ADVANCED DATA MINING AND REDICTIVE ANALYTICS

REPORT

**Group 6:**

**Individual Contribution:**

**Niharika Dobanaboina:** Collinearity check, Model selection, and implementation for Regression Analysis

**Zachariah Alex:** Model selection and implementation for classification analysis.

**Rachana Kurra:** EDA, feature selection for regression, and presentation.

**Tilak Kumar Bonala:** EDA, features selection for classification, and Report.

**The goal of the Project:**

The objective of this project is to minimize investment risks and losses while maximizing profits for financial investors. This will be achieved by analyzing client profiles and their probability of loan default. The analysis will evaluate the profits and losses incurred by investors due to defaulting customers, including loan amount, interest rates, and tenure period. Based on the capital investment and loss incurred, investors will determine the feasibility of lending money to customers with a lower probability of default, thereby ensuring a steady flow of interest and principal gains. This analysis will enable investors to make informed decisions on loan approvals while balancing the potential profits and risks associated with lending money to both non-defaulting and defaulting customers.

**Overview of data, including data exploration analysis**

The bank loan dataset contains columns with unique customer identification numbers (IDs) and a column for loss. However, the rest of the columns in the dataset have poor labeling, making it difficult to identify the variables and their purpose. The project objective is to reduce financial investor losses by analyzing the per capita income and investment risks associated with each client.

The dataset consists of 80000 rows and 764 variables, with 17.5% of the columns containing missing values. Specifically, 521 columns contain 40 570 observations with null values, accounting for less than 50% of the total null values, and 10 variables have no variance. To prepare the data for analysis, the first step was to identify the variables with numerical values. Subsequently, the columns with the highest percentage of missing values were removed, and the remaining null values were imputed with the median value.

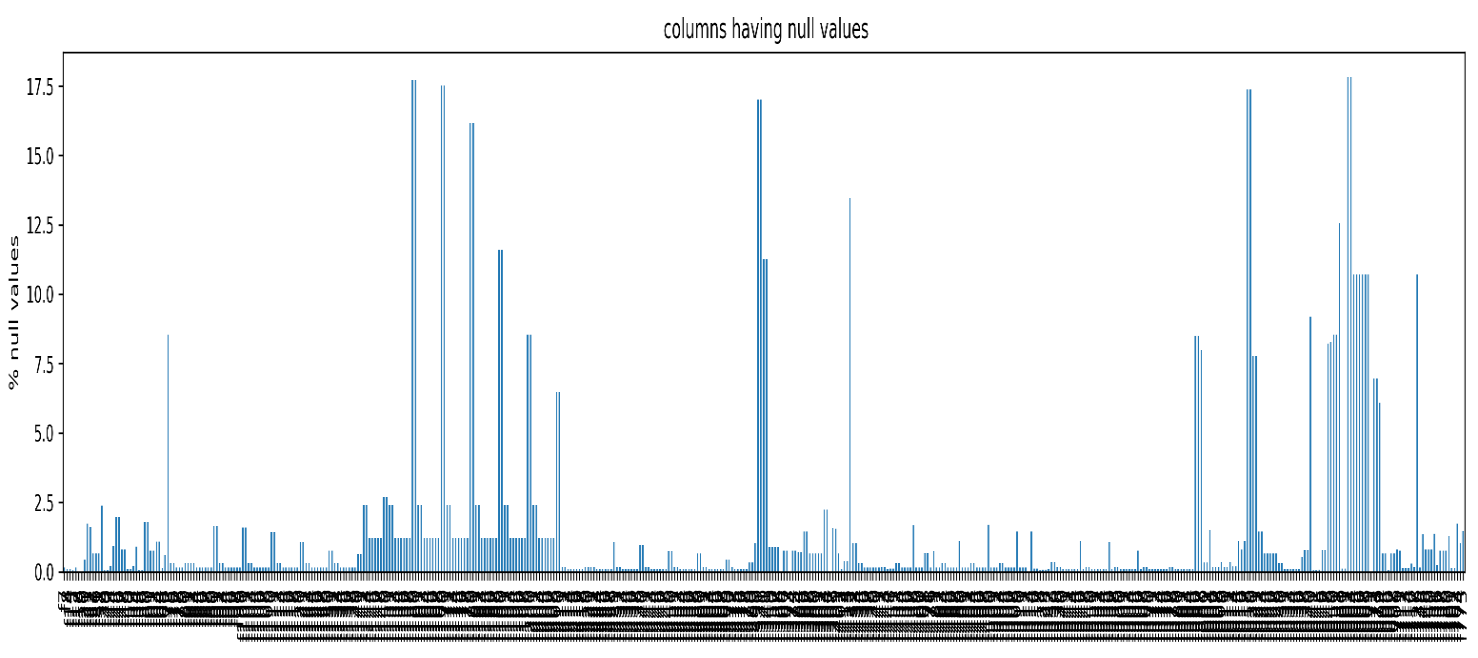
Overall, this approach aims to create a clean and organized dataset that can be effectively analyzed to identify the variables that have the most significant impact on financial investor losses and enable data-driven decision-making.

