BFSI –Credit Risk Assignment

Domain Overview :

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 risk of a bank.

Banks are allowed to use their own methodologies and incorporate factors relevant to their specific business operations. Some banks may choose to use historical data and statistical models to estimate the components of ECL calculation, while others may rely on expert judgement. The choice of the method can vary depending on factors such as the bank's risk appetite, the nature of the loans and the available data. Additionally, some banks may include certain external factors, such as macroeconomic conditions, in their calculations, while others may not.

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.

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

The LGD value for a loan, given its collateral and assuming that the customer has already made some repayments, is given by:

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

Problem statement:

For this assignment, we will focus on the Loss Given Default (LGD) component of the ECL computation. The objective is to build a statistical model to estimate borrowers' LGD.

As a business analyst working for a bank, you have been tasked with developing a model that can estimate borrowers' LGD. To develop this model, you have been provided with relevant data sets that include information about defaulted accounts and the amount of money that has been retrieved from them using collaterals and other collection methods.

Please note that in the real-worls, the LGD model will be used in conjunction with the PD (estimated using an existing PD model) to calculate the ECL. This will provide the bank with a more accurate understanding of the potential losses from its loans, thus allowing it to make more informed decisions about credit risk management and provisioning. The bank will also be able to identify high-risk loans and take appropriate actions to mitigate the risks.

Steps Followed for develop Model for LGD –Credit Risk:

1. Understand the business problem deeply in the context of the current business and domain.
2. Build a model that can predict the loss given default (LGD) for defaulted accounts and evaluate it based on the performance metric that will be described in the subsequent segments.
3. Thoroughly understand the data sets provided and familiarize with the variables in the data set, data types and data distribution. Ensure that you have a thorough understanding of the target variable, which is the LGD of the accounts, and how it is calculated. The collection data is provided in a separate data set and needs to be aggregated and merged to gather relevant information. It is essential to ensure that the data types of the variables are identified correctly, as Python is notorious for running into errors owing to data type mismatch errors.
4. Clean and pre-process the data. This includes handling missing values, outliers and any other issues that may affect the model's performance. Also, perform any necessary feature engineering or feature extraction to create new variables that may be useful for the model.

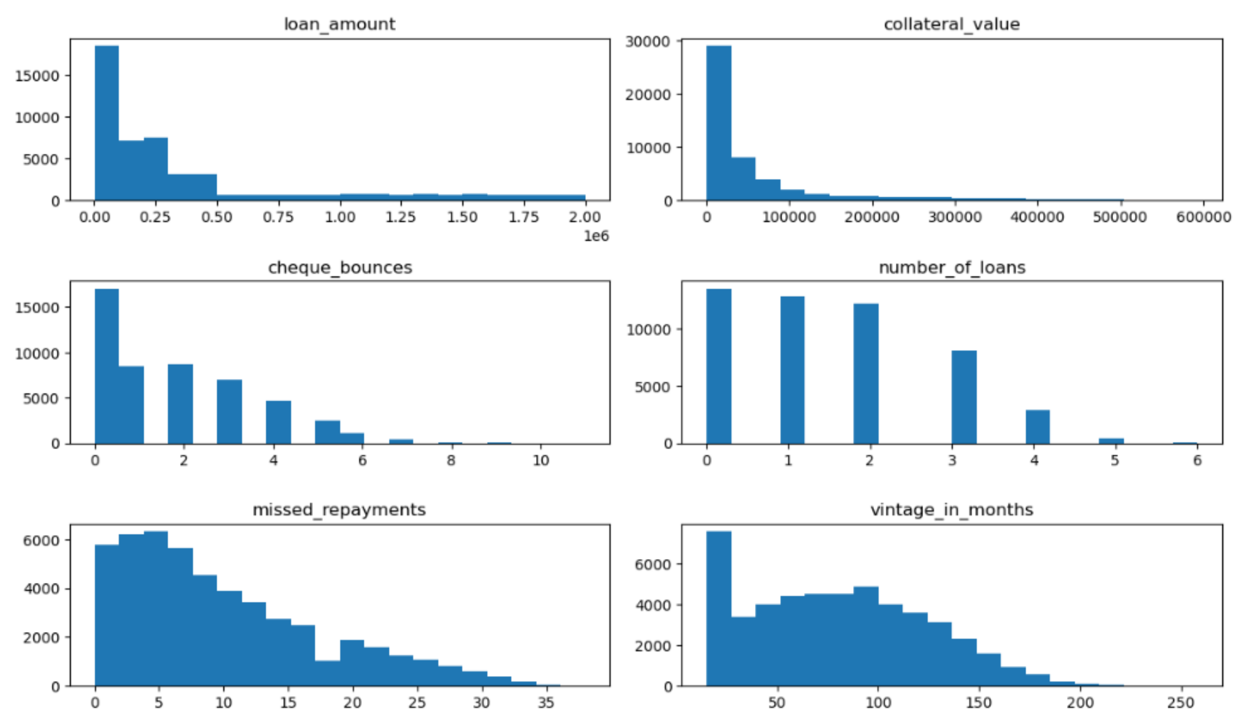
5. The target variable LGD is not specified directly and, hence, Calculate the LGD using the following formula:

