BFSI Credit Risk Assignment

BY,
NAREN HAZARE

Index

- Objective
- Background
- 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)**

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

BACKGROUND

- Credit risk analytics in the context of the banking sector and model a commonmetric 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*

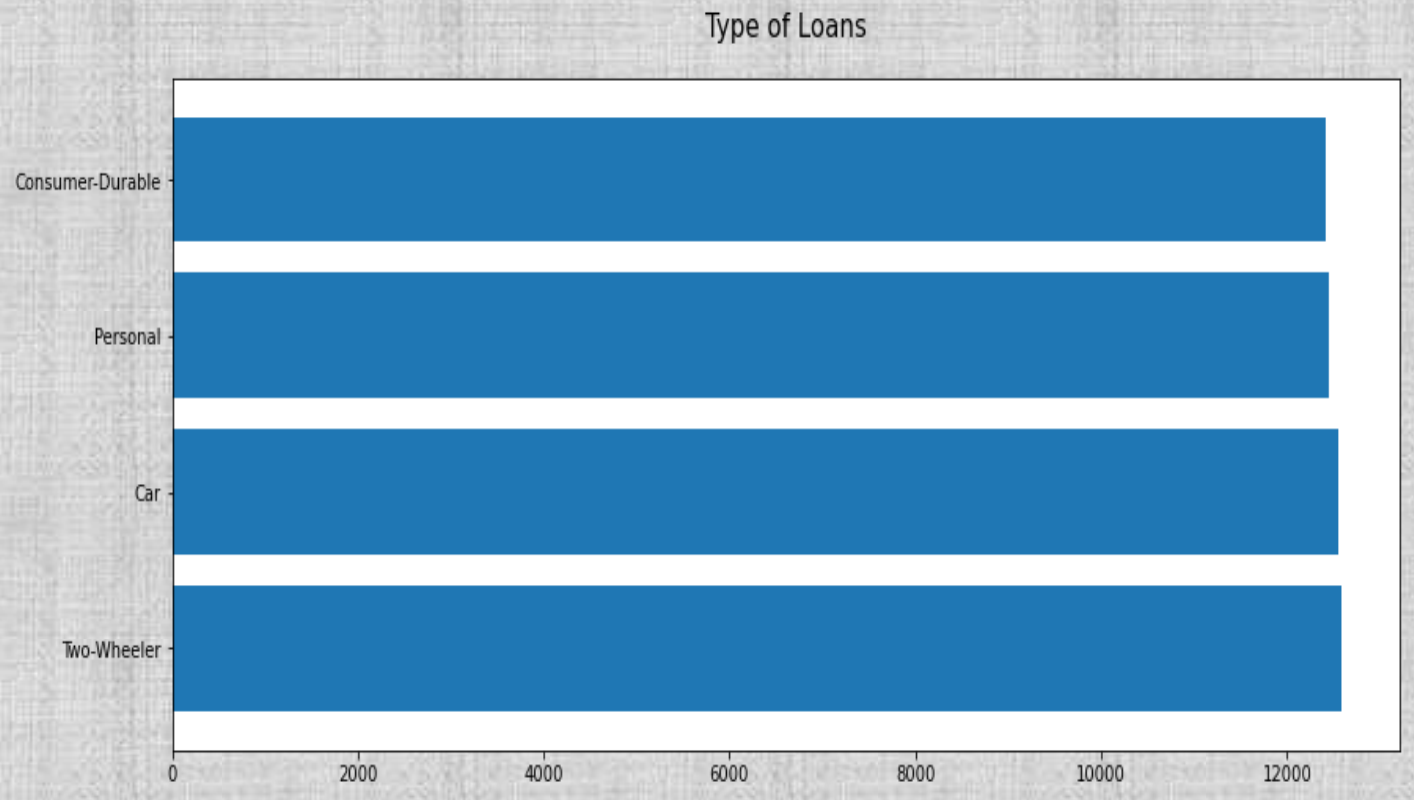
- Used 3 Data sets for model Building
  - The main_loan_base data set contains information about loan accounts and otherrelevant information for the corresponding borrowers.
  - The repayment_base data set contains information about the repayments received bythe banks in the form of EMIs or through other collection efforts.\
  - The monthly_balance_base contains the information pertaining to the monthlybalance statements in the borrower’s accounts.

PRE PROCESSING OF DATA

- The data underwent type conversion as needed for each dataset, ensuring consistency and compatibility across variables.
- Deletion and imputation methods were applied to manage null values, while duplicate entries were eliminated from the datasets.
- Integration of datasets was executed, establishing the target variable (LGD) for analysis and modeling.
- Comprehensive Exploratory Data Analysis (EDA) was conducted to gain insights and identify patterns within the datasets.
- Variable transformation techniques were employed to enhance data distribution and improve model performance.
- Utilization of dummy encoding facilitated the conversion of categorical variables into numerical equivalents, enabling inclusion in predictive models.
- Standard scaling was applied using the Standard Scaler method to normalize the features and mitigate the impact of differing scales across variables.