Figure 1. Null value distribuion

After preparing the dataset, we conducted a correlation analysis between the variables. Interestingly, we observed that the loss column in the dataset had two distinct values, namely "ZEROS" and "NUMERICAL VALUE." The presence of zeros in the loss column indicates that certain clients/customers were able to pay their loans without defaulting, while the numerical values correspond to the percentage of loan default for other customers.

The data distribution image Figure 2 gives us information that the percentage of clients who paid the complete loan maximum than the defaulters count. Considering that the customers had low-interest rates, they may have a steady income and other financial support.

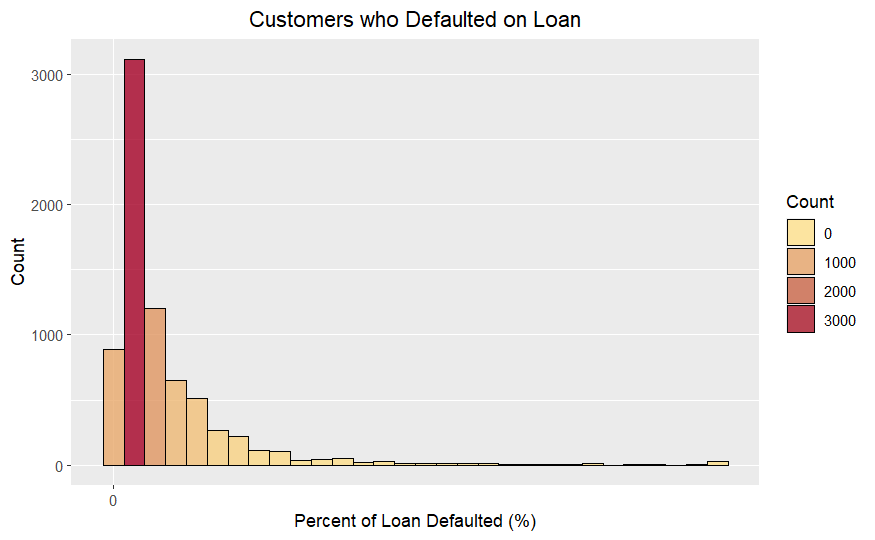


Figure 2. Percentage of defaulted customers

Defaulted and Non-Defaulted clients/ customers

As per the data we classified customers into defaulted and non-defaulted out of 80000 clients/customers 72621 customers cleared their loans and only 7379 customers needed to pay their loans/ defaulted. The plot of Figure 3 says that 9.22% is the default rate for a given data set.

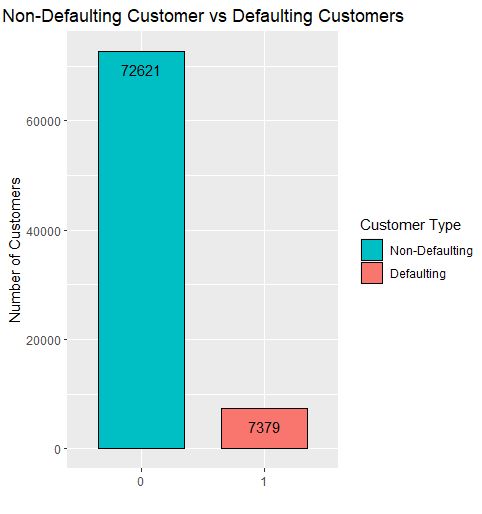


Figure 3. Number of defaulted and non-defaulted customers.

