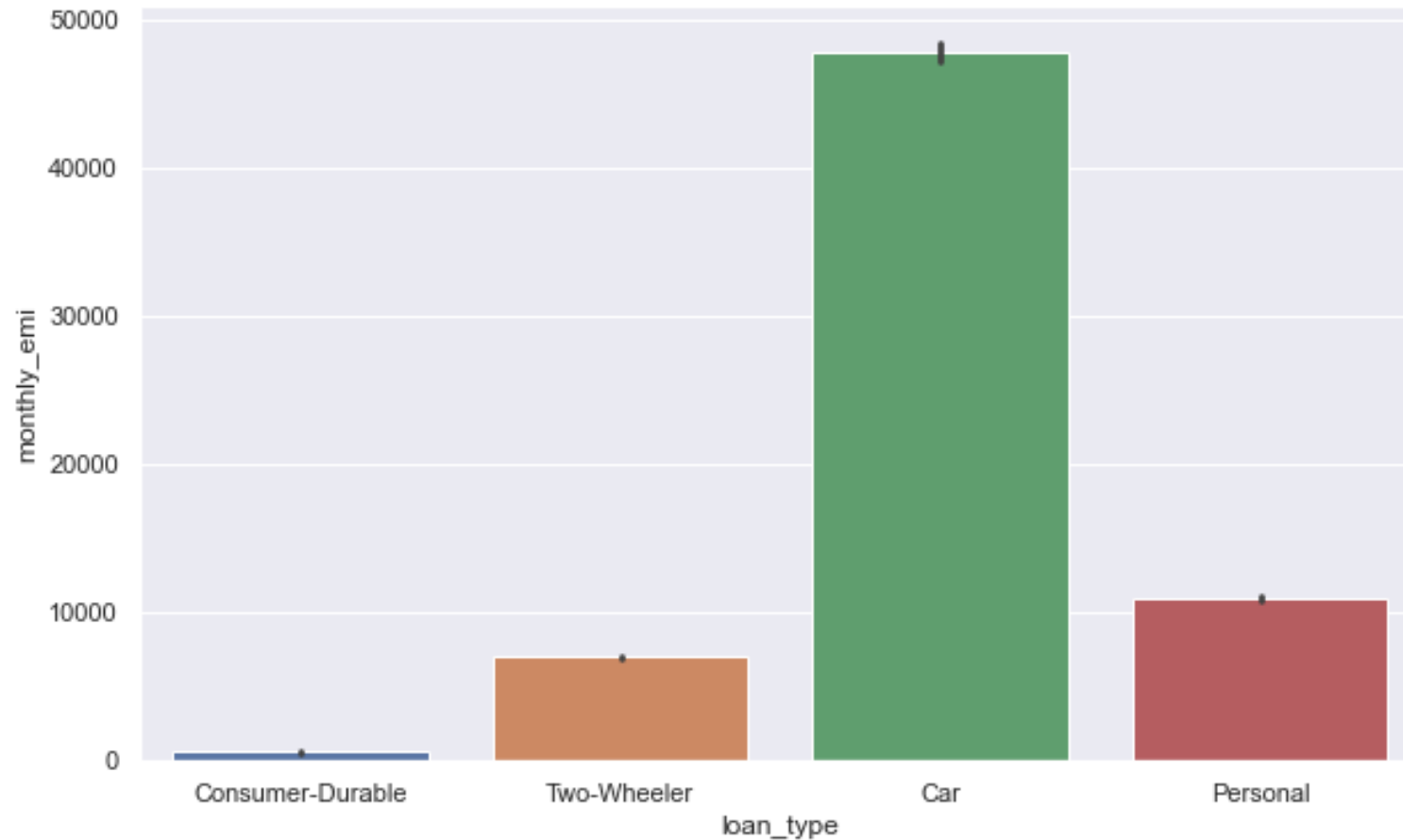
EDA

But, the loan amount of car loan is the highest.