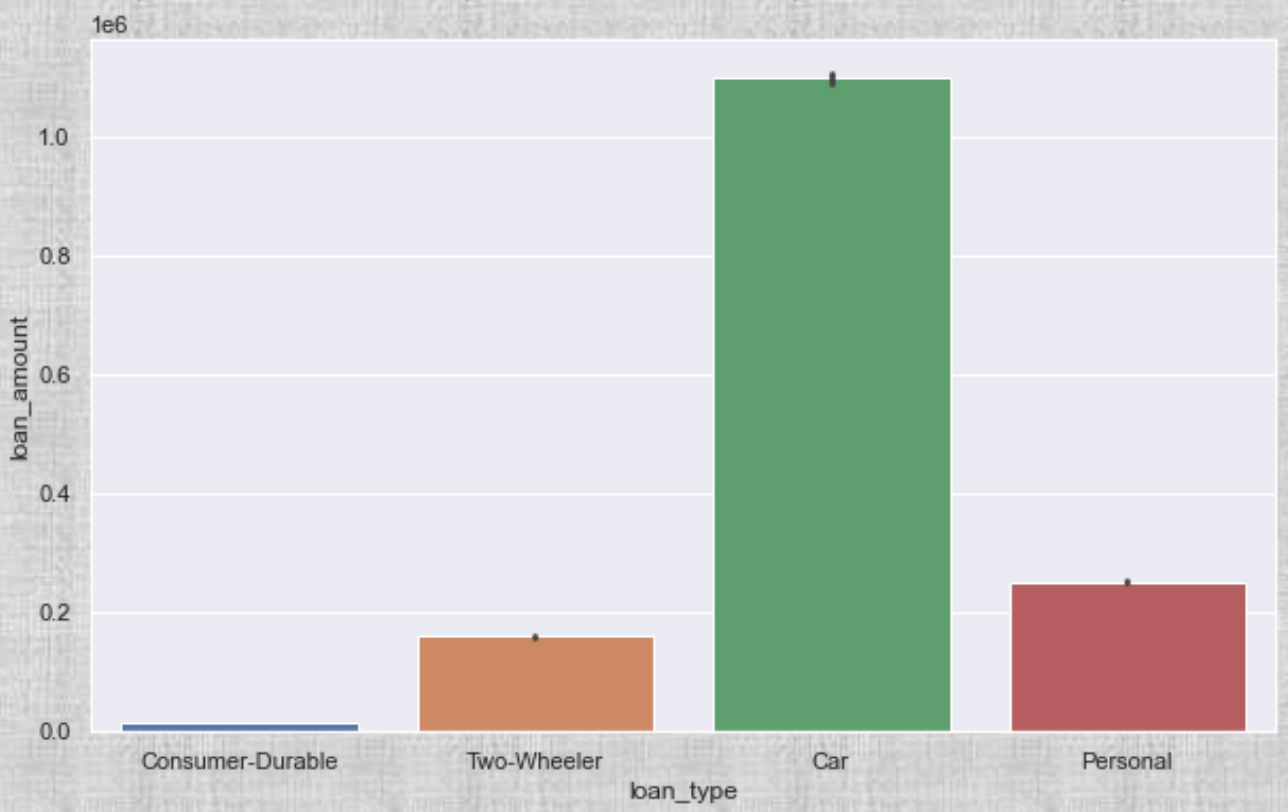
EDA

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

