BFSI: CREDIT RISK ASSIGNMENT

By: Shreevatsa Hegde, Shraddha Nerlekar, Shruthi J

Contents

- 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)**

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

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

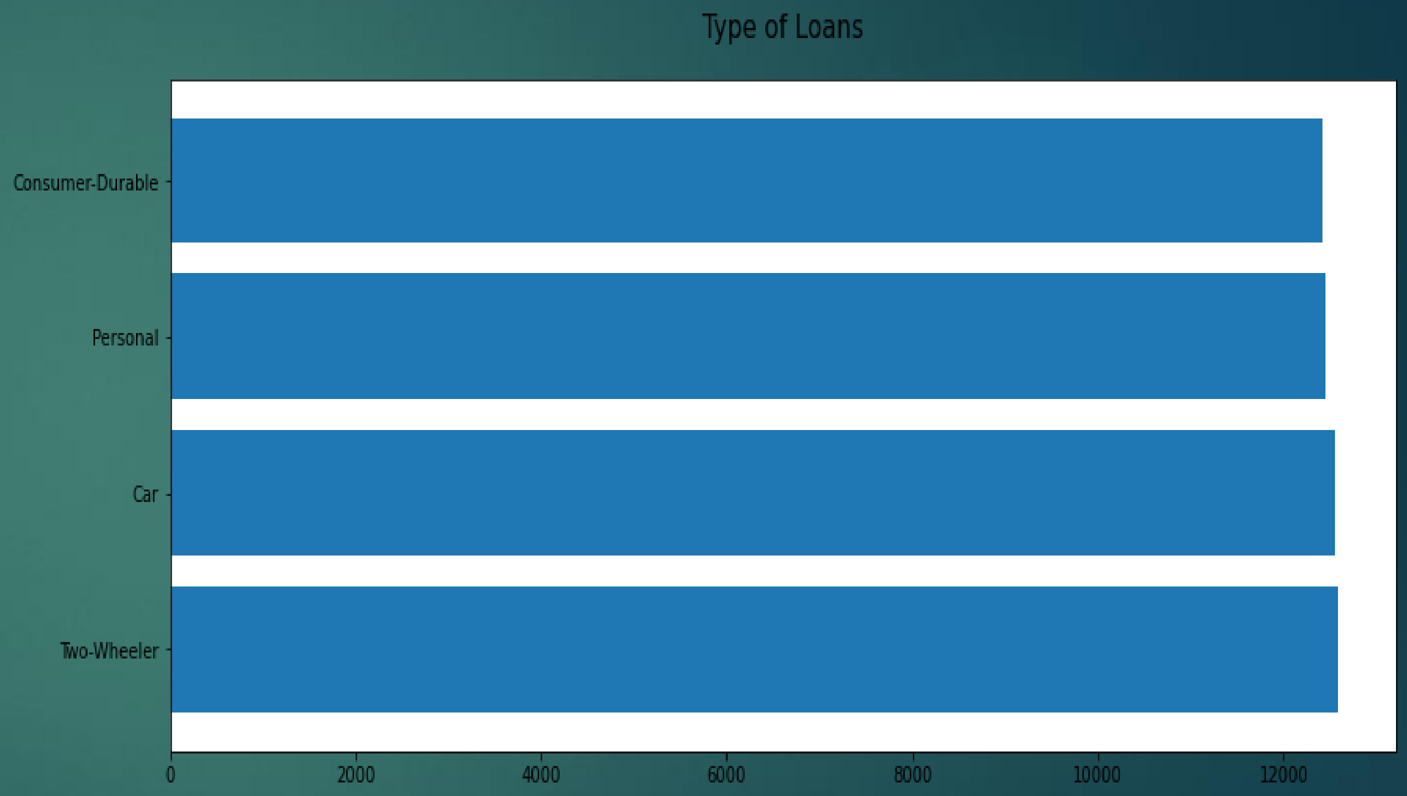
- Used 3 Data sets for model Building
  - The main_loan_base data set contains information about loan accounts and other relevant information for the corresponding borrowers.
  - The repayment_base data set contains information about the repayments received by the banks in the form of EMIs or through 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

