Domain Oriented Assignment - Banking Case Study

Insights Document

Submitted by:

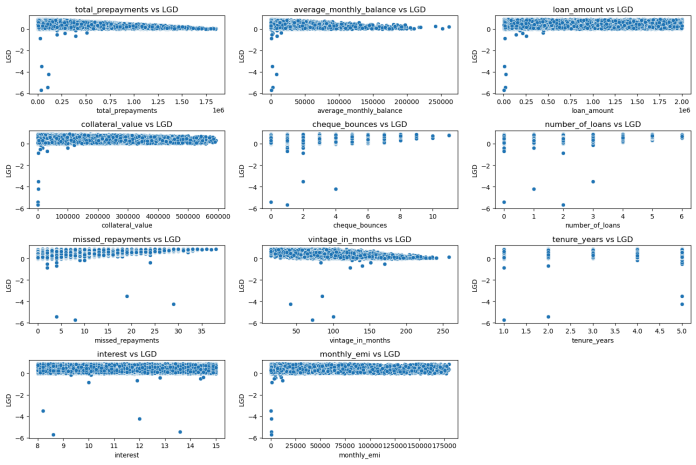
Sankalp Chaudhary
Kunal Salunke
Hanumanth Sai Aditya

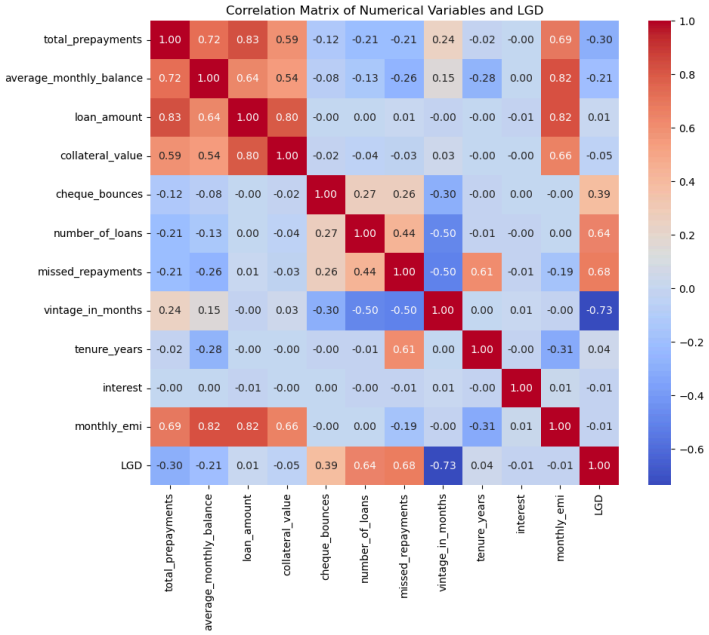
upGrad & IIITB | Data Science Program - April 2023

Our objective was to build a model that can predict the Loss Given Default (LGD) for defaulted accounts.
LGD= {Loan Amount-(Collateral Amount+Total sum of all pre-payments)}/Loan Amount
Where:
- **Loan Amount**: The total amount of the loan or credit exposure.
- **Collateral Amount**: The value of collateral or security provided against the loan.
- **Sum of total prepayments:** The total amount of repayments made on the loan before default.

- We calculated total repayment and average monthly balance against each of the loan amount and then merged the datasets using a left join to the main loan dataset. We then calculated LGD using the above formula and it became our target variable.

Found the correlation matrix of LGD against the variables.





Using RFE we eliminated the high correlation variables and came up with the following equation for simple linear regression machine learning model

Equation of the Linear Regression model:
LGD = 0.0108 * cheque_bounces + 0.0387 * number_of_loans + 0.0202 * missed_repayments + -0.0016 * vintage_in_months + -0.0566 * tenure_years + 0.4723

- **Higher instances of cheque bounces** signal financial instability, suggesting a heightened likelihood of default due to potential cash flow issues.
- **Increasing loan count indicates higher indebtedness**, potentially leading to overleveraged and difficulties managing multiple debt obligations, thus elevating default risk.
- **History of missed repayments reflects poor creditworthiness**
- **Vintage in Months:** Longer tenure as a customer signifies a more established relationship with the bank, potentially indicating a reliable borrower with a lower risk of default.
- **Tenure Years:**Shorter loan tenures may imply higher default risk due to borrowers having less time to repay loans or facing elevated repayment obligations, thus increasing the likelihood of default.

