

# BFSI: CREDIT RISK ASSIGNMENT


## PREPARED BY:

RISHIKA BANSAL

SARTHAK VERMA

RITA PILAYI

VARUN AGGARWAL

- 
- ▶ Objective
  - ▶ Background
  - ▶ Data Analysis
    - Pre Processing of Data
    - EDA
    - Model Building
    - Interpreting the Results
    - Recommendations

# OBJECTIVE

PREDICTING LOSS GIVEN DEFAULT (LGD): THE PRIMARY BUSINESS OBJECTIVE IS TO DEVELOP A PREDICTIVE MODEL FOR LOSS GIVEN DEFAULT (LGD) FOR DEFAULTED ACCOUNTS. LGD REPRESENTS THE PROPORTION OF THE EXPOSURE THAT A LENDER IS UNABLE TO RECOVER AFTER A BORROWER DEFAULTS. ACCURATE PREDICTION OF LGD IS CRUCIAL FOR RISK MANAGEMENT AND FINANCIAL PLANNING.

PERFORMANCE EVALUATION: EVALUATE THE PREDICTIVE MODEL BASED ON A DEFINED PERFORMANCE METRIC. THE PERFORMANCE METRIC SHOULD ALIGN WITH BUSINESS GOALS AND ACCURATELY MEASURE THE MODEL'S EFFECTIVENESS IN PREDICTING LGD.

UNDERSTANDING DATA SETS: GAIN A DEEP UNDERSTANDING OF THE PROVIDED DATA SETS, INCLUDING VARIABLES, DATA TYPES, AND DISTRIBUTIONS. RECOGNIZE THE RELEVANCE OF EACH VARIABLE TO LGD PREDICTION. THIS UNDERSTANDING IS ESSENTIAL FOR BUILDING A MEANINGFUL AND ACCURATE MODEL.

DATA AGGREGATION AND MERGING: AGGREGATE AND MERGE RELEVANT INFORMATION FROM THE COLLECTION DATA SET TO ENHANCE THE PREDICTIVE POWER OF THE MODEL. THE MERGED DATA SHOULD PROVIDE COMPREHENSIVE INSIGHTS INTO THE FACTORS AFFECTING LGD.

DATA TYPES AND QUALITY: ENSURE ACCURATE IDENTIFICATION OF VARIABLE DATA TYPES TO PREVENT DATATYPE MISMATCH ERRORS. CLEAN AND PREPROCESS THE DATA TO HANDLE MISSING VALUES, OUTLIERS, OR ANY ISSUES THAT MIGHT NEGATIVELY IMPACT THE MODEL'S PERFORMANCE.

FEATURE ENGINEERING AND EXTRACTION: CONDUCT FEATURE ENGINEERING AND EXTRACTION TO CREATE NEW VARIABLES THAT COULD ENHANCE THE MODEL'S PREDICTIVE CAPABILITIES. INTRODUCE FEATURES THAT CAPTURE THE INTRICACIES OF DEFAULTED ACCOUNTS AND CONTRIBUTE TO A MORE ACCURATE LGD PREDICTION

# 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.

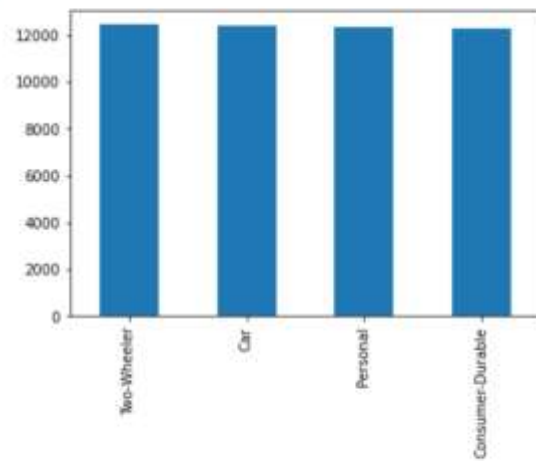
# PRE PROCESSING OF DATA

- ▶ 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

```
In [315]: # Checking the distribution of loans by loan type  
df_train["loan_type"].value_counts().plot.bar()
```