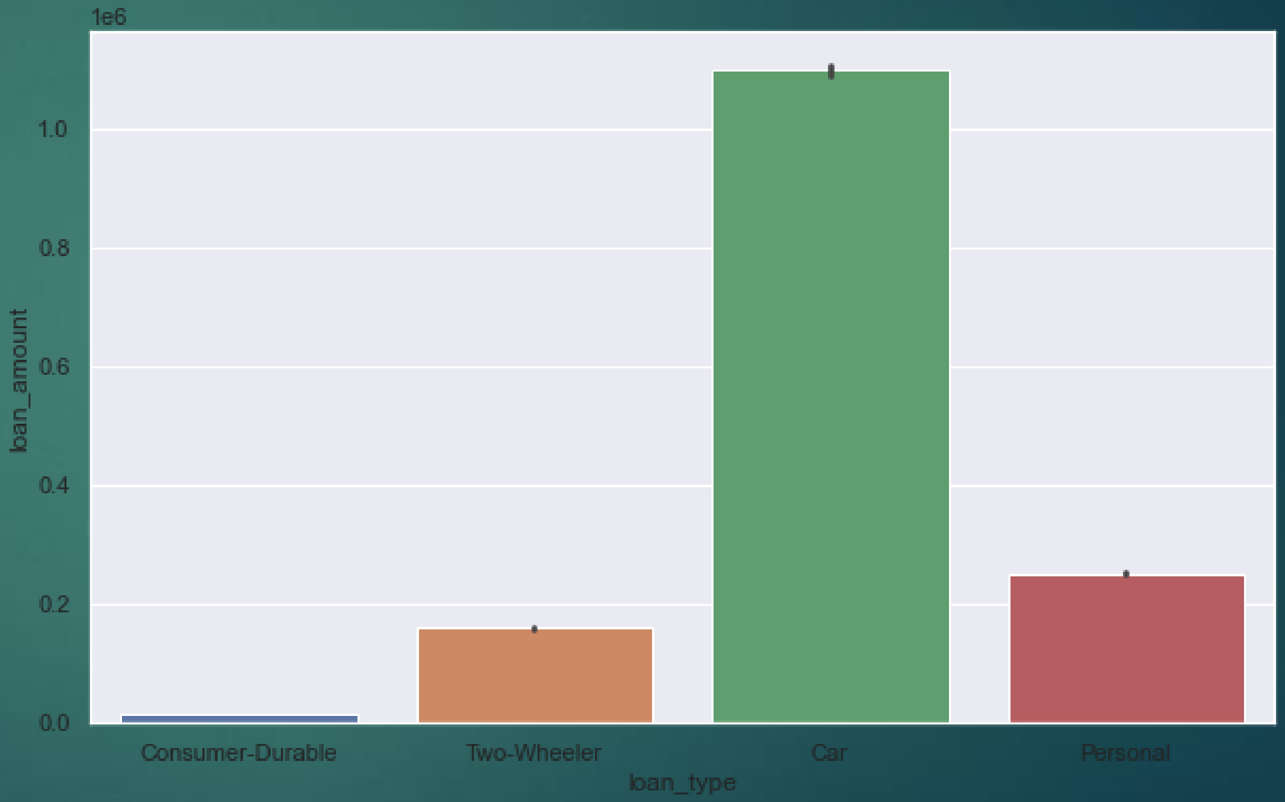
EDA

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



