

BFSI: CREDIT RISK ASSIGNMENT

Prepared By:
Priyanka Ohdar
Bhupendra Gautam
Amirdha Parthasarathy

CONTENTS

- Objective
- Importance of LGD Model
- Background
- Data Analysis
 - Pre Processing of Data
 - EDA
 - Model Building
 - Interpreting the Results
 - Recommendations and Conclusion

OBJECTIVE

The problem statement involves building a statistical model to estimate the Loss Given Default (LGD) for defaulted accounts in a bank. The LGD represents the proportion of the total loan amount that is expected to be lost in the event of default and ranges from 0 to 1, where 0 indicates no expected loss and 1 indicates the entire loan amount is expected to be lost.

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

Importance of LGD Model:

LGD is a critical component in estimating the Expected Credit Loss (ECL) for a bank. By accurately estimating the LGD, the bank can better understand the potential losses from its loans and make informed decisions about credit risk management and provisioning. An effective LGD model can help the bank identify high-risk loans and take appropriate actions to mitigate risks.

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

The problem statement provides relevant data sets that include information about defaulted accounts and the amount of money retrieved from them using collaterals and other collection methods. It is important to thoroughly understand the data sets, including variables, data types, and data distribution. The collection data needs to be aggregated and merged to gather relevant information for building the LGD model.

- 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.