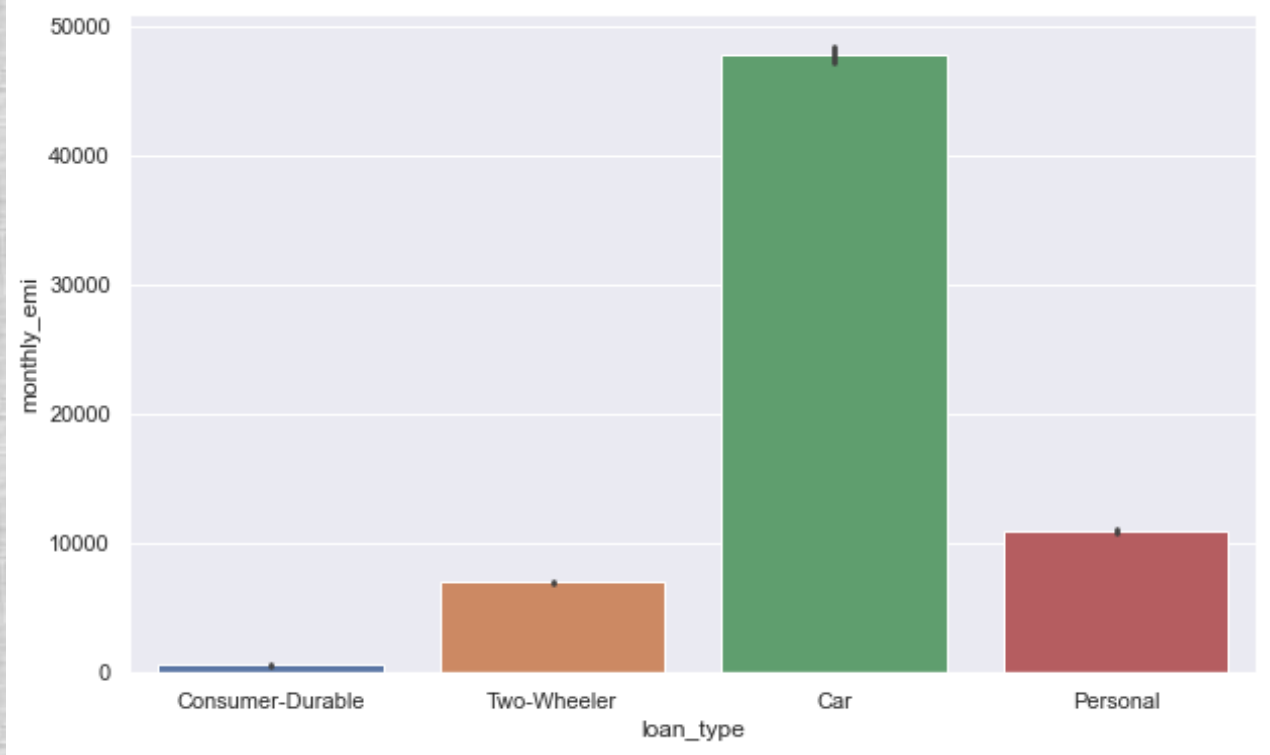


EDA

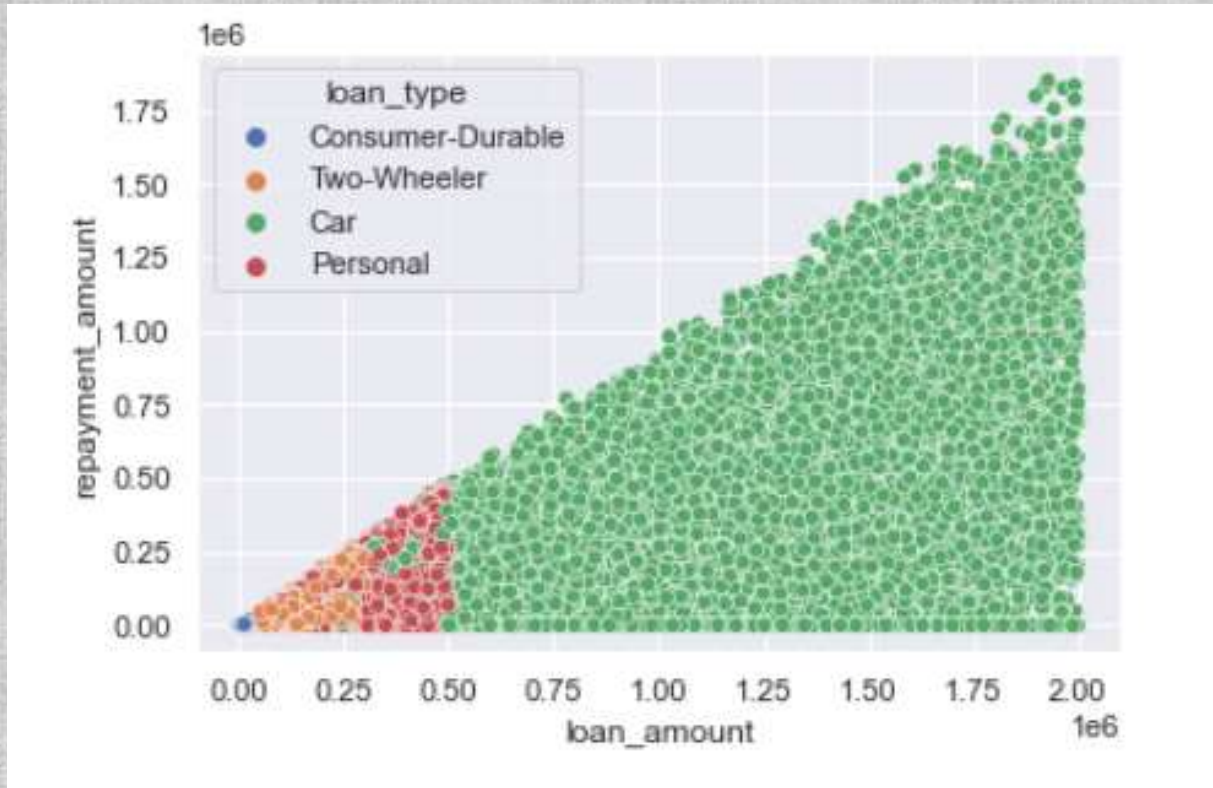
But, the loan amount of car loan is the highest.