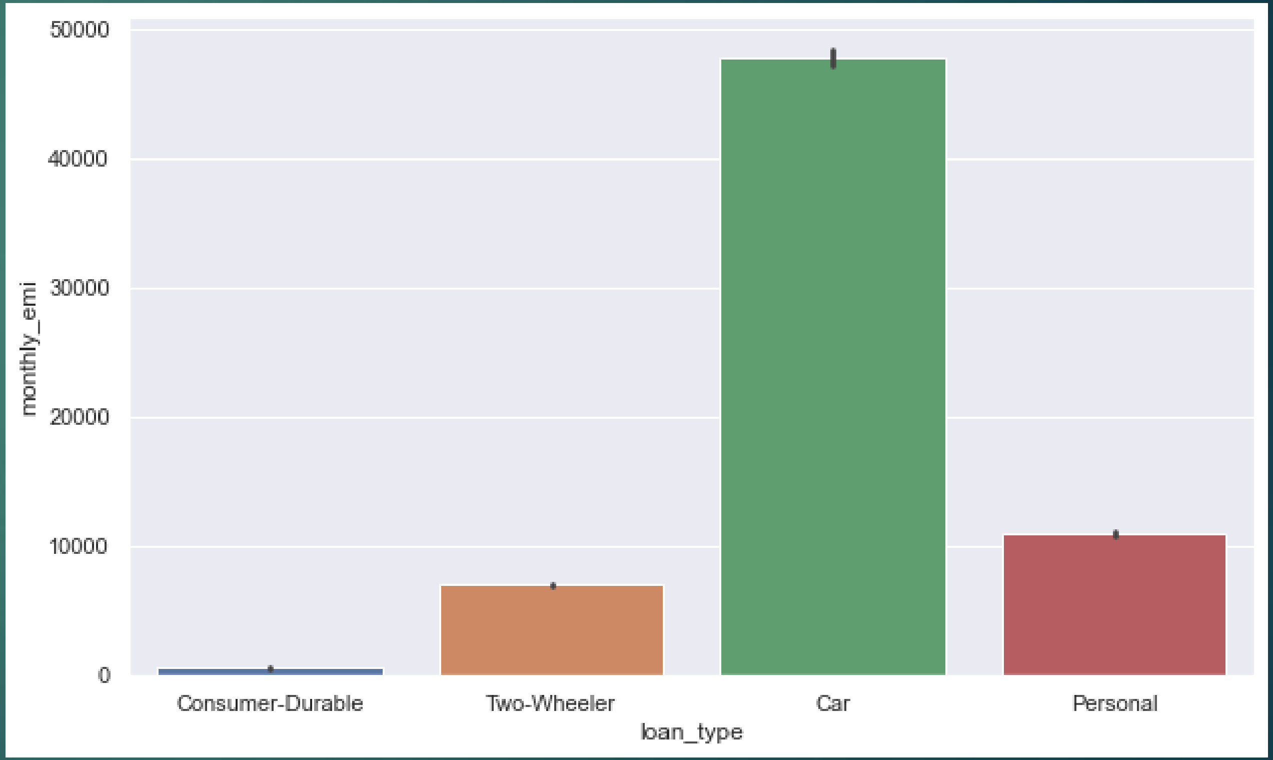
EDA

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

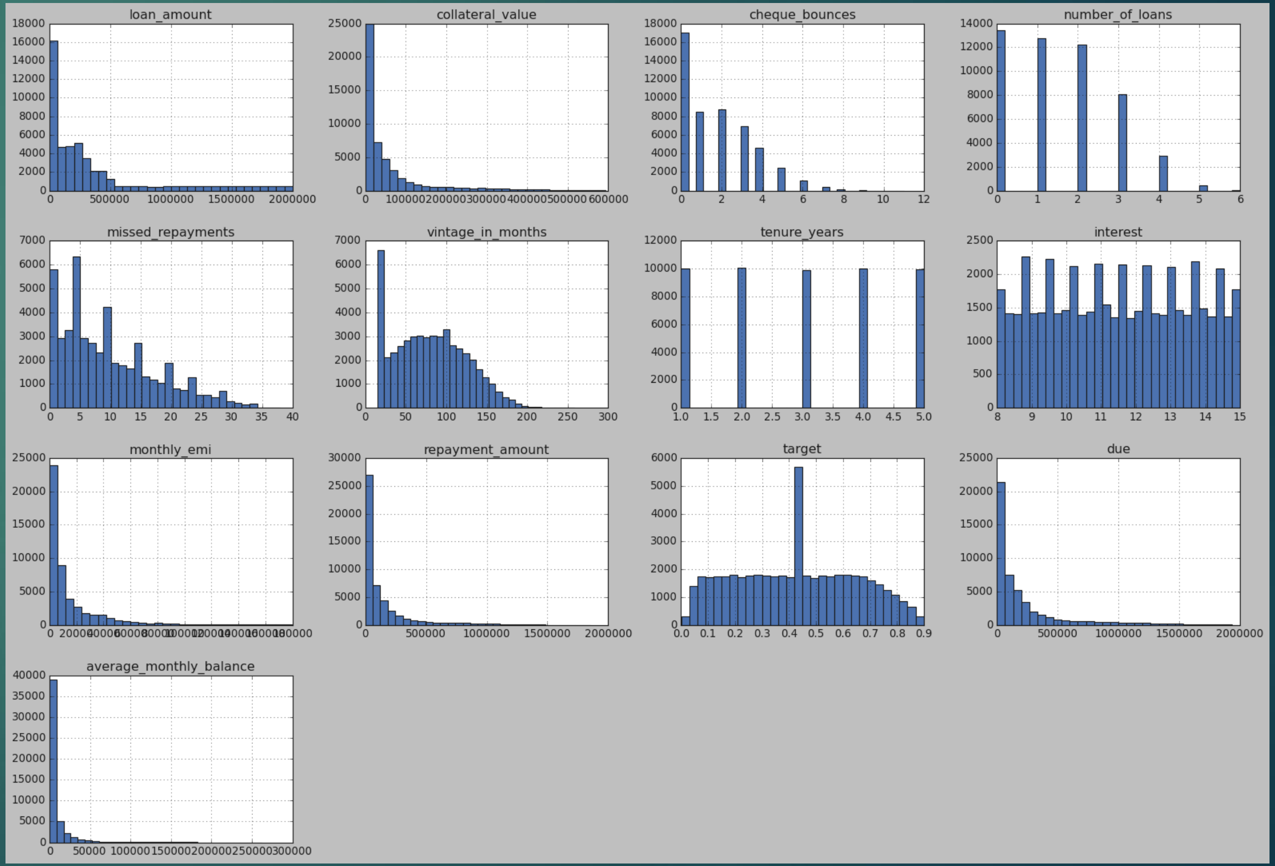