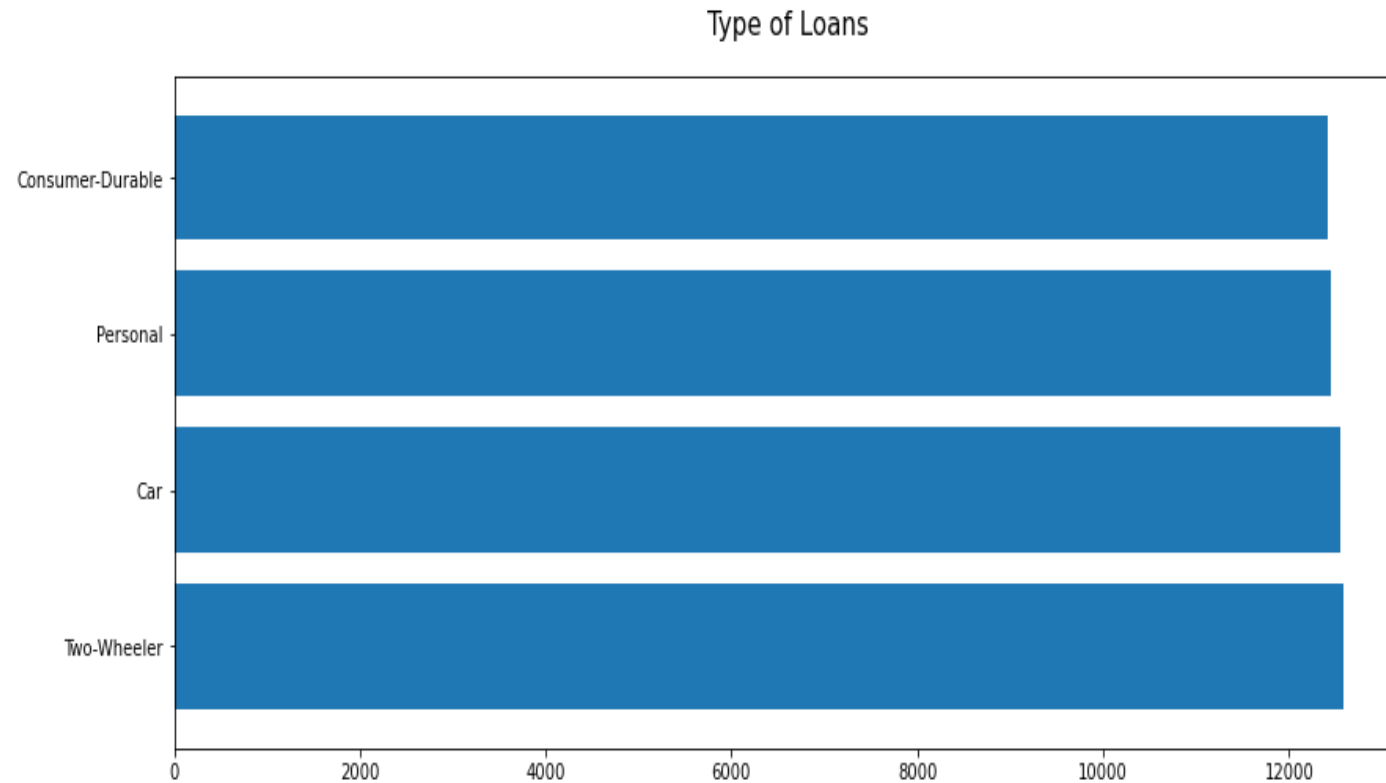
PRE PROCESSING OF DATA

Data cleaning and pre-processing are crucial steps in building an accurate LGD model. This includes handling missing values, outliers, and other issues that may affect the model's performance. Additionally, feature engineering or feature extraction may be necessary to create new variables that could be useful in the model.