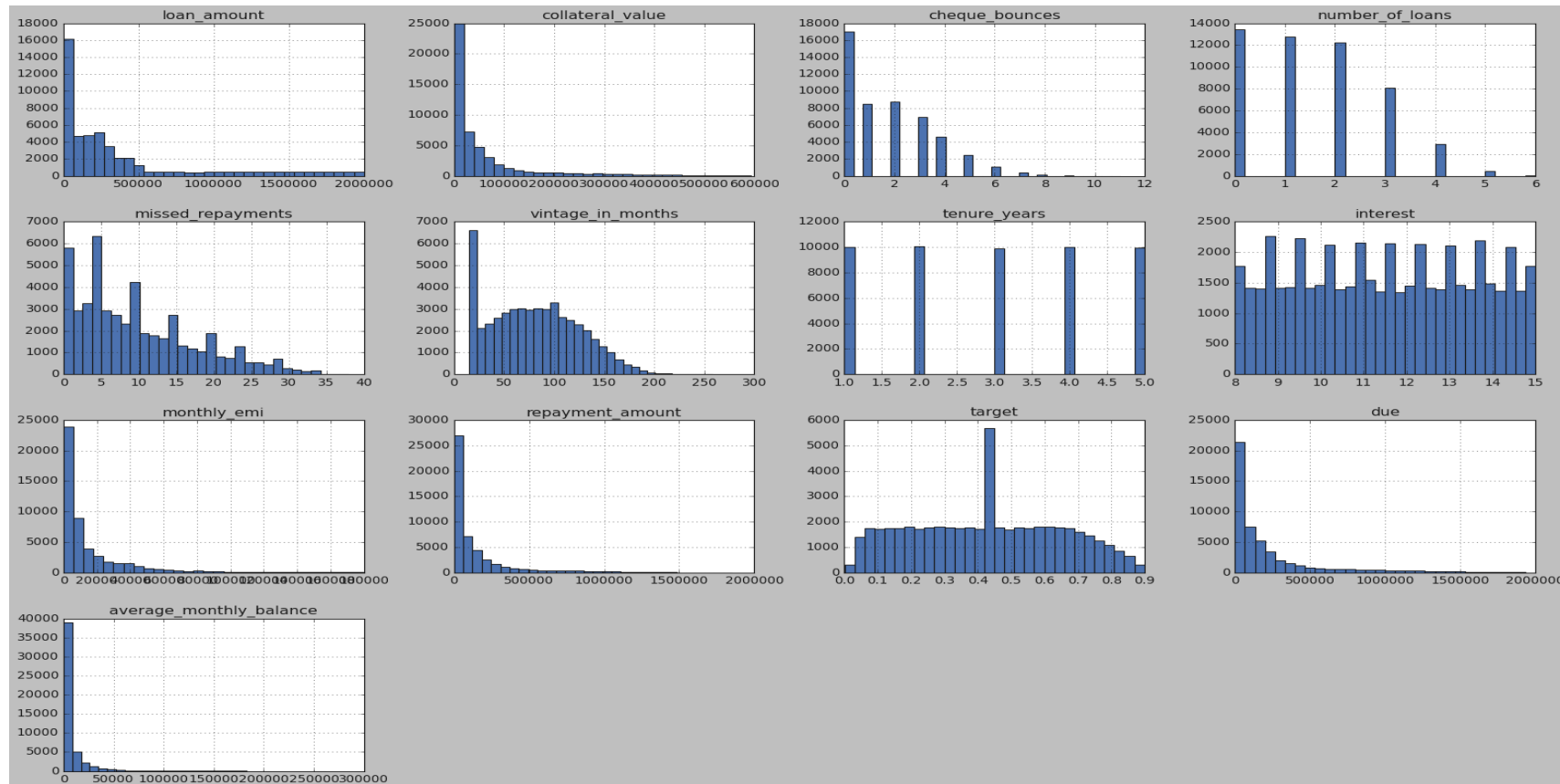
Monthly EMI

Car loan is much higher compared to other loans.