EDA

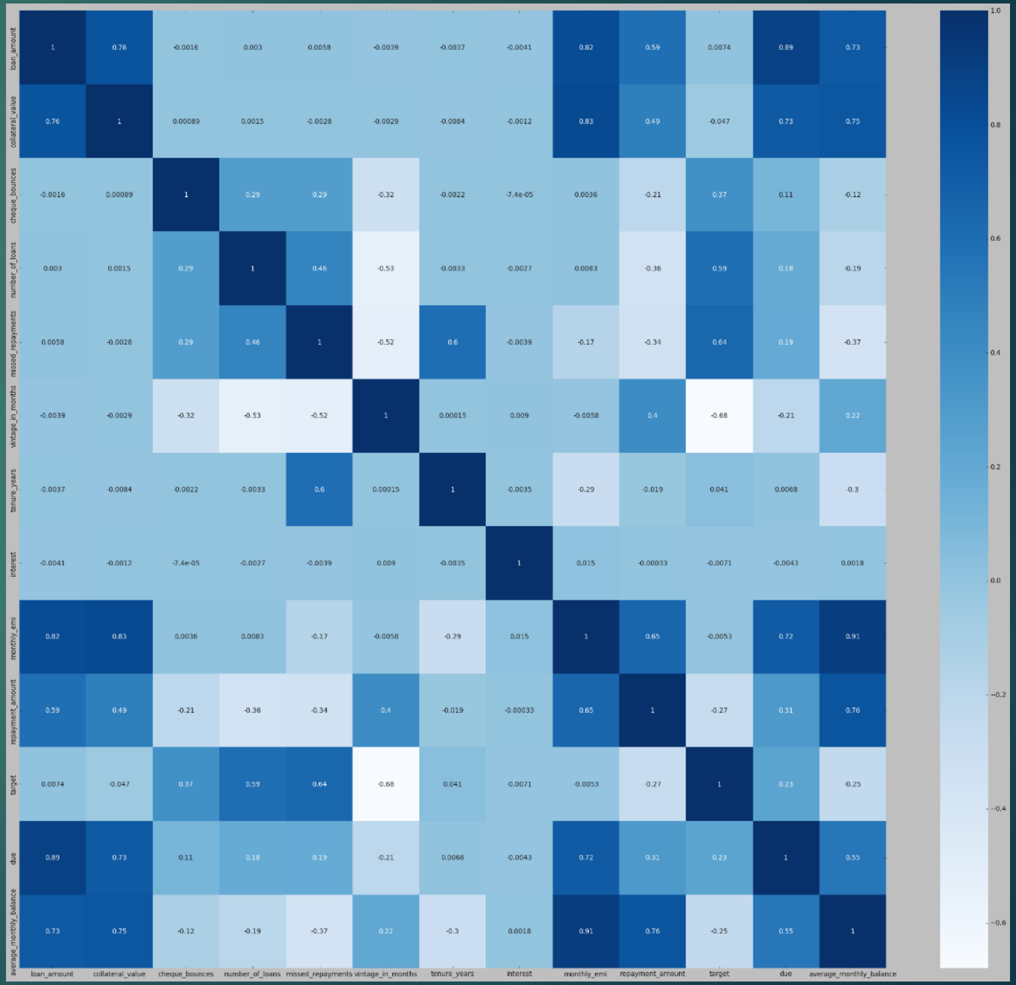
Monthly EMI also car loan is much higher compared to other loans.