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

6. Used multiple statistical and machine learning technique to develop a model for predicting the borrowers' LGD.

Data Visualization:

Loan amount, collateral, monthly_emi, repayment have outliers. The majority of the data is concentrated on the left side, it is an asymmetrical distribution. Mean would be a better representation of the central tendency here.



Avg monthly balance, repayment total, vintage in months have negative correlation with target.

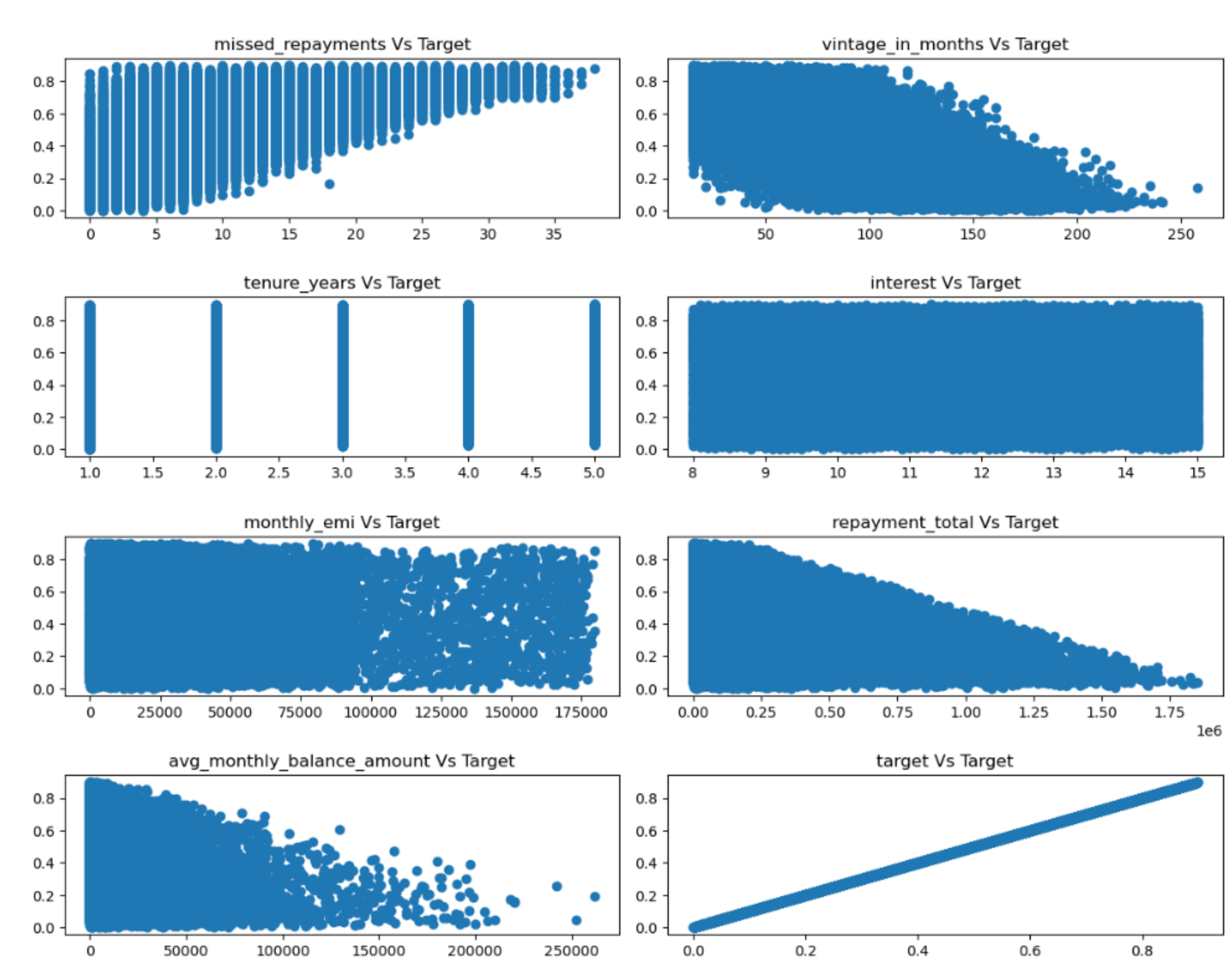
Customers with higher monthly balance to have lower chances of defaulting the loan.