Out[315]: <AxesSubplot:>



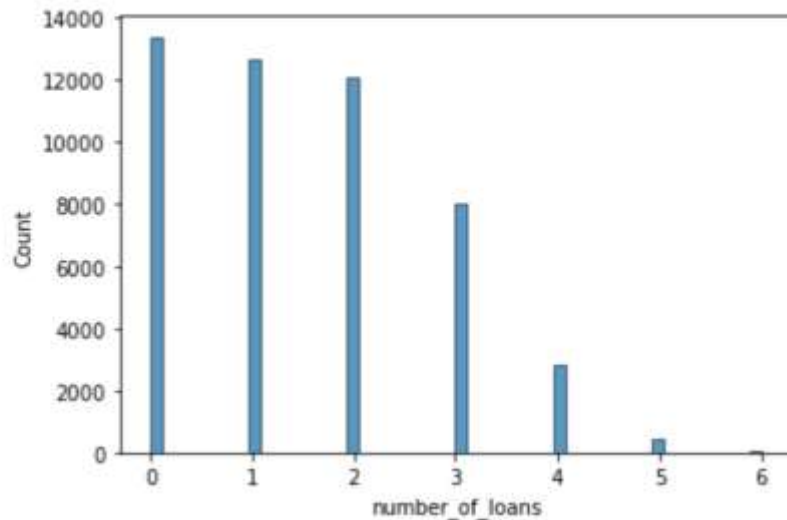
In terms of loan types all loan types have a equal distribution

# EDA

Majority of customers have 1- 3 loans, which shows that they are living off high credit

```
In [316]: sns.histplot(df_train["number_of_loans"])
```

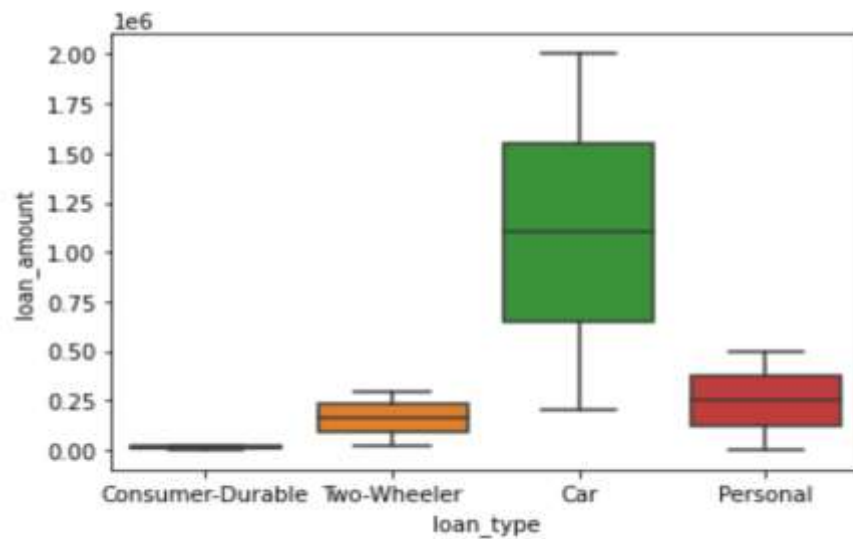
```
Out[316]: <AxesSubplot:xlabel='number_of_loans', ylabel='Count'>
```





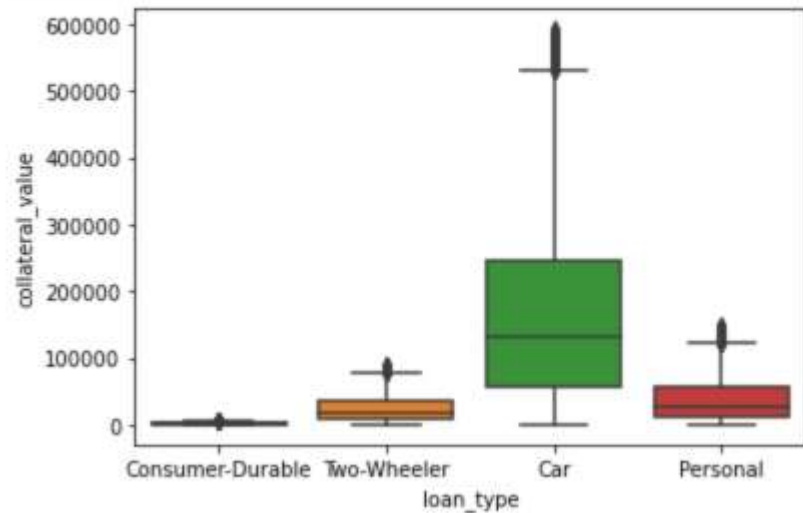
The loan amount for car loans are the highest even though car loans contribute to ~25% of loans

```
#Bivariate Analysis  
sns.boxplot(data=df_train,y="loan_amount",x="loan_type")  
plt.show()
```



