BFSI Credit Risk Assessment - LGD COMPUTATION

- ## Expected Credit Loss

  Expected credit loss (ECL) computation is a method used in credit risk management to determine the amount of loss a bank is expected to incur in the event a borrower defaults on their loan. Different banks may use different methodologies for calculating the expected credit loss (ECL) and provisioning of a bank. Banks are allowed to use their own methodologies and incorporate factors relevant to their specific business operations.

  The formula for ECL typically used in practice is as follows:
  ECL = EAD x PD x LGD
  Expected credit loss = Exposure at default x Probability of Default x Loss given default

  ECLs are calculated based on the exposure at default (EAD), probability of default (PD) and the loss given default (LGD) for each borrower. Banks can calculate the ECL for different points in time based on their risk management strategy and regulatory requirements.

- ## What is LGD?

  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. In a dataset containing the historical data of loans defaulted, such as the value of the collateral (if any), loan tenure, number of missed repayments, etc. the bank tries to estimate the approximate amount that it stands to lose if a borrower defaults.
  It might seem evident that this can be directly calculated by subtracting the collateral amount and the repayments received against the loan from the present outstanding loan amount.
  However, banks make collection efforts even after a borrower has defaulted. Therefore, the LGD is estimated using historical data from the default loan base using statistical methods.

- ## Business Objective

  To build a statistical model to estimate LGD of borrower for the defaulted accounts.
  For this assignment, we are focusing only on LGD component of ECL computation.

- ## Target Variable

  Target variable is the LGD of the accounts, which is calculated using the following formula: -

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

  The LGD would be represented as a decimal value ranging from 0 to 1. The value represents the proportion of the total loan amount that is expected to be lost in the event of default.

  A value of 0 indicates that no loss is expected, whereas a value of 1 indicates that the entire loan amount is expected to be lost.

  The collateral value and collected amount through repayments need to be utilized to calculate this.

- ## Dataset Description

  - The train_data folder contains three data sets in the form of .csv files titled: main_loan_base.csv, repayment_base.csv, and monthly_balance_base.csv.
  - 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 pertaining to the loan accounts in the main_loan_base.

- ➢ The monthly_balance_base contains the information pertaining to the monthly balance statements in the borrower's accounts.
- ➢ The collection data is provided in a separate data set and needs to be aggregated and merged to gather relevant information.
- ➢ Similarly, we also have TEST datasets provided.

• Analysis and Model Building

1. Cleaning & Pre-processing:

For Main dataset:

- ➢ It is observed that, for **314** accounts, monthly balance main dataset related columns have missing values. This can be replaced with their mean/average value.
- ➢ It is observed that, for **3977** accounts the repayment_amount value is missing. We can infer that the account holder has not paid any instalment. So, we will have to replace NaN value of repayment_amount with 0. This will automatically take care of missing values for due_amount as it is derived from repayment_amount.
- ➢ Here, we cannot perform mean/mode imputation as the value repayment_amount for each account number will depend on respective loan amount, tenure, EMI and interest rate.

For Test dataset:

- ➢ It is observed that for **57** accounts records related to monthly balance dataset is missing in merged data frame. Here, we can replace the null values of mean of all respective columns in the test data set.
- ➢ It is observed that, for **768** accounts the repayment_amount value is missing. We can infer that the account holder has not paid any instalment. So we will have to replace NaN value of repayment_amount with 0. This will automatically take care of missing values for due_amount as it is derived from repayment_amount.
- ➢ Here, we cannot perform mean/mode imputation as the value repayment_amount for each account number will depend on respective loan amount, tenure, EMI and interest rate

2. Derived features:

Repayment amount = Sum of All Repayment Entries for Unique Loan Account Numbers.
Due amount = Loan amount – Repayment amount.
Balance amount = Latest Entry of monthly balance for each unique loan account numbers.
Average Monthly balance
Count entries = Number of entries of monthly balance for each unique loan account numbers.
Month since default = Months of a difference of default date and disbursal date. [default – disbursal]

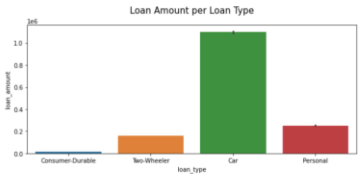
3. Calculated LGD using the formula.

We got **78** rows with negative value of LGD. It is observed that the **Loan Amount < (collateral value + repaid amount).**

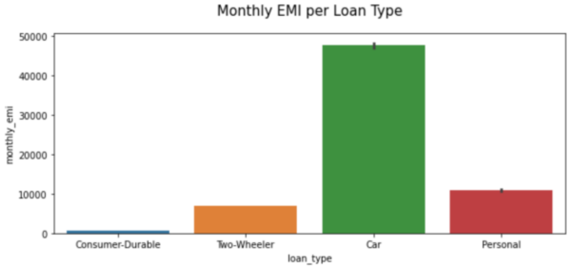
In these cases, Bank/Financial Institute will be able to recover loan amount with help of collateral. These are exceptional cases where even if borrower fails to repay the full amount of loan, lender can recover it by selling collaterals. Hence, it should be feasible to drop rows with negative LGD as it will skew the model results.

LGD values are in the range of 0 to 1.