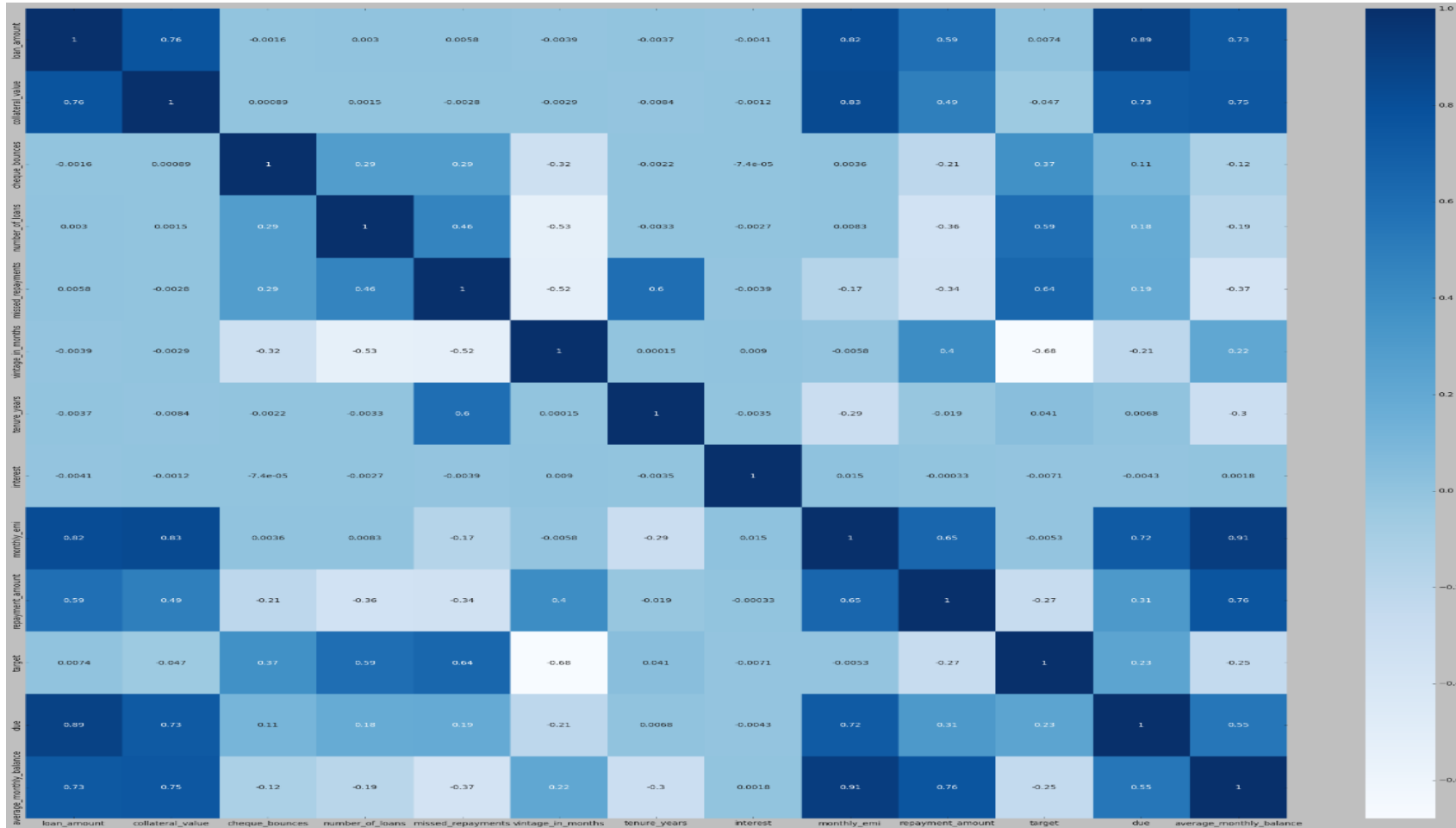
GRAPHICAL PRSENTATION:

Plotted histograms for the numerical columns to understand the distribution of data.



HEATMAP PRESENTATION:

Created Heatmap to understand the correlation between the variables.