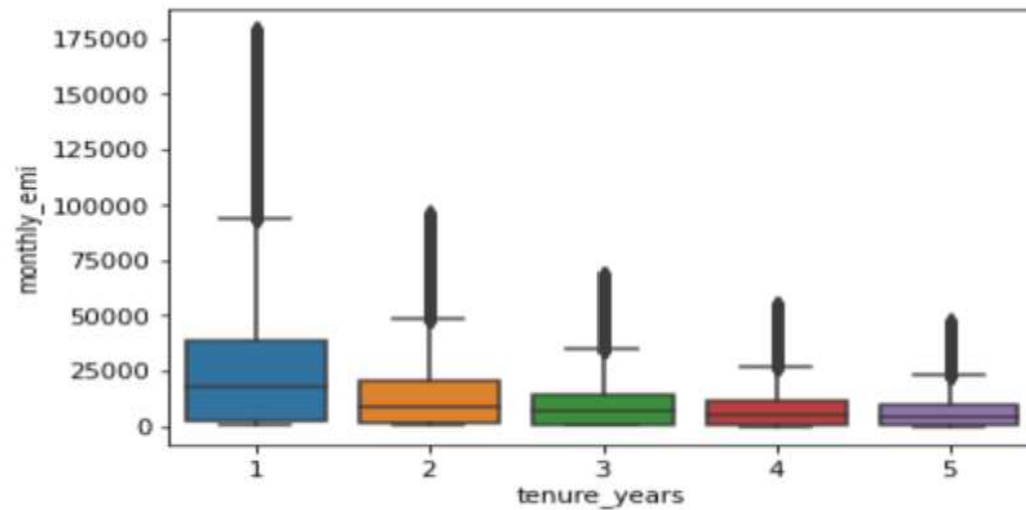
Similar distribution for collateral values show that car loans require higher collateral (e.g. The car itself)

```
#Bivariate Analysis  
sns.boxplot(data=df_train,y="collateral_value",x="loan_type")  
plt.show()
```



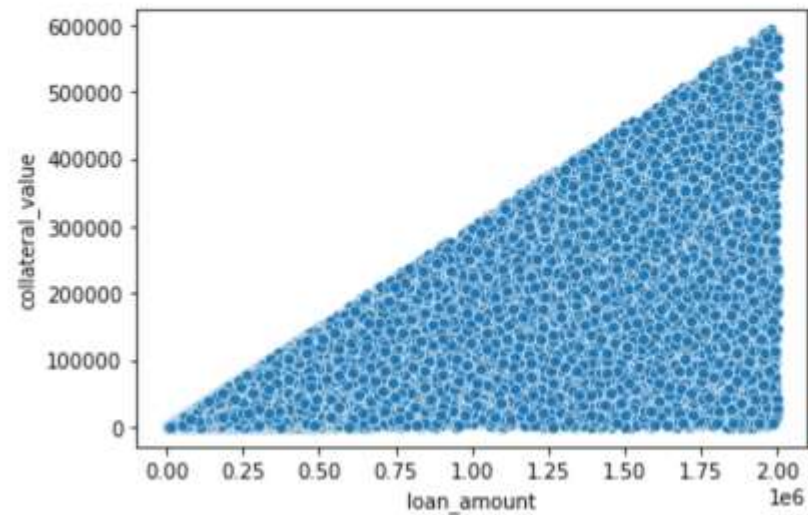
As expected longer loans have lower monthly emi

```
#Bivariate analysis  
sns.boxplot(data=df_train,y="monthly_emi",x="tenure_years")  
plt.show()
```