4. Exploratory Data Analysis (EDA) :

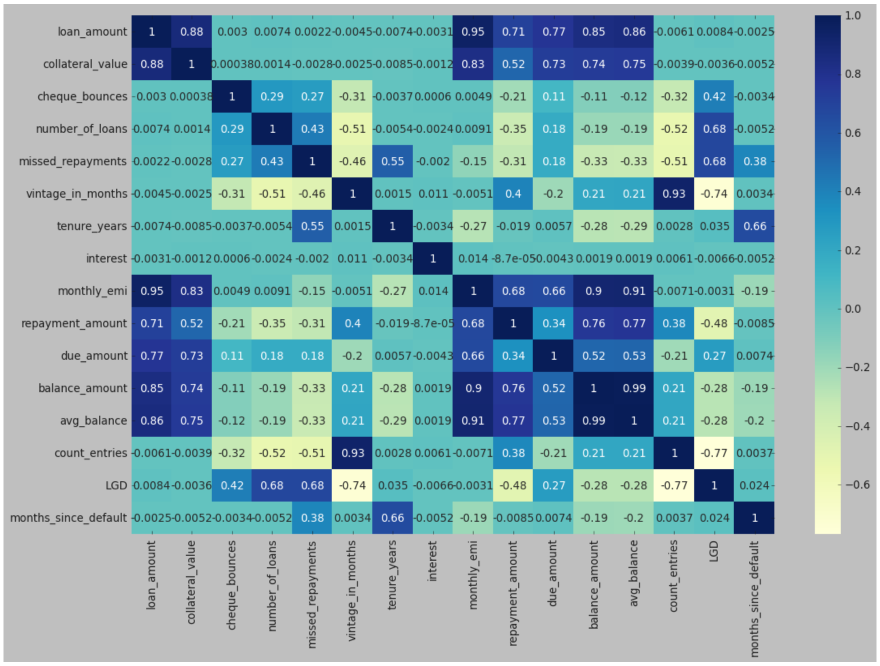


1. Loan amount for Car is higher amongst all.
2. Loan amount for Consumer-durable is small.



1. Monthly EMI for car loan is highest among all.
2. Monthly EMI for Consumer-Durable is small among all.

- ➢ Multivariate analysis for checking the correlation coefficients to see which variables are highly correlated.



- ➢ LGD is highly positively correlated with number_of_loans and missed_repayments.
- ➢ LGD is highly negatively correlated with vintage_in_months and count_entries

5. Performed feature transformation

- ➢ Used **Log transformation** for right skewed variables “monthly_emi","balance_amount", "avg_balance", 'vintage_in_months', "missed_repayments", "months_since_default" and made them normally distributed.
- ➢ Applied **Square Root power transformation** for the rest of the variables ‘cheque_bounces','missed_repayments','vintage_in_months'
- ➢ Applying **Power transformation using fifth root** for 'loan_amount','repayment_amount','collateral_value' to make them normally distributed.
- ➢ Same steps are performed on Test data as well.

6. Performed one -hot encoding technique for Categorical variables in Main and Test data sets.

7. The given target variable is a continuous variable, hence this fall under the category of Supervised Machine Learning so we used the following models:

- Recursive Feature Elimination (RFE),
- Multilinear regression,
- Random Forest Regression,
- Gradient Boost Regression
- Extreme Gradient (XG) Boost Regression

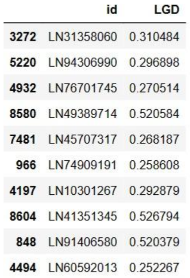
8. After building the model, we evaluated its performance with the Test dataset and got the results as follows:

Model Name	**R^2 value**	**Model**	**R^2 Train**	**R^2 Test**
RFE	0.86	Random Forest Regressor	0.77	0.76
Linear Regression	0.863	Gradient Boost Regression	0.92	0.91
		XG Boost Regression	0.995	0.997

9. We have finalized XG Boost Regression model as it is giving best R2 Metric for both train and test data sets.

10. Applying the model to the unseen data and we have saved the CSV to the local path. Please find the attached "predicted_LGD.csv" in the zipped folder.

11. The predicted LGD values are as follows:



	id	LGD
3272	LN31358060	0.310484
5220	LN94306990	0.296898
4932	LN76701745	0.270514
8580	LN49389714	0.520584
7481	LN45707317	0.268187
966	LN74909191	0.258608
4197	LN10301267	0.292879
8604	LN41351345	0.526794
848	LN91406580	0.520379
4494	LN60592013	0.252267

Insights and Recommendations

1. Significant features affecting estimated LGD are:
   - vintage_in_months - No of months for which borrower is connected with bank
   - missed_repayments - no of instalments missed by borrower
   - no_of_loans - no of other loans customer has
   - tenure_years - tenure of loan repayment in years
   - repayment_amount - total amount repaid by borrower
   - cheque_bounces - no of cheques bounced
2. BFSI should focus on above features in order to control Credit Loss by borrower.
3. LGD is directly proportional to ECL, so the higher the LGD, higher the ECL , so the organization can focus on borrowers with high predicted LGD as they are likely to cause more Credit Loss.
4. In case,
   - When LGD is equal to 0, loan amount can be recovered with Collateral value and chances of BFSI losing the loan amount is less.
   - When LGD is equal to 1, indicates that the entire loan amount is expected to be lost.
5. By using these extreme cases and predicted LGD values, organization can segment borrowers in maybe 2 clusters based on a threshold LGD value.

6. The BFSI organization will not expect any loss and will be able to recover loan amount for borrowers who are below threshold, in case of default.
7. However, BFSI should focus on the borrowers whose LGD is above threshold, for which organization can expect loss in case of default.
8. In order to avoid the Credit Loss, Collateral value should be given more importance.
9. In order to protect banks against possible default of the customers, ECL method is used to provide capital buffer provision for banks. The ECL can be calculated for the new borrowers based on the significant variables and also with the help of PD and EAD. The values can be compared with the standard BASEL III norms , based on which an organization can come up with strict compliance norms which can be followed stringently.

Case Study Submitted by –
1. Divya Shah
2. Deepak Chakala
3. Naga Sai Siva Kumar.
DS C 44 – BA Track