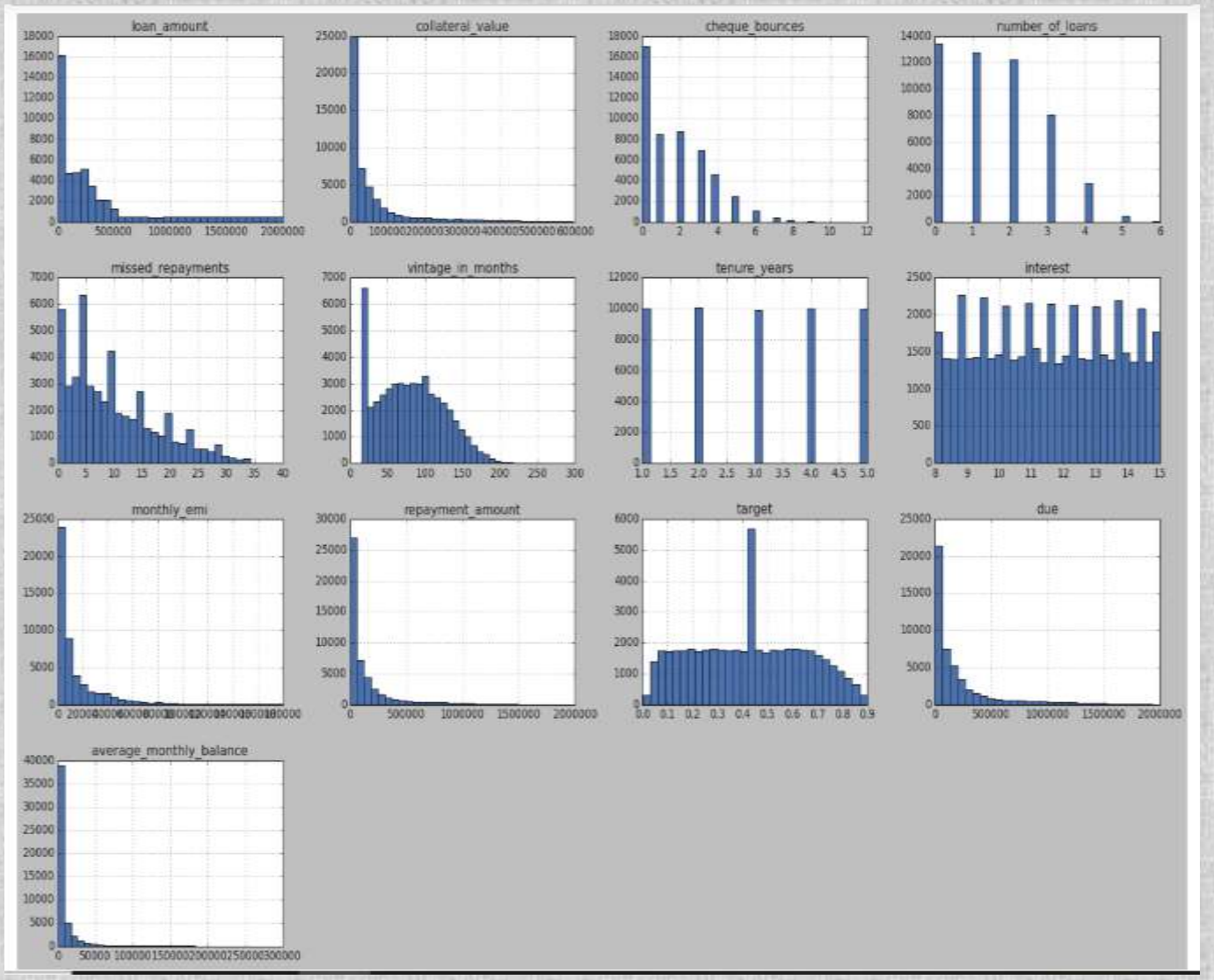
Monthly EMI also car loan is much higher compared to other loans.