- 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

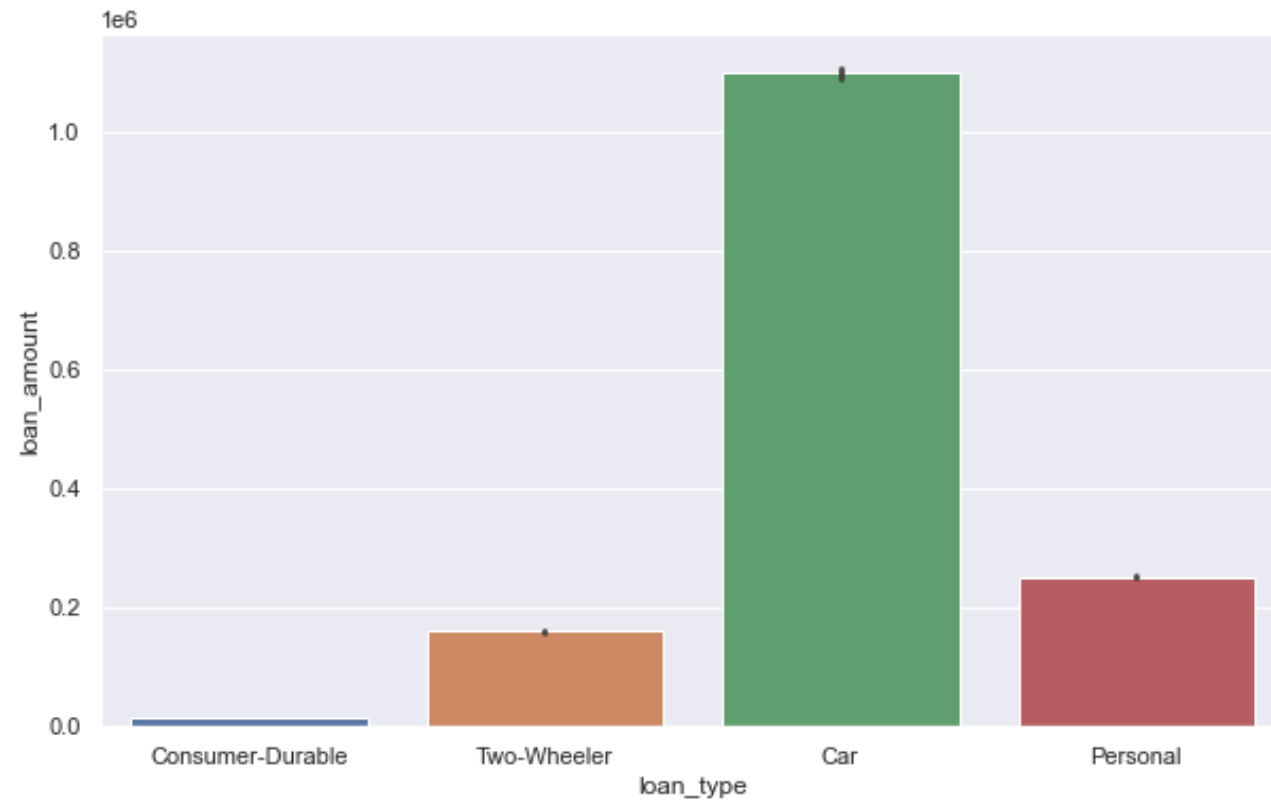
EDA

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