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



Created Heatmap to understand the correlation between the variables



	loan_amount	collateral_value	cheque_bounces	number_of_loans	missed_repayments	vintage_in_months	tenure_years	interest	monthly_emi	repayment_amount	target	due	average_monthly_balance
loan_amount	1	0.76	-0.0016	0.003	0.0058	-0.0039	-0.0037	-0.0041	0.82	0.59	0.0074	0.89	0.73
collateral_value	0.76	1	0.00089	0.0015	-0.0028	-0.0029	-0.0084	-0.0012	0.83	0.49	-0.047	0.73	0.75
cheque_bounces	-0.0016	0.00089	1	0.29	0.29	-0.32	-0.0022	-7.4e-05	0.0036	-0.21	0.37	0.11	-0.12
number_of_loans	0.003	0.0015	0.29	1	0.46	-0.53	-0.0033	-0.0027	0.0083	-0.36	0.59	0.18	-0.19
missed_repayments	0.0058	-0.0028	0.29	0.46	1	-0.52	0.6	-0.0039	-0.17	-0.34	0.64	0.19	-0.37
vintage_in_months	-0.0039	-0.0029	-0.32	-0.53	-0.52	1	0.00015	0.009	-0.0058	0.4	-0.68	-0.21	0.22
tenure_years	-0.0037	-0.0084	-0.0022	-0.0033	0.6	0.00015	1	-0.0035	-0.29	-0.019	0.041	0.0068	-0.3
interest	-0.0041	-0.0012	-7.4e-05	-0.0027	-0.0039	0.009	-0.0035	1	0.015	-0.00033	-0.0071	-0.0043	0.0018
monthly_emi	0.82	0.83	0.0036	0.0083	-0.17	-0.0058	-0.29	0.015	1	0.65	-0.0053	0.72	0.91
repayment_amount	0.59	0.49	-0.21	-0.36	-0.34	0.4	-0.019	-0.00033	0.65	1	-0.27	0.31	0.76
target	0.0074	-0.047	0.37	0.59	0.64	-0.68	0.041	-0.0071	-0.0053	-0.27	1	0.23	-0.25
due	0.89	0.73	0.11	0.18	0.19	-0.21	0.0068	-0.0043	0.72	0.31	0.23	1	0.55
average_monthly_balance	0.73	0.75	-0.12	-0.19	-0.37	0.22	-0.3	0.0018	0.91	0.76	-0.25	0.55	1