STEPS PERFORMED:

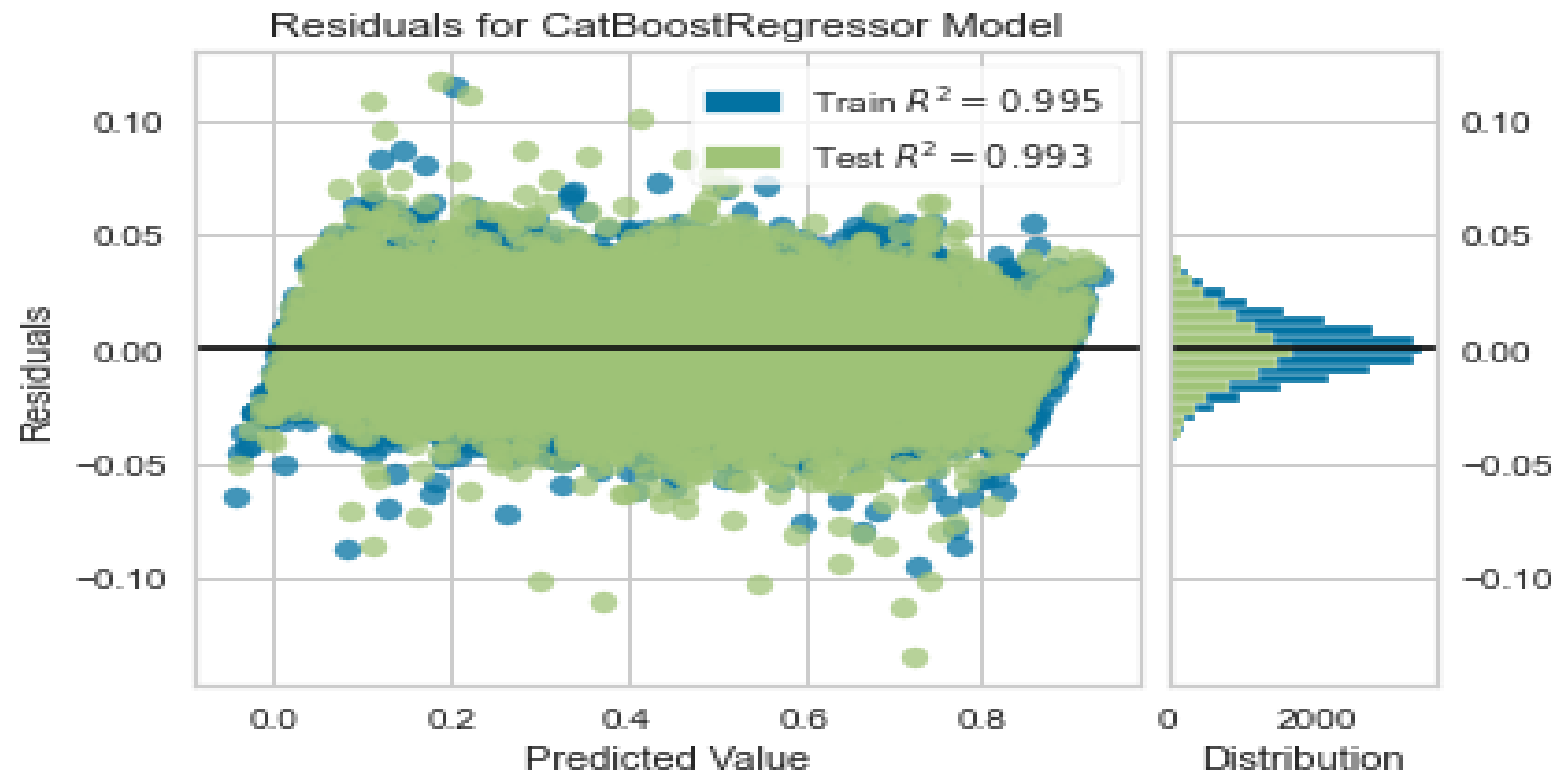
- ❖ Used Power transformation to make numerical variables Normally distributed
- ❖ Dropped unnecessary columns for model building.
- ❖ Used One-hot encoding technique and created dummy variable for necessary categorical variables.