Customers with lower repayment total that means with higher due amount are likely to have higher chances of defaulting hence higher LGD.

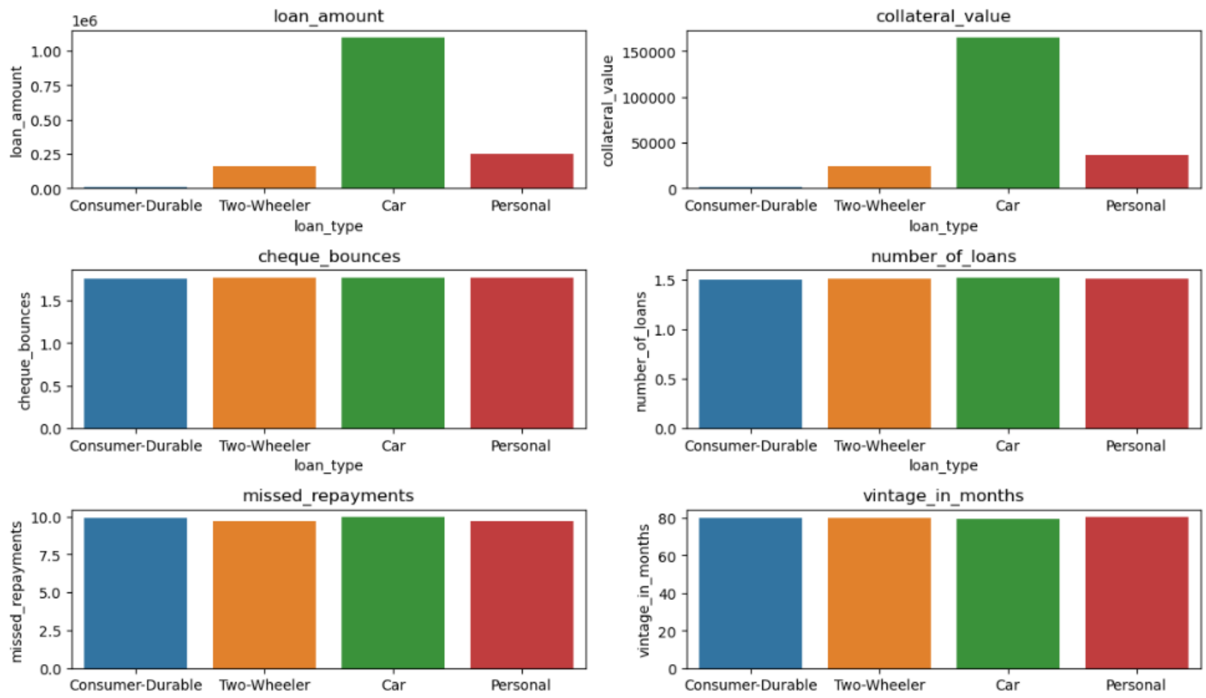
The longer a customer is associated with the bank, the lesser chances of defaulting hence lesser LGD.