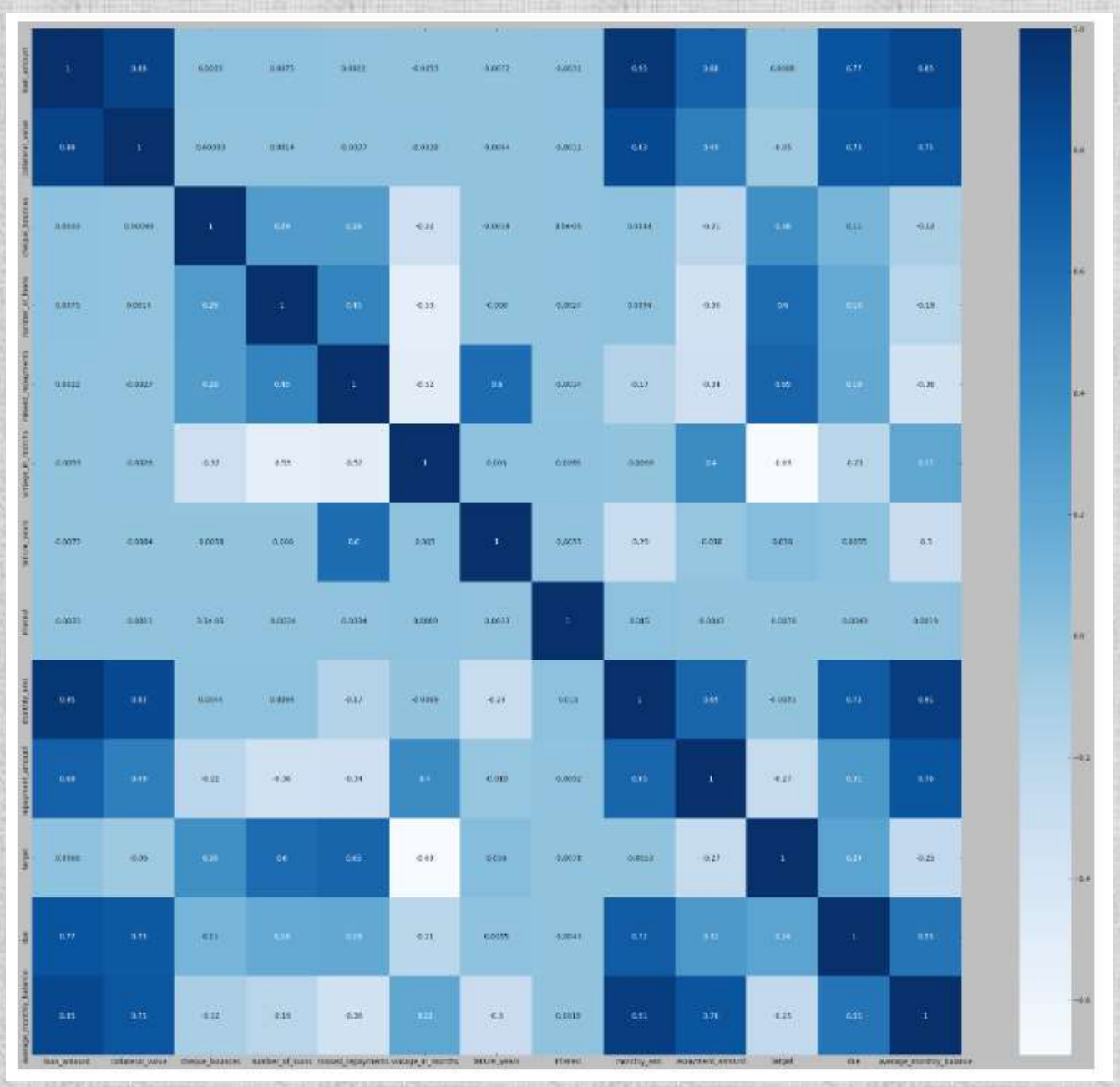
Two-Wheeler loan is greater than others, Car loan comprised the maximum loan amount



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



Created Heatmap to understand the correlation between the variables



Steps Performed

In the data preprocessing phase, several techniques were employed to enhance the quality of the dataset. Firstly, power transformation was applied to numerical variables to normalize their distributions, ensuring that they adhere to the assumptions of many statistical models. This transformation helps in stabilizing variance and making the data more symmetrical. Additionally, irrelevant columns that did not contribute significantly to the modeling process were removed to streamline the dataset and reduce noise.

Furthermore, categorical variables were encoded using the one-hot encoding technique. This method converts categorical variables into binary vectors, creating dummy variables for each category. By doing so, it preserves the categorical information in a format that is suitable for machine learning algorithms, preventing the introduction of unintended ordinality or hierarchy. These preprocessing steps collectively contribute to optimizing the dataset for further analysis and modeling, facilitating the development of accurate and robust predictive models.