MODEL BUILDING

- ❖ Used various models like Multiple Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, Adaboost Regressor, Elastic Net : Hybrid Regularized Model, LightGBM for model building.
- ❖ Used R Squared as a performance metrics.
- ❖ XGBoost has given us 99.5% R squared on test data across the models.

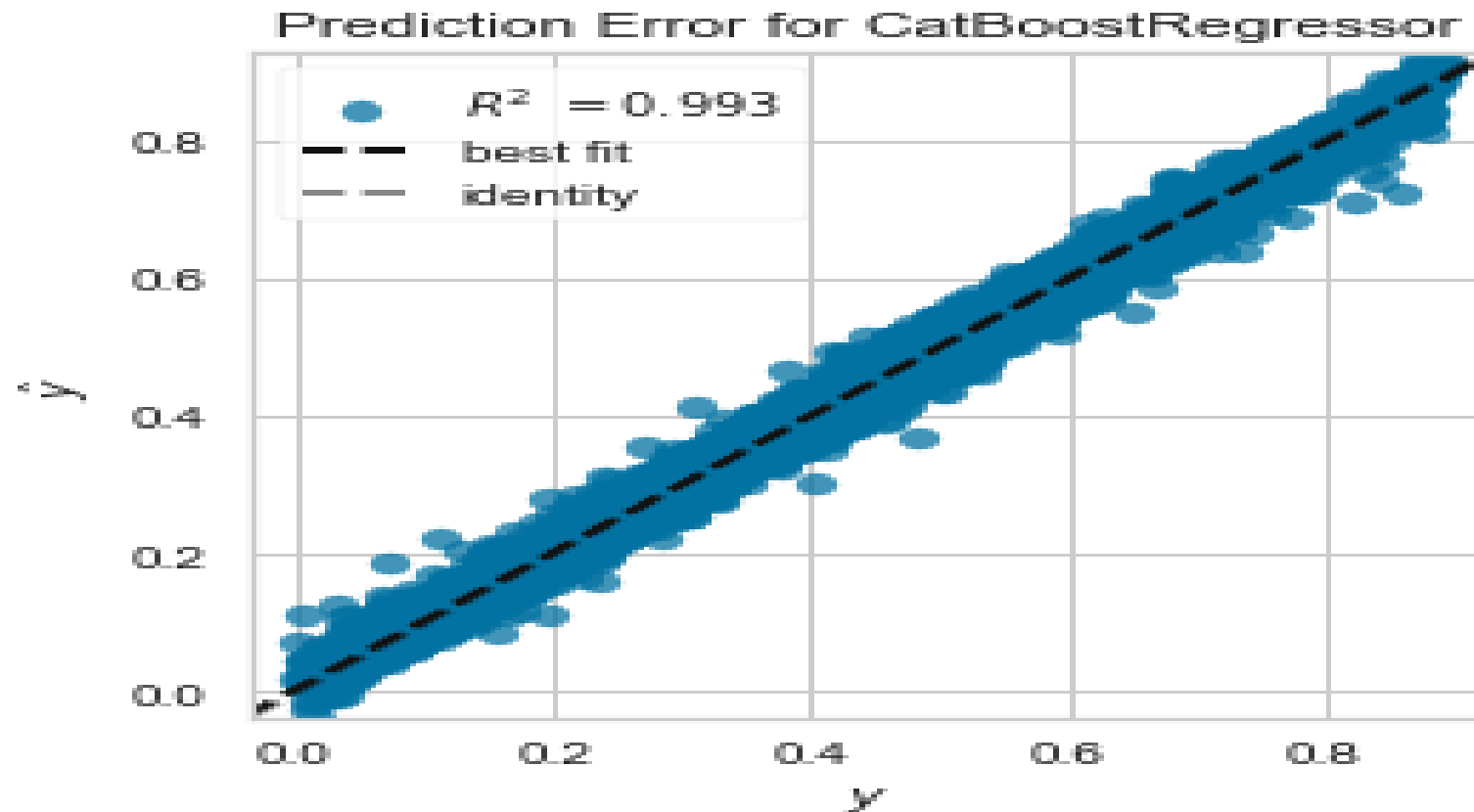
REGRESSION INTERPRETATION

❖ Residual plot of the finest model.