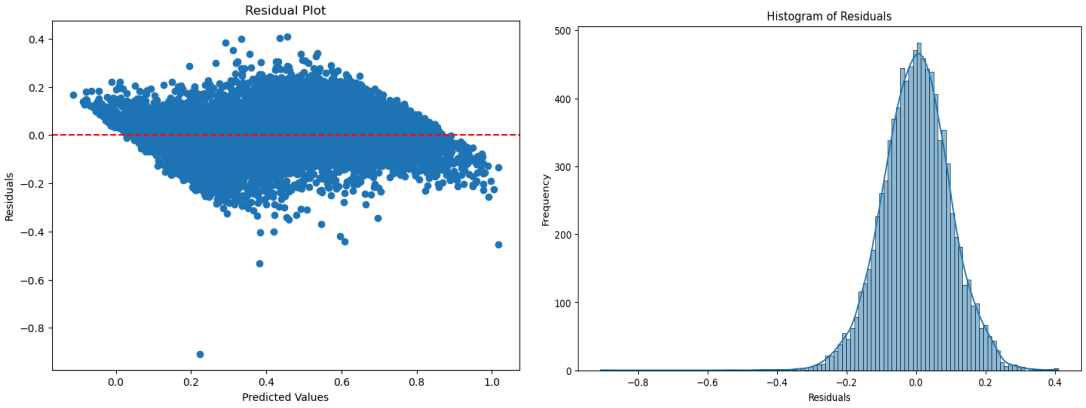
Model Performance metrics

R-squared (R2) Score: 0.82

Mean Squared Error (MSE): 0.0095

Root Mean Squared Error (RMSE): 0.09

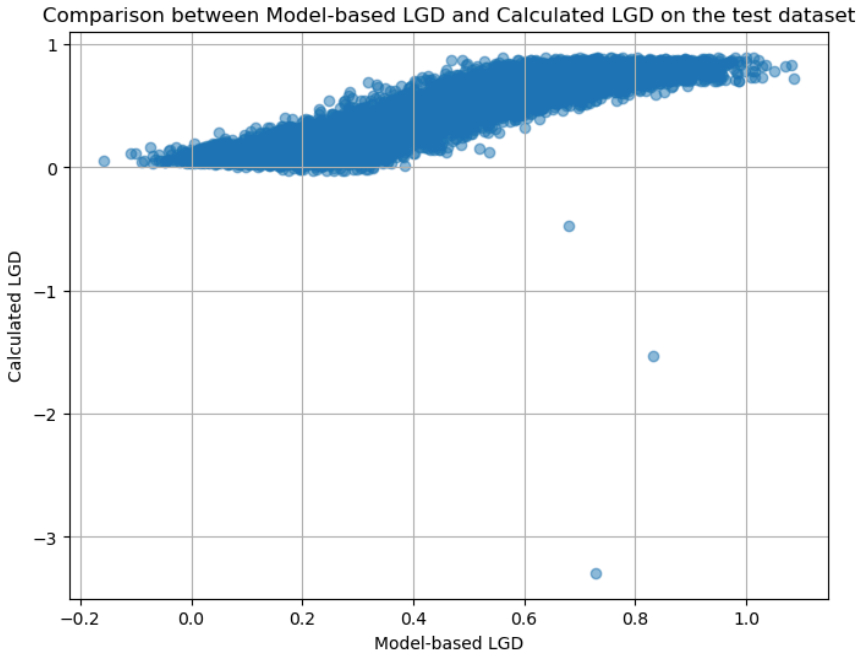
Adjusted R-squared (Adjusted R2) Score: 0.8226231345400887



The model is aligned with the assumptions of Simple Linear regression

Insights on Model Deployment

- We calculated the LGD based on the formula above using the same technique and also calculated the LGD using our prediction model. To compare the accuracy of our predictions we created a plot of calculated LGD and predicted LGD.



As expected the values are between zero and one.

- Mean Absolute Error (MAE): 0.079
- Mean Squared Error (MSE): 0.012
- Root Mean Squared Error (RMSE): 0.1
- R-squared (R2) Score: 0.79
- Pearson Correlation Coefficient: 0.88

The high R-squared score (approximately 0.79) suggests that the model explains a significant portion of the variance in LGD values. A strong positive linear correlation (Pearson correlation coefficient approximately 0.8899) between predicted and calculated LGD values further validates the model's effectiveness.

The model can enhance risk management by providing accurate estimates of LGD, enabling banks to allocate appropriate reserves for potential losses.Compliance with regulatory standards, is facilitated by the model's ability to quantify credit risk and inform decision-making processes.Areas for improvement may include refining feature selection, exploring additional predictive variables, and conducting further validation and testing to ensure robustness.