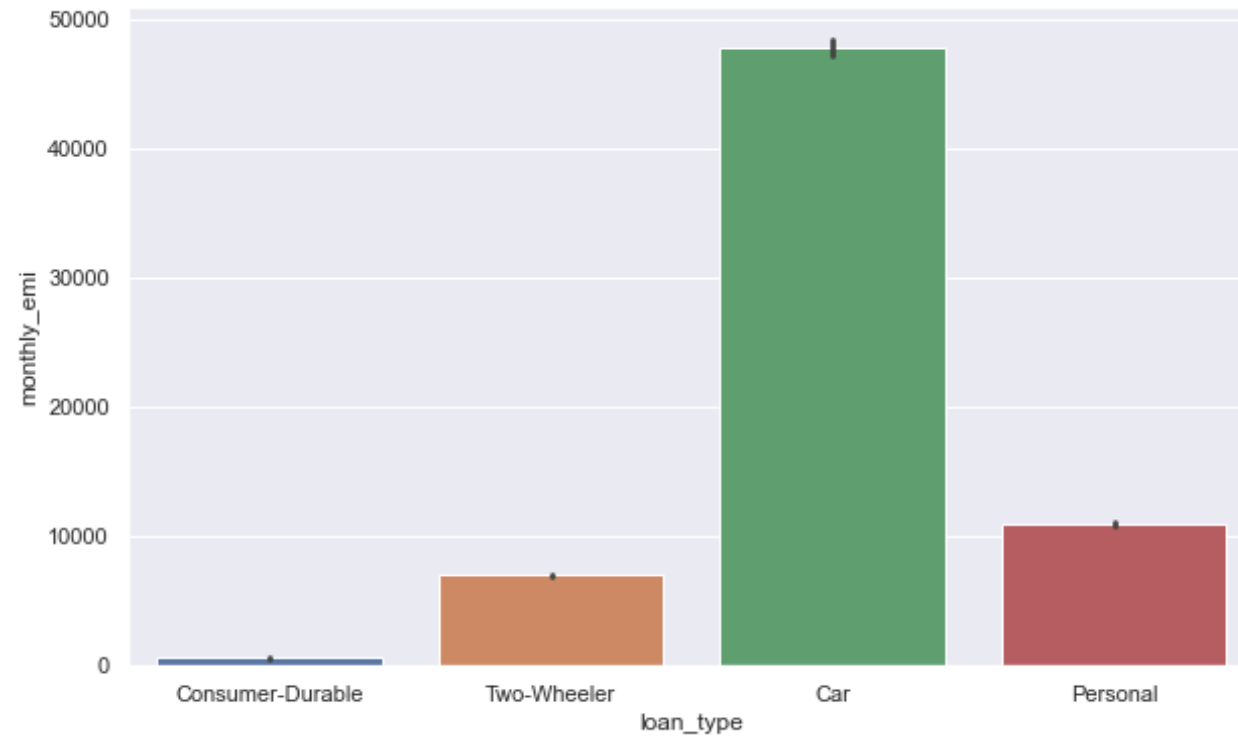


EDA

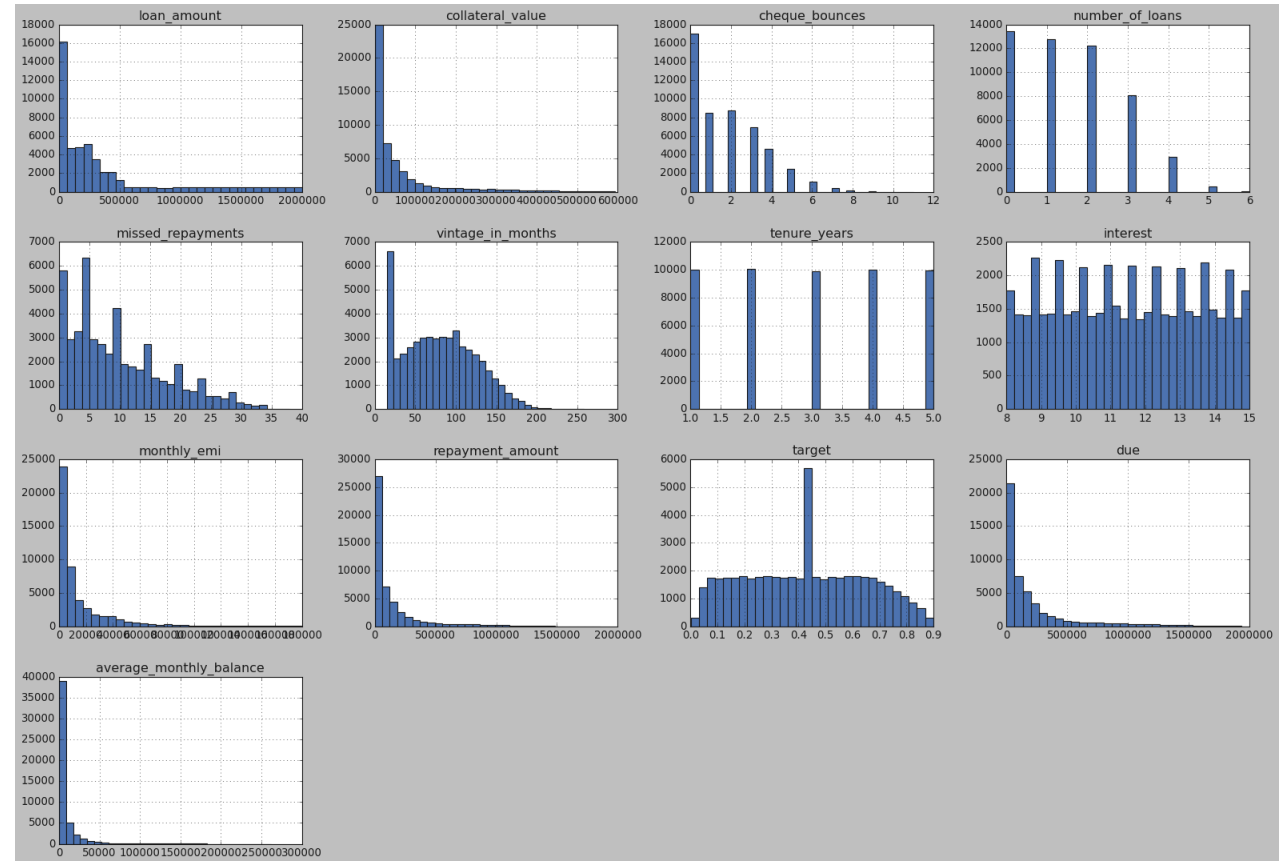
But, the loan amount of car loan is the highest.