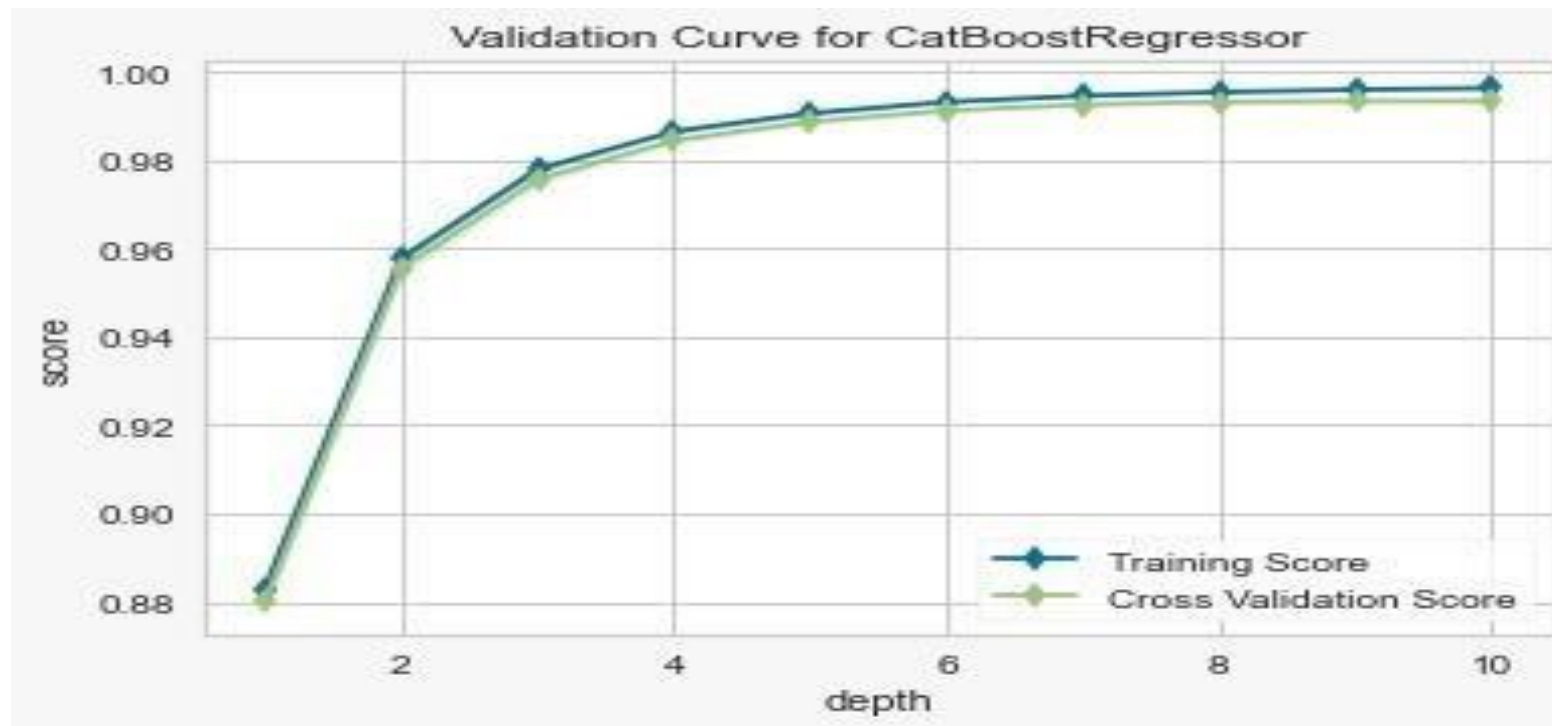
REGRESSION INTERPRETATION

- ❖ Best fit line corresponding the prediction error.