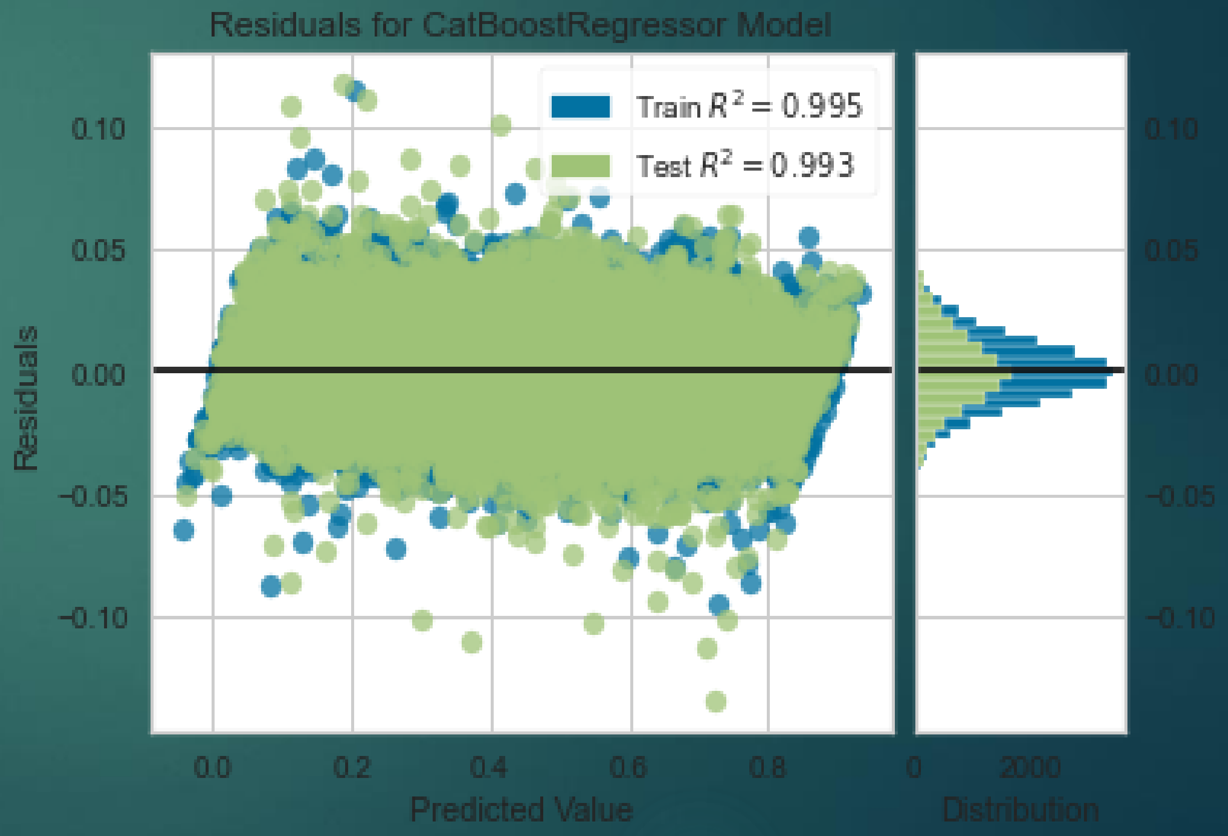
- used Power transformation to make numerical variables Normally distributed
- Dropped unnecessary columns for model building
- Used One-hot encoding technique and created dummy variable for necessary categorical variables.

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.