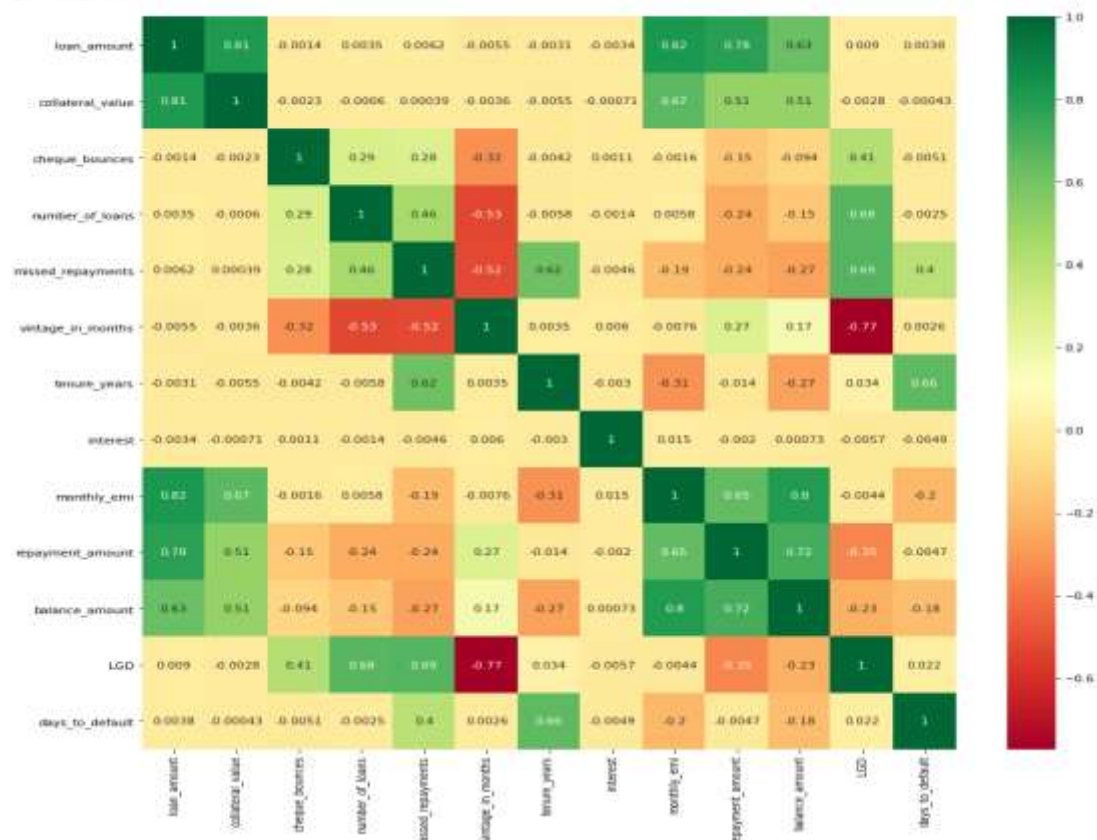
Higher Loan amounts require higher collateral values

```
#Bivariate Analysis  
sns.scatterplot(data=df_train,x="loan_amount",y="collateral_value")  
plt.show()
```



```
In [327]: #Checking for correlation bewteen variables
plt.figure(figsize=(15,15))
sns.heatmap(df_train.corr(),cmap="RdYlGn",annot=True)

Out[327]: <AxesSubplot: >
```



The above matrix shows that:

Long term customers have lower missed repayments and lower LGDs which makes them better candidates for loans

Higher missed repayments, Number of loans, cheque bounces and LGD are positively correlated which shows that customers which have higher missed payments and #loans tend to greater loss