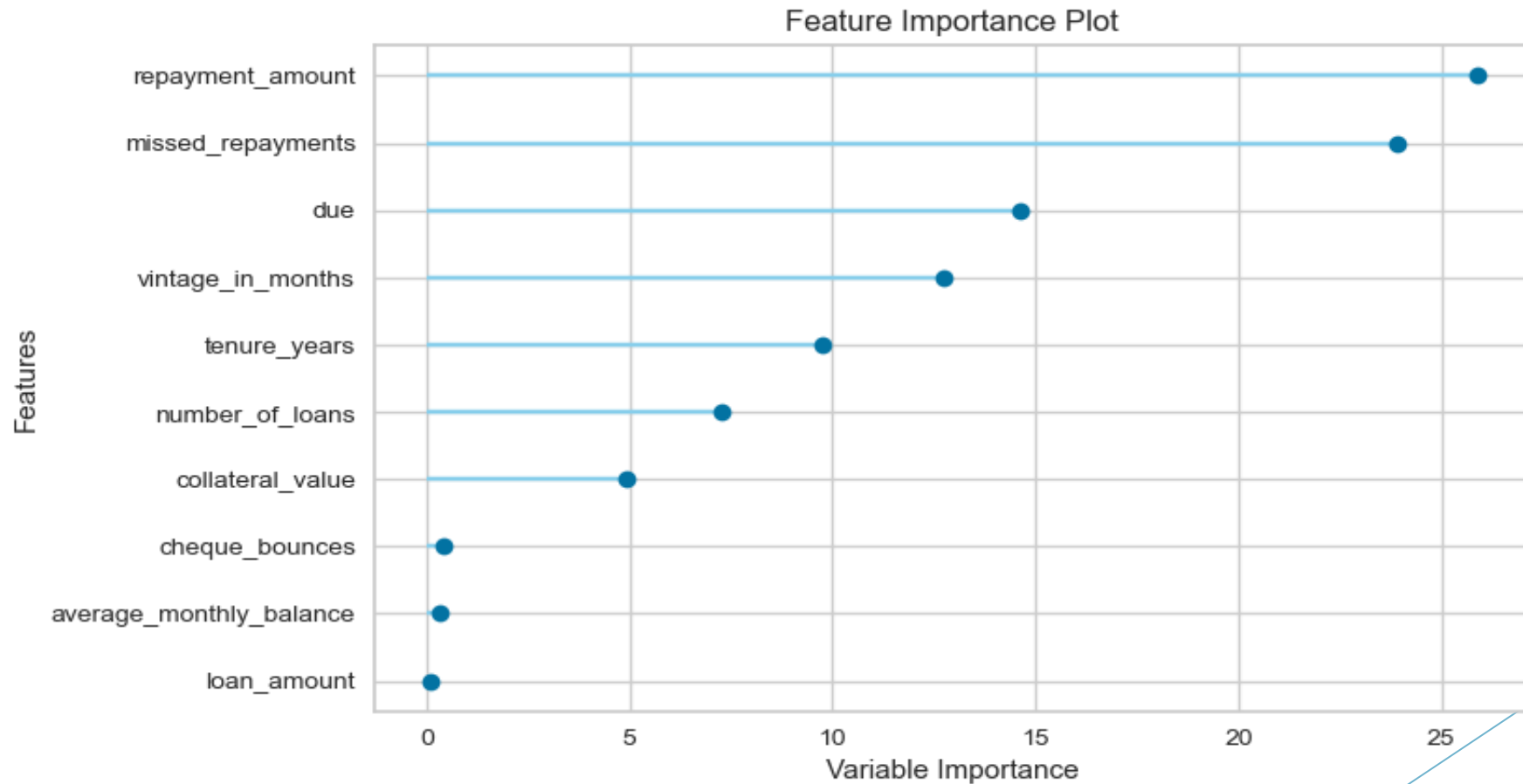
REGRESSION INTERPRETATION

❖ Validation Curve



REGRESSION INTERPRETATION

❖ Feature Importance



RECOMMENDATIONS :

- ❖ We should focus more on Car and Two-wheeler loan types.
- ❖ Missed Repayment customers with high repayment amount should be highlighted.
- ❖ Customer's due factors and tenure are another subset of influencers to predict the Loss Given Default of the customers.