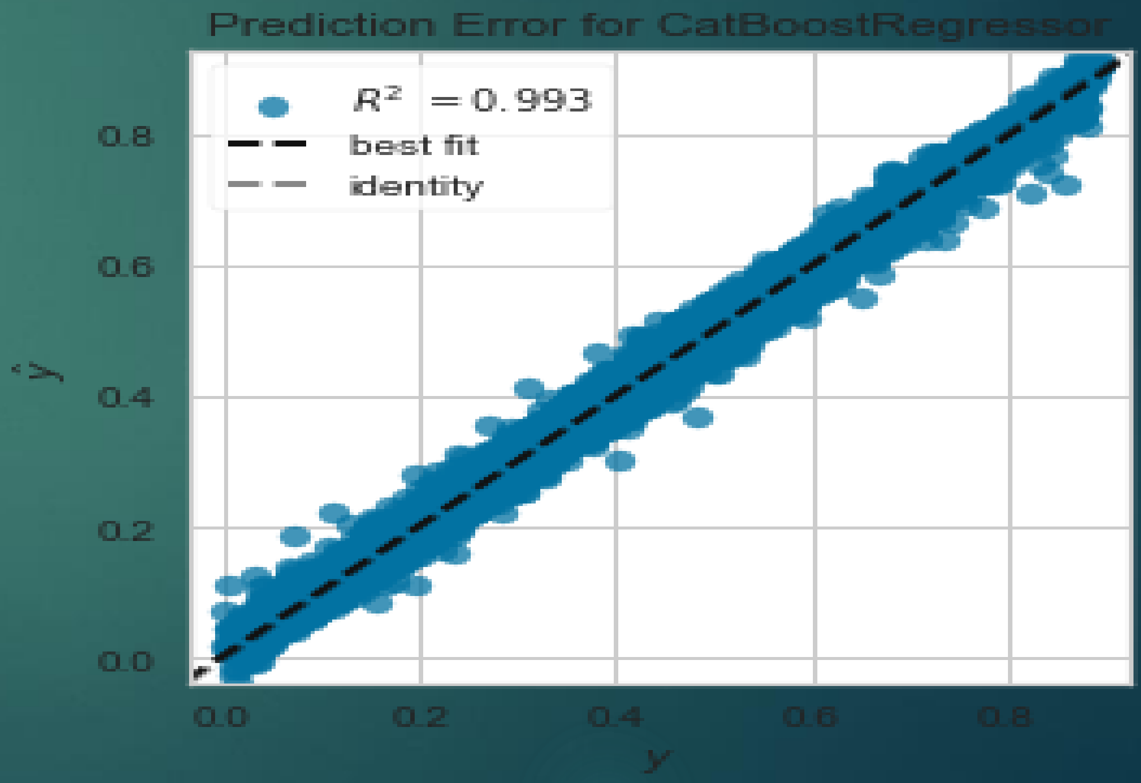
REGRESSION INTERPRETATION

- Residual plot of the finest model



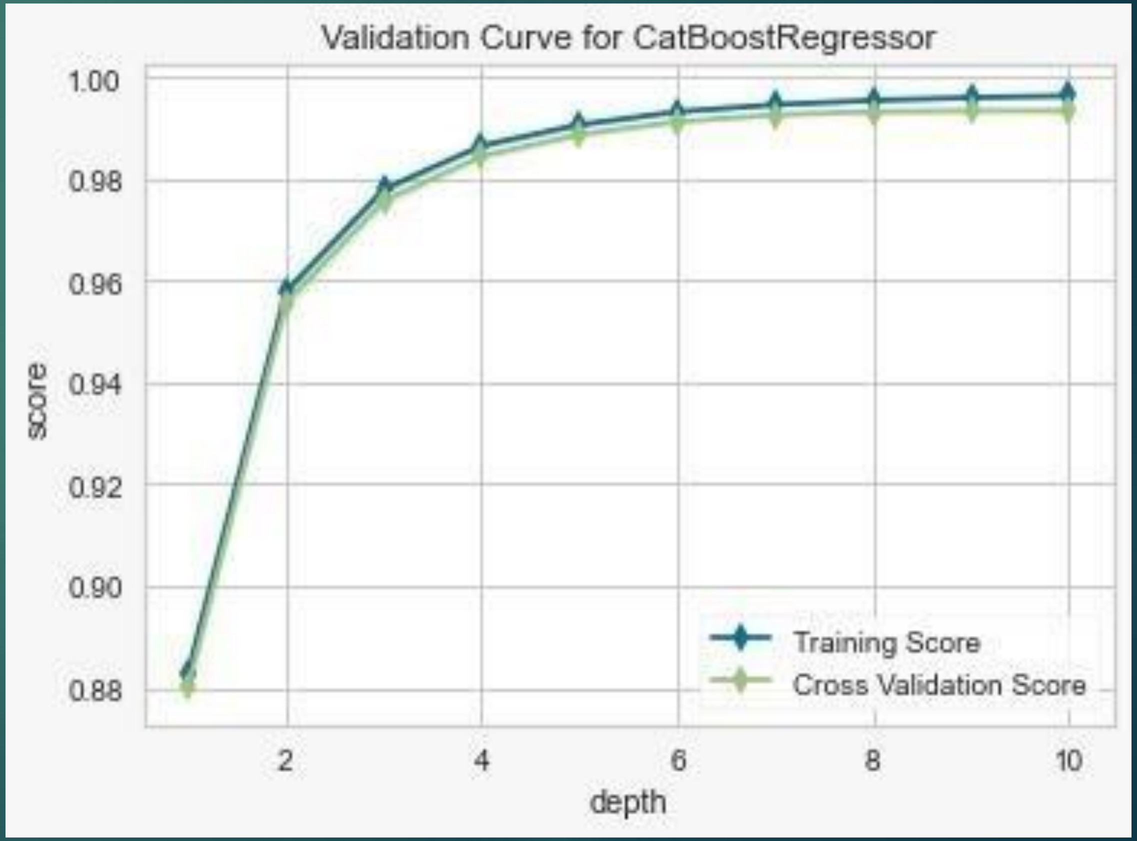
REGRESSION INTERPRETATION