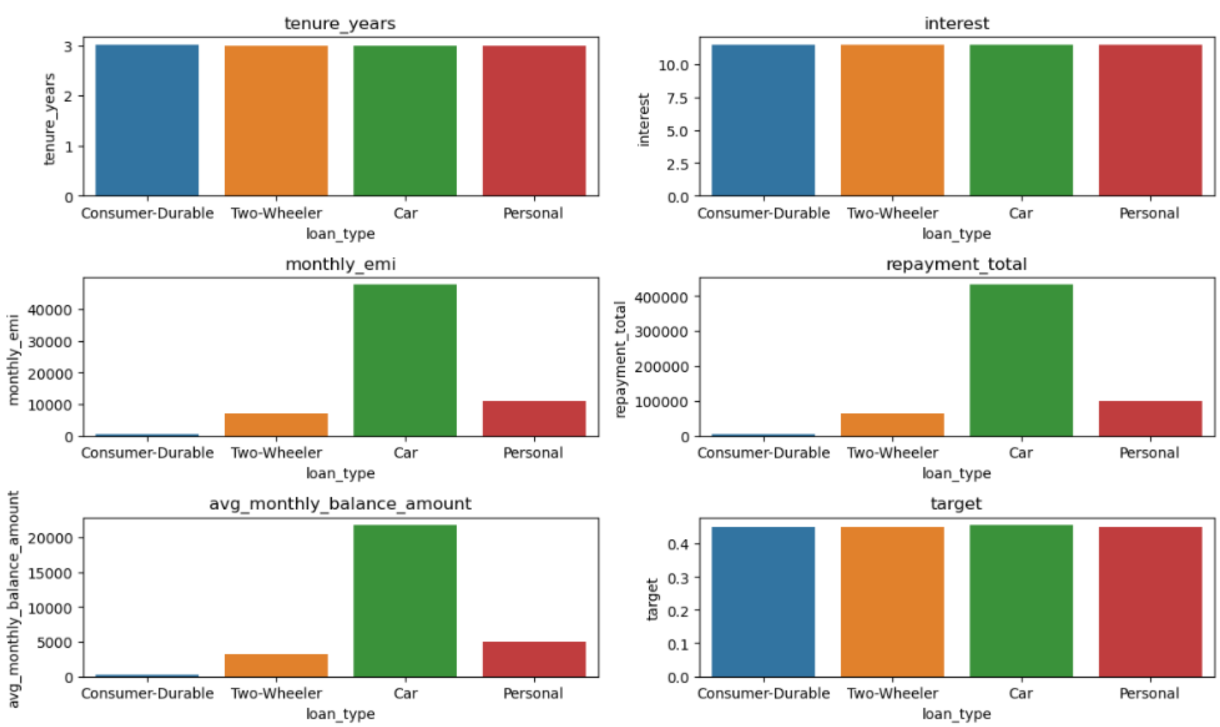
The risk of LGD is same till the no of loans are 3. The risk of LGD increases when the no of loans are higher than 3.

The higher the missed repayments, the higher chances of LGD.