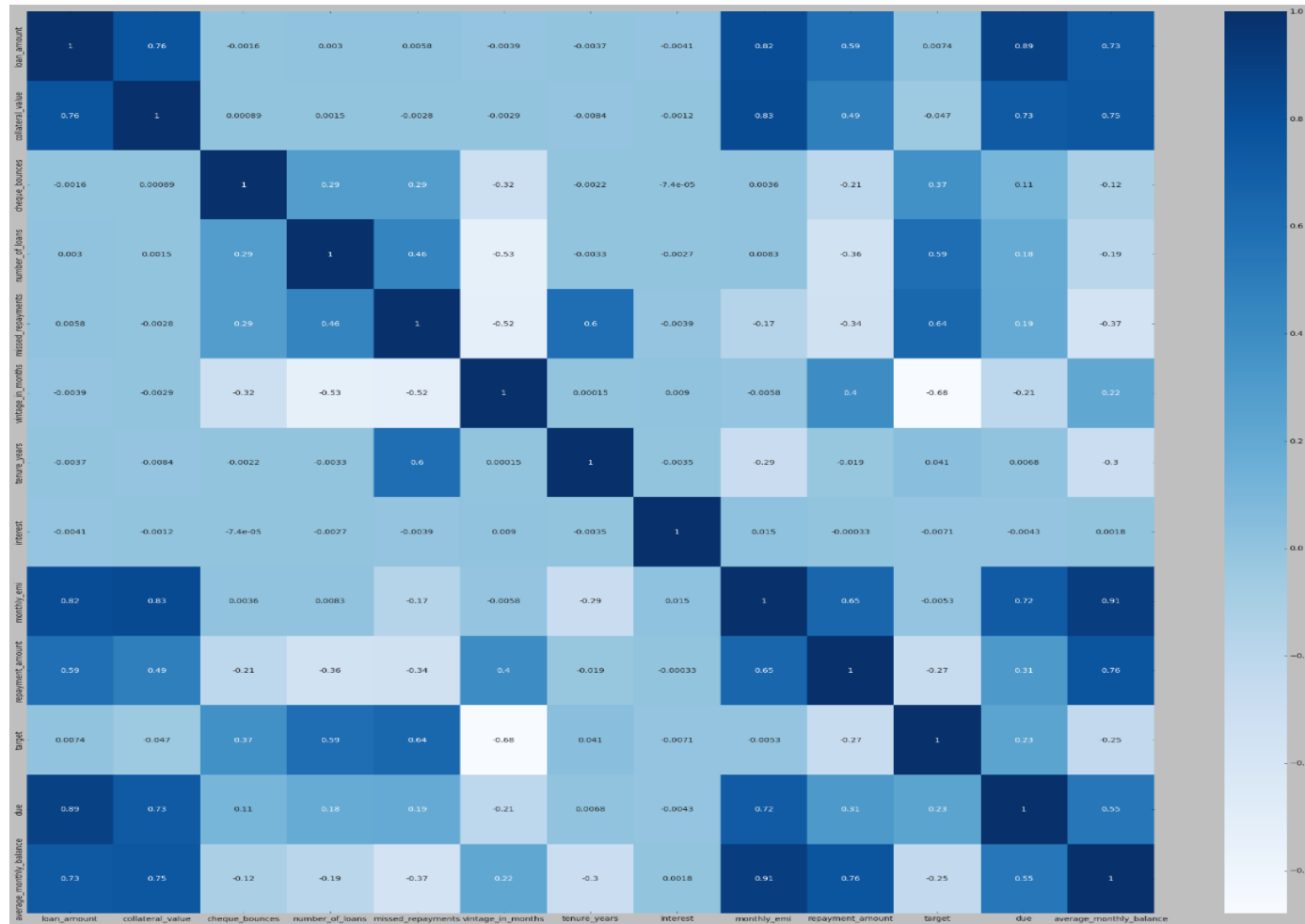
Monthly EMI also car loan is much higher compared to other loans.



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



Created Heatmap to understand the correlation between the variables



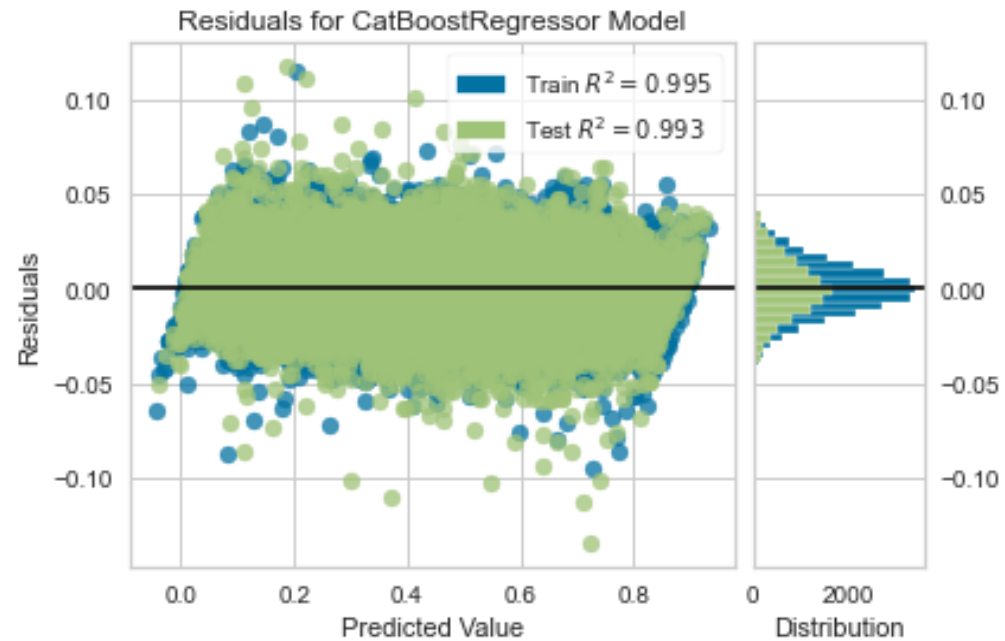
- Used Power transformation to make numerical variables Normally distributed
- Dropped unnecessary columns for model building
- Used One-hot encoding technique and created dummy variable for necessary categorical variables.