MODEL BUILDING

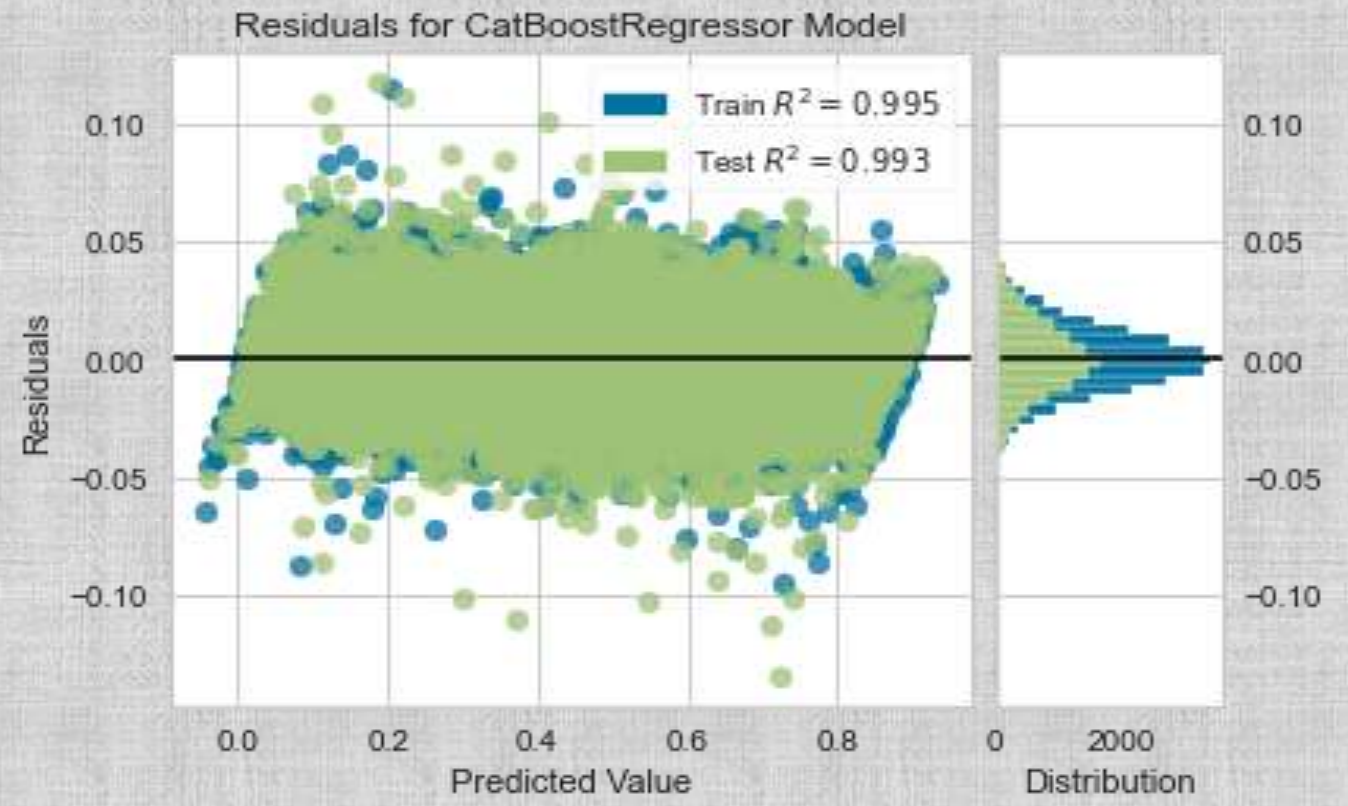
A diverse range of regression models was employed in the model building process, including Multiple Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, Adaboost Regressor, ElasticNet (a hybrid regularized model), and LightGBM. Each model offers unique strengths and capabilities, allowing for comprehensive exploration of the dataset and optimization of predictive performance.

To evaluate the performance of these models, the coefficient of determination (R-squared) was utilized as a performance metric. R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables, providing insight into the goodness of fit of the models.

Notably, XGBoost Regressor demonstrated exceptional performance, achieving an impressive R-squared value of **99.5%** on the test data across all models. This high R-squared value indicates that the XGBoost model explains a significant portion of the variance in the dependent variable, highlighting its efficacy in capturing the underlying patterns and relationships within the dataset. Overall, the