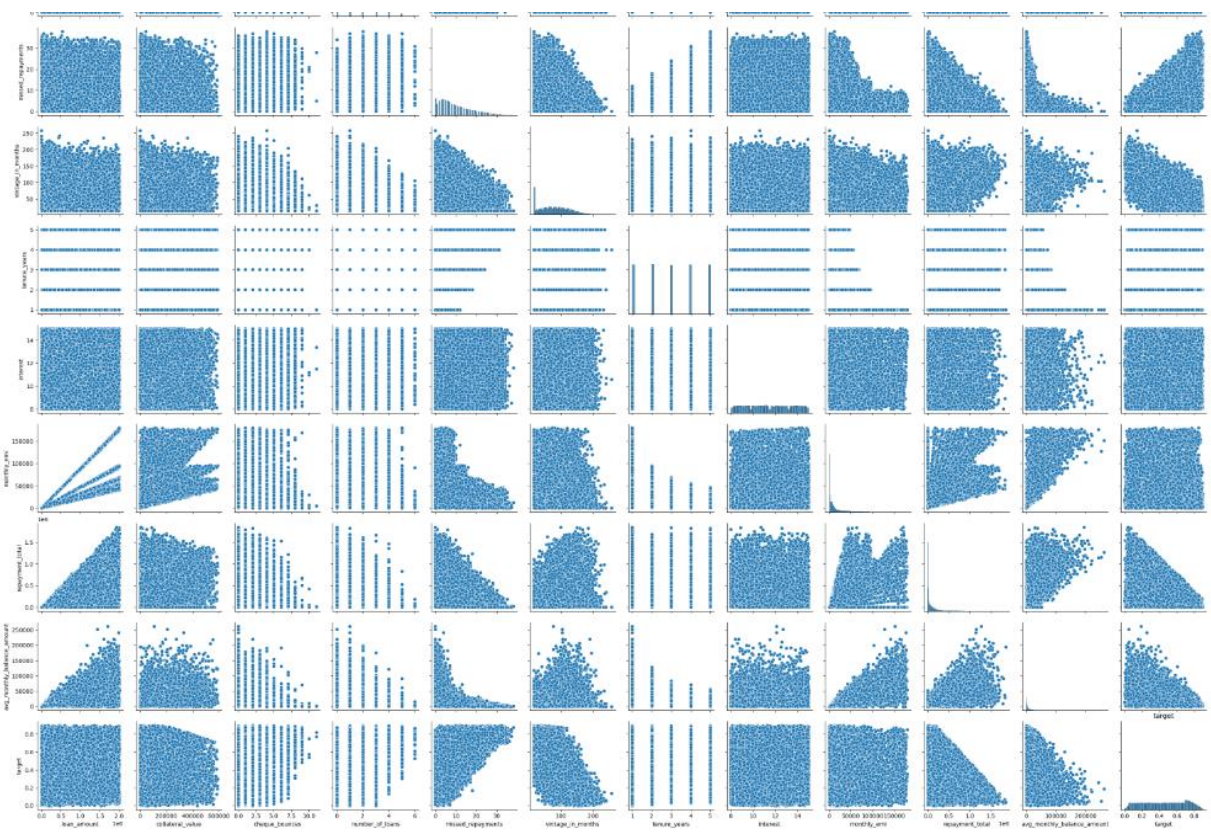


The maximum loan amount was given for 'Car' category. The amounts for repayment, monthly balance, collateral have similar trend as loan amount.