- Best fit line corresponding the prediction error



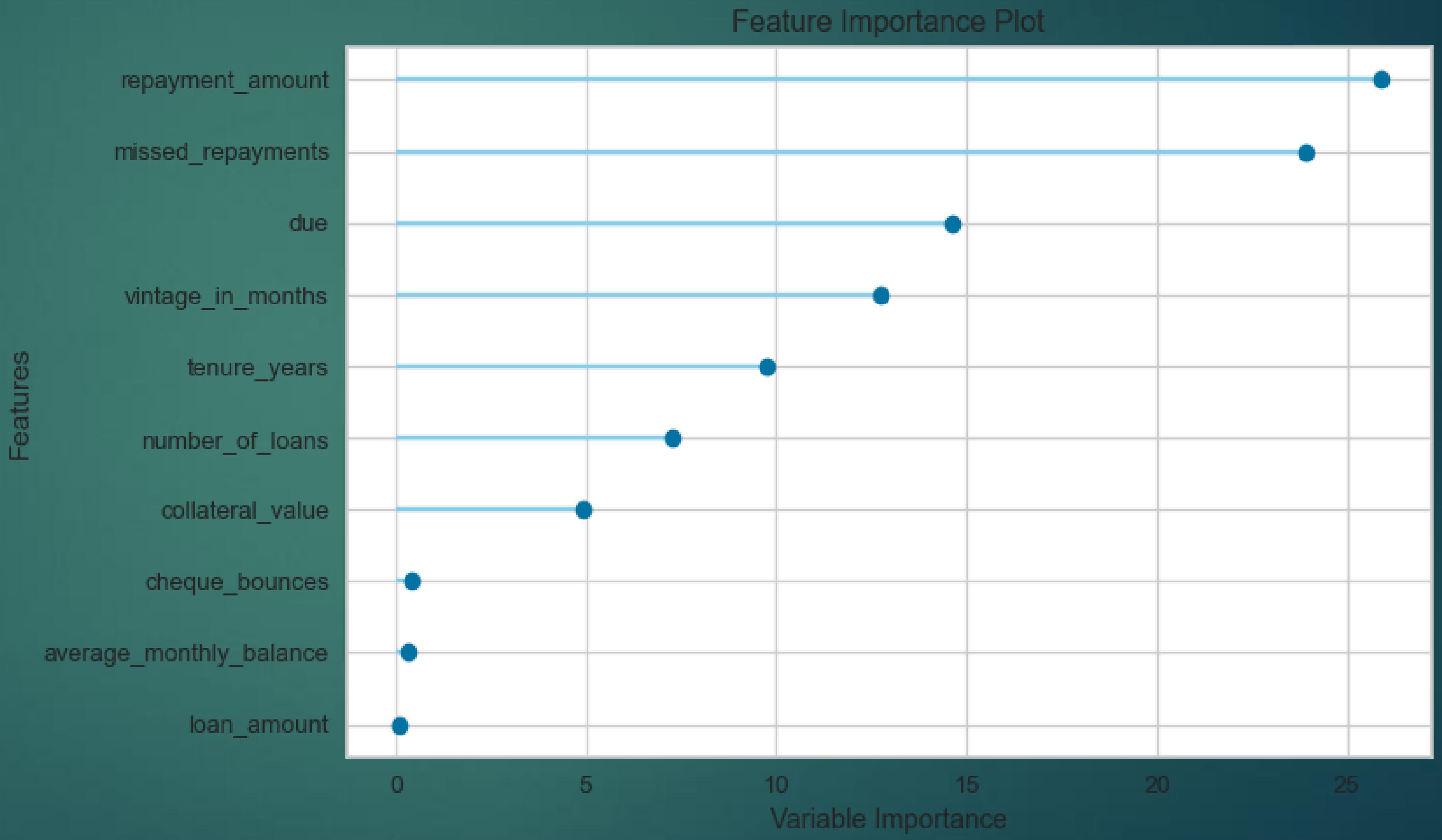
REGRESSION INTERPRETATION

- 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 D efault of the customers.

THANK YOU