Customers with higher bank balance tend to less loss at default

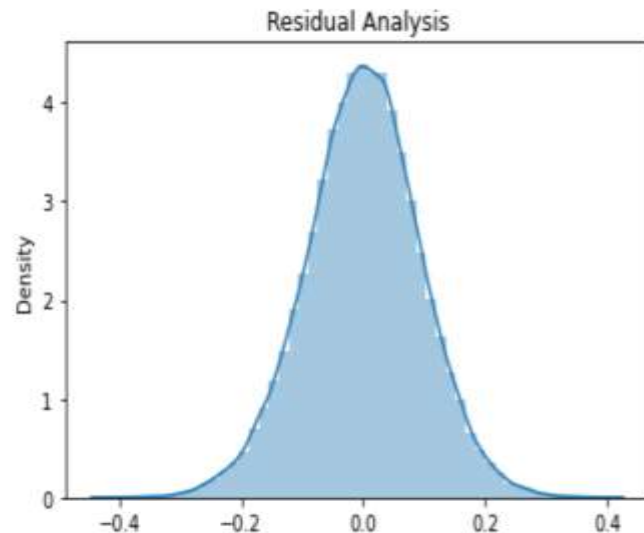
# MODEL BUILDING

- USED VARIOUS MODELS LIKE MULTIPLE LINEAR REGRESSION, RANDOM FOREST REGRESSOR, RESIDUAL ANALYSIS OF THE MODEL  
USED R SQUARED AS A PERFORMANCE METRICS.
- XG BOOST HAS GIVEN US 99.5% R SQUARED ON TEST DATA ACROSS THE MODELS.

# REGRESSION INTERPRETATION

## RESIDUAL PLOT OF THE FINEST MODEL

```
In [800]: res=y_train-y_train_pred  
sns.distplot(res)  
plt.title("Residual Analysis")  
plt.show()
```



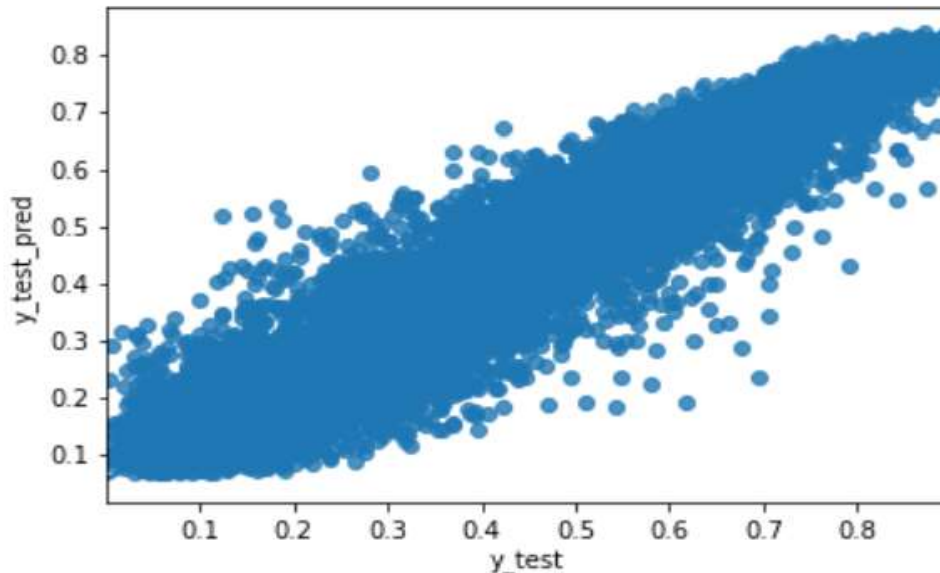
The Residual analysis is normally distributed with a mean of 0 which shows that the model is a good fit

# REGRESSION INTERPRETATION

BEST FIT LINE CORRESPONDING THE PREDICTION ERROR

```
: #Plotting the y_pred vs y_test
```

```
sns.regplot(x=y_test,y=y_test_pred_rf)  
plt.xlabel("y_test")  
plt.ylabel("y_test_pred")  
plt.show()
```





# Conclusions

- WE DEVELOPED MULTIPLE REGRESSION MODELS: MLR AND RF TO CALCULATE LGD
- THE RF MODEL IS A BETTER PREDICTOR OF LGD WITH GIVEN INPUT VARIABLES
- WE HAVE HYPER TUNED THE PARAMETERS OF THE RF MODEL AND THE RF\_BEST IS WHAT WE HAVE SELECTED AS FINAL MODEL



**Thank you**