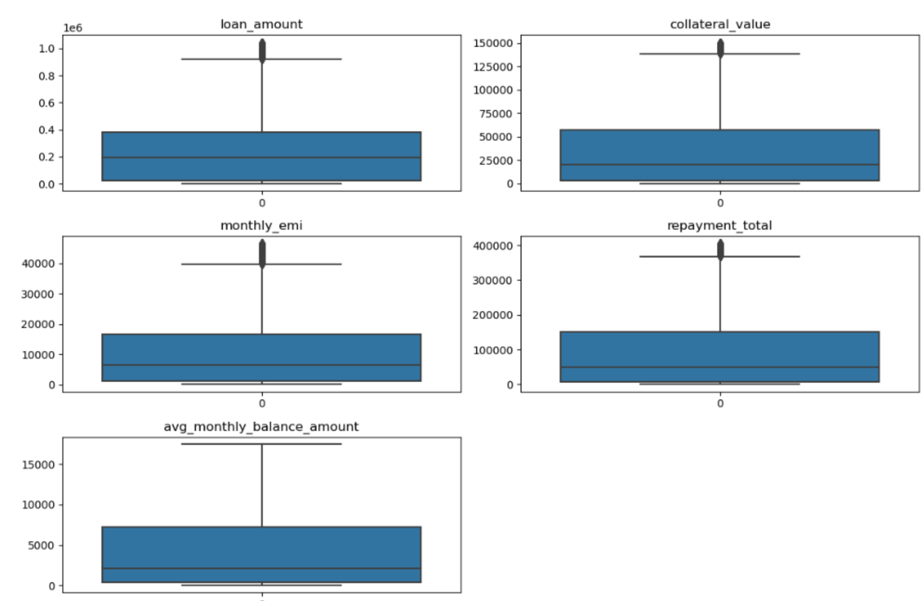




There is correlation between few independent variables.



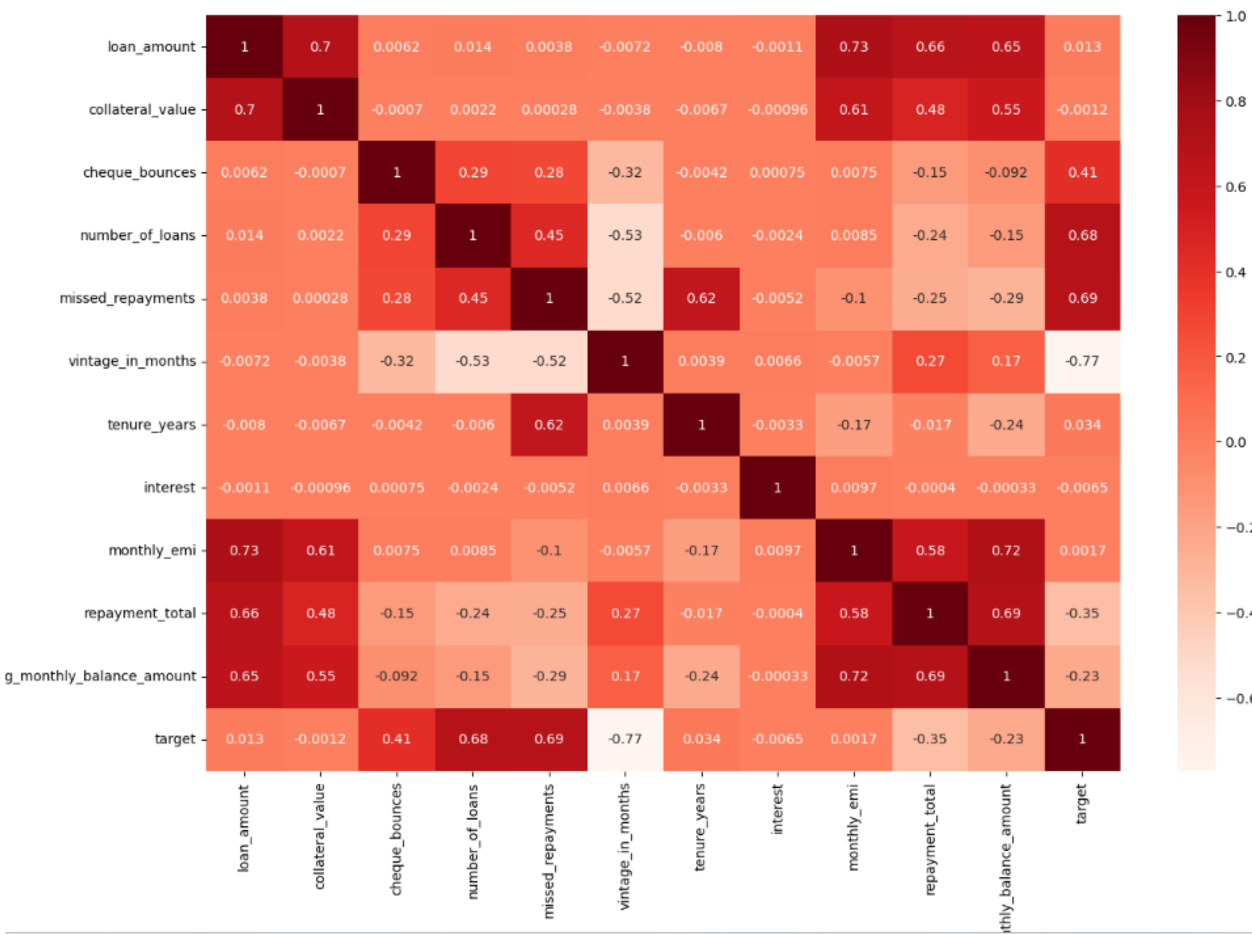
Outliers are replaced with the mean value.