MODEL BUILDING

- Used various models like Multiple Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, Adaboost Regressor, ElasticNet :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.

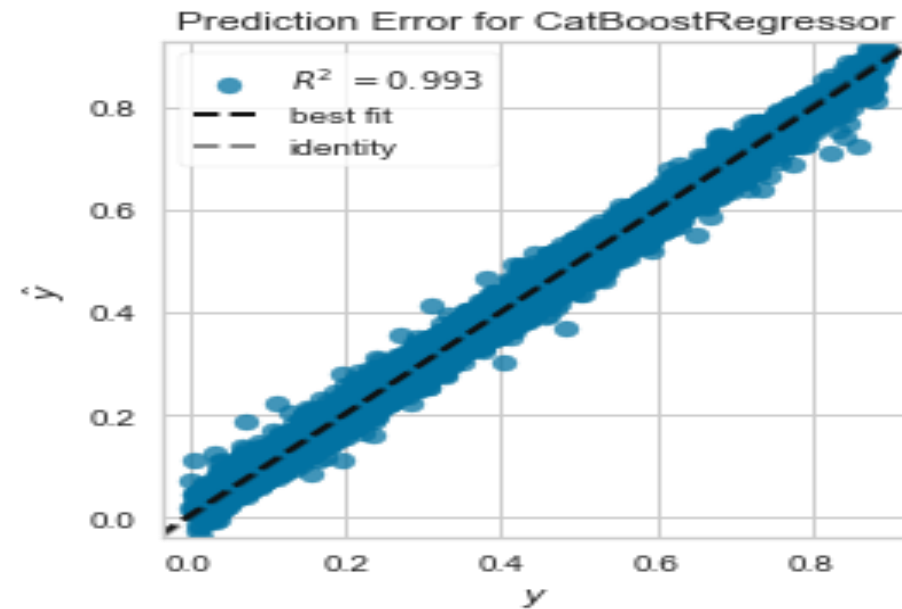
REGRESSION INTERPRETATION

- Residual plot of the finest model



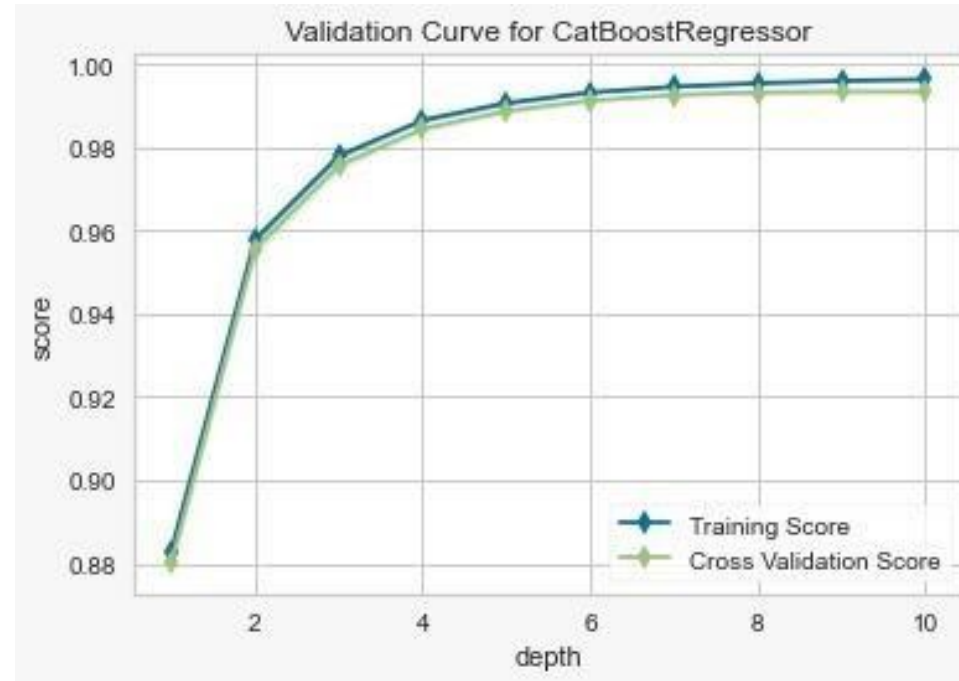
REGRESSION INTERPRETATION

- Best fit line corresponding the prediction error



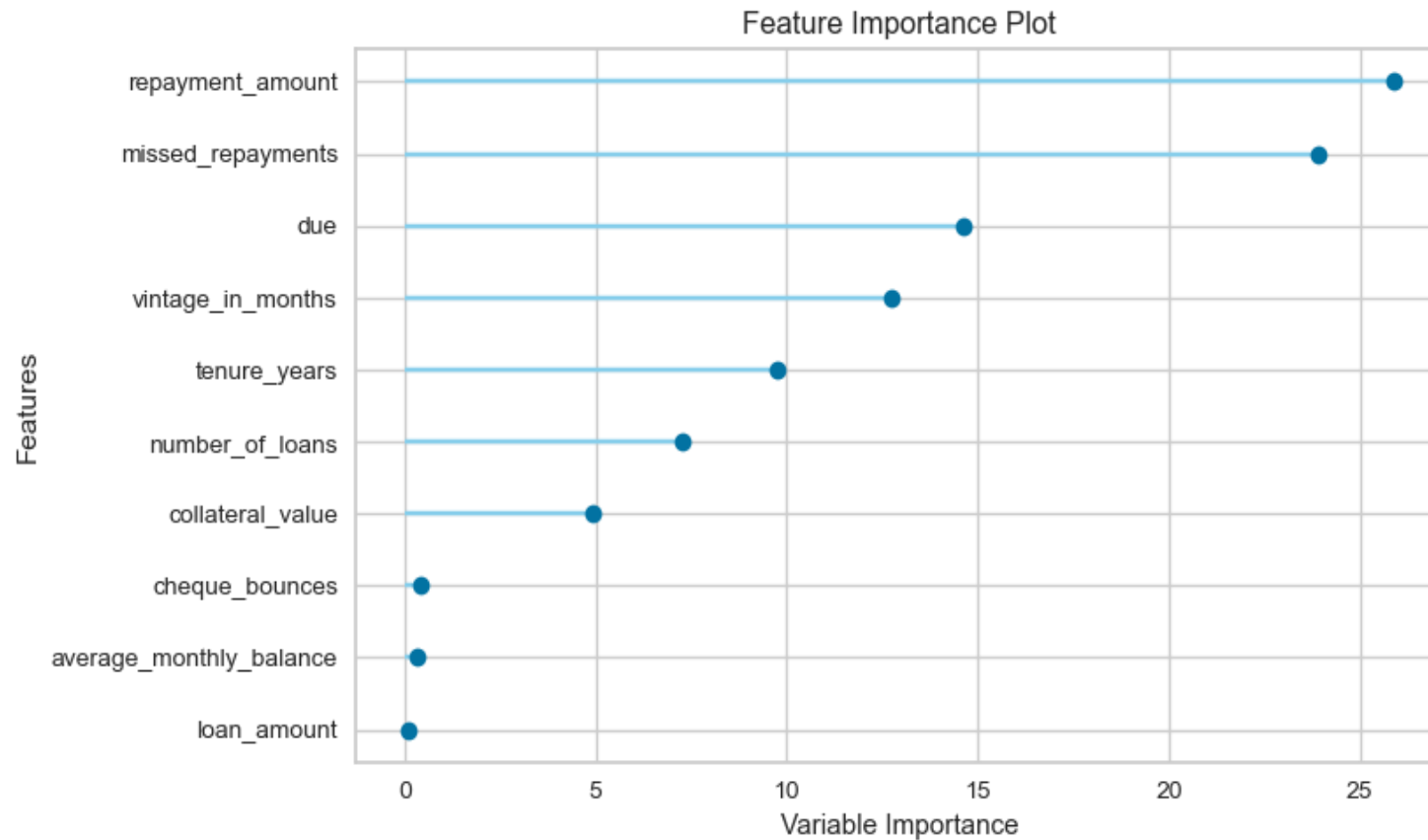
REGRESSION INTERPRETATION

- Validation Curve



REGRESSION INTERPRETATION

- Feature Importance



RECOMMENDATIONS

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

The LGD model has significant implications for the bank's business operations, particularly in enhancing risk management and compliance with regulatory standards such as Basel norms. Accurate LGD estimates can help the bank allocate appropriate provisions for potential loan losses, manage credit risk effectively, and comply with regulatory requirements.

THANKYOU