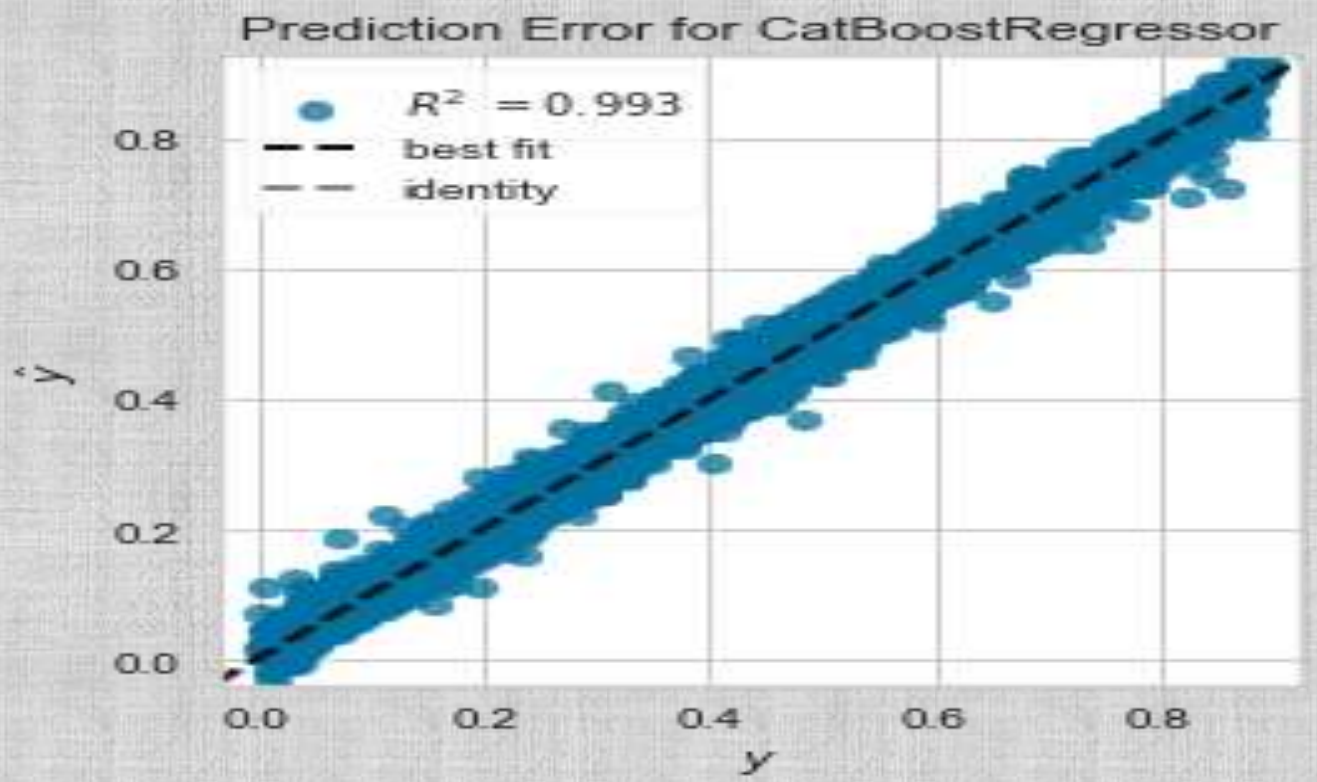
REGRESSION INTERPRETATION

- Residual plot of the finest model



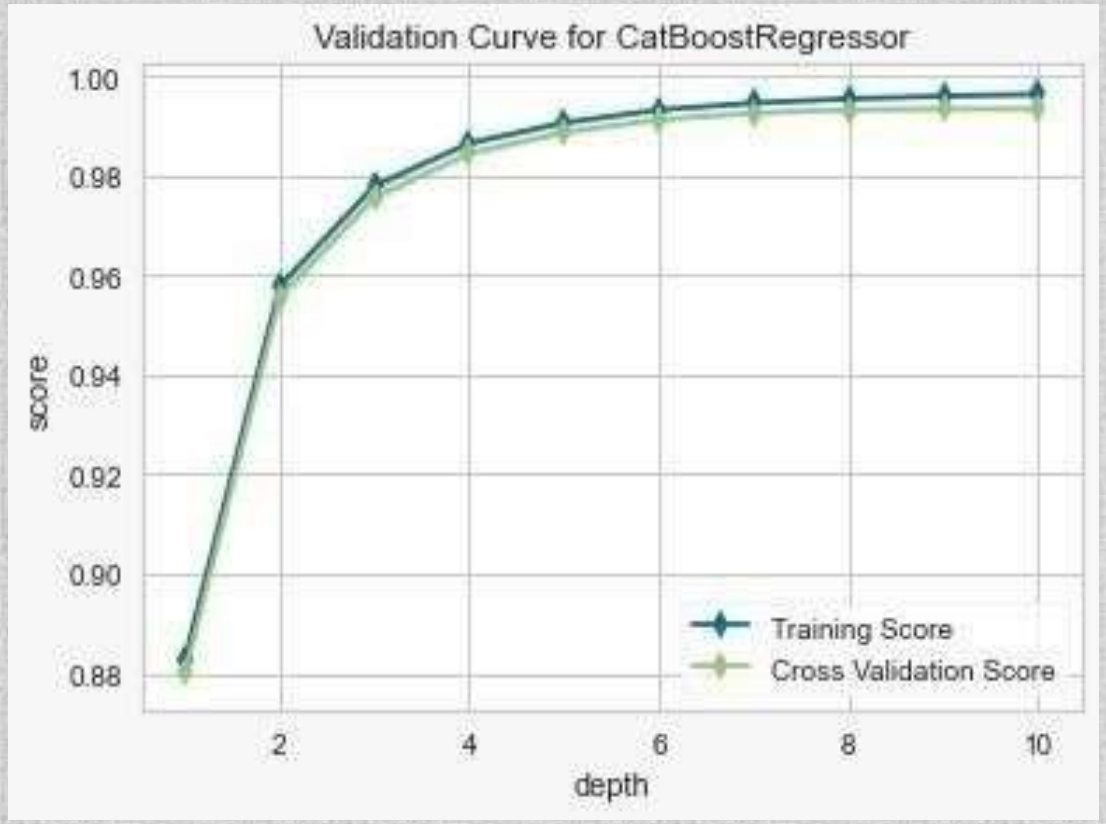
REGRESSION INTERPRETATION

- Best fit line corresponding the prediction error



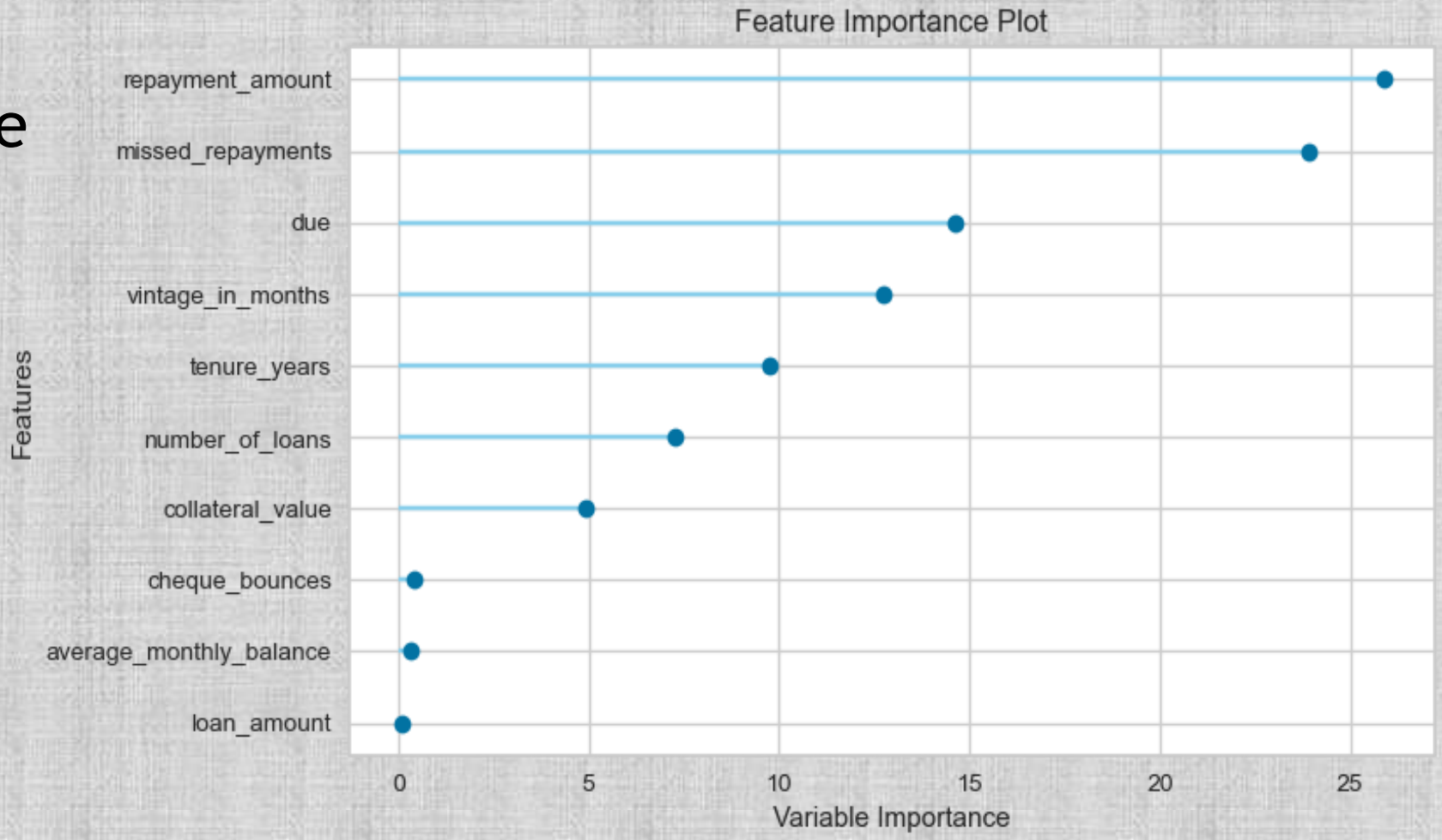
REGRESSION INTERPRETATION

- Validation Curve



REGRESSION INTERPRETATION

- Feature Importance



Final Predictions

In [146]:
```
#Making predictions
final_predictions = XGB.predict(test_data)
final_prediction_series = pd.Series(final_predictions)
```

In [147]:
```
#Combining the results into dataframe
submission_df = pd.DataFrame({'id':test['loan_acc_num'].values, 'LGD':final_prediction_series.values})
```

In [148]:
```
submission_df.sample(10)
```

Out[148]:

	id	LGD
2697	LN61230359	0.055809
9176	LN85931981	0.239187
5647	LN19460566	0.107201
7391	LN32548392	0.405380
9173	LN70910974	0.422389
3103	LN85162535	0.435993
4000	LN91919615	0.153061
1494	LN52303333	0.647433
9172	LN30034915	0.692726
5022	LN65358503	0.186089

In [ ]:
```
submission_df.to_csv("C:\\Users\\naren\\Downloads\\BFSI Credit Risk Assignment\\submission.csv",index=False)
```

RECOMMENDATIONS

- Emphasize prioritizing Car and Two-wheeler loan categories for increased attention.
- Highlight customers with missed repayments, particularly those with high repayment amounts.
- Consider customer factors such as dues and tenure as influential predictors of Loss Given Default (LGD).
- Place greater emphasis on Car and Two-wheeler loan segments for enhanced scrutiny.
- Identify and flag customers with missed repayments, especially those with significant repayment obligations.
- Evaluate customer attributes like dues and tenure to better forecast Loss Given Default for individuals.
- Prioritize Car and Two-wheeler loan portfolios for heightened monitoring and analysis.
- Identify customers with overdue payments, particularly those with substantial repayment obligations, for closer examination.
- Analyze customer characteristics such as outstanding dues and loan tenure to refine LGD predictions.


Thank You