The Target variable (LGD) has a high positive correlation with "Number of Loans", "Missed Repayments" and a high negative correlation with "Vintage in Months“.

The customers associated with the bank for a long period of time tend to have a lower probability of default and hence the proportion of LGD decreases due to them paying back their loan on time.

People who have a longer loan tenure tend to have more missed repayments.



Various Model Evaluation value:

	ML Model:	R Square
0	Multiple Linear Regression	0.9056
1	Multiple Linear Regression with RFE	0.8022
2	Gradient Boosting Regressor	0.9006
3	Gradient Boosting Regressor - Hyperparameter t...	0.9354
4	XGBoost Regressor	0.8989
5	XGBoost Regressor - Hyperparameter tuned	0.9438
6	Random Forest Regressor	0.9074
7	Random Forest Regressor - Hyperparameter tuned	0.9056

Conclusion:

Best fit model : XGBoost Regressor with Hyperparameter tuning has given the best performance of 94% on the unseen data.

Feature importance : The features selected by the XGBoost Regressor with Hyperparameter tuning

Vintage_in_months: The model considers vintage_in_months as a significant predictor for predicting lower losses given default for customers with longer relationships with the institution.

No of loans: Customers with a higher number of loans may exhibit distinct patterns in loan outcomes.

Missed repayments: This implies that customers who have a higher number of missed repayments are likely to face larger losses in the event of default. It suggests that a history of missed repayments may indicate a higher level of credit risk and a greater likelihood of experiencing significant losses.

Tenure years: The model suggests that the loss given default risk is lesser for the loans with higher tenure. The loans with lower tenure are more likely to be default.

Repayment total: The higher the repayments are done, the lesser is the loss given default.

By considering the impact of each feature, the model can help identify customers with the loss given default value. This will help the business to make more informed decisions related to risk management and mitigation strategies.

Recommendations to the business:

Customer Segmentation: The business can use the features to segment customers based on their risk profile. This segmentation can help tailor risk mitigation measures.

Risk Assessment: Assign appropriate risk levels based on factors such as vintage_in_months, number of loans and missed repayments. This can help manage the risk-reward tradeoff and optimize profitability while minimizing potential losses.

Early Warning System: Monitors key features such as missed repayments and number of loans to identify customers likely to default and offer them financial counseling or restructuring options. This approach can help reduce the likelihood of default and mitigate potential losses.

Customer Retention: Offer incentives, loyalty programs, or personalized services to encourage long-term relationships with the bank. The long tenure lowers the default risk.

Proactive measures: Revise collateral requirements or loan terms for customers with higher risk profiles, such as those with a higher number of loans or missed repayments.