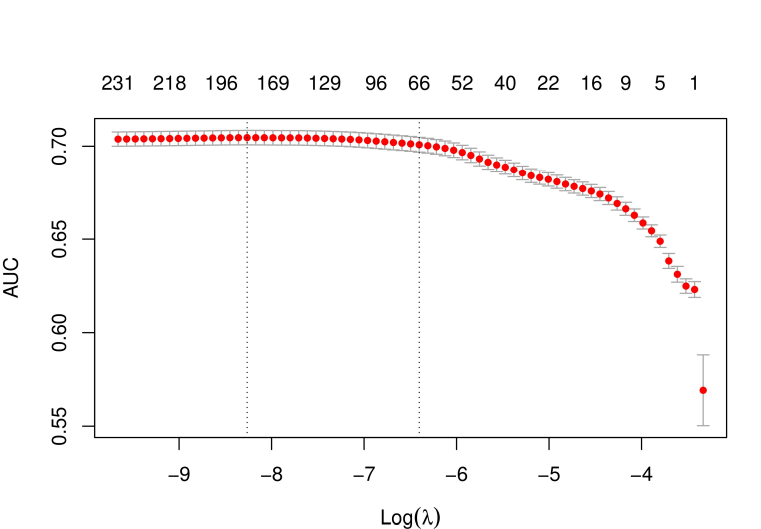
**3. Details of our modelling strategy**

After removing the null values and making the data ready for analysis, we created a new **Default** column with values of 1’s and o’s based on loss value. if the loss is 0 then the default value is 0 and if the value in the loss column is other than 0 then the default column is assigned to be 1 to classify customers into default and non-default.

3.1 **Classification Model:**

***Feature Selection:***

To select the most significant variables, we run a lasso regression analysis on 261 variables for features selection. After running the lasso model, out of 261 variables, 78 variables were eliminated and only 183 variables were selected for further analysis.



***Data Partition:***

We partitioned the data into Training and Validation with 85% and 15% respectively. Our idea behind this was to have more data samples for Training.

***Model Selection:***

We choose Gradient Boosting Machine (GBM) algorithm with 10-fold cross-validation.

to train our model having followed hyperparameters:

Number of Trees: 100

Interaction Depth: 9

Shrinkage: 0.02

Bag Fraction: 0.8

Train Fraction: 0.85

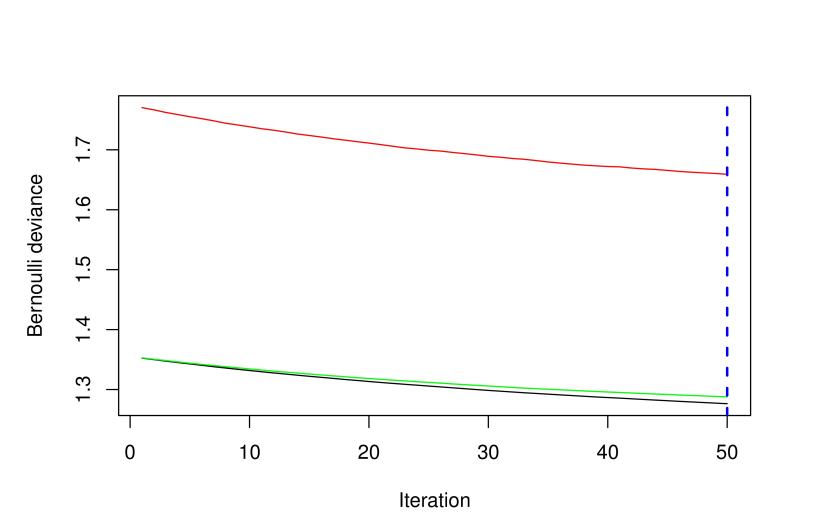
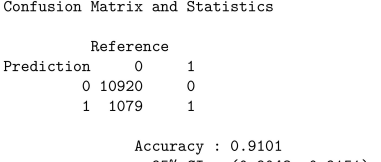


Figure 5. Iterations Deviance

Bernoulli Distribution plot of the model shows that the model overfits beyond 50 trees. Hence, the number of trees was set to 50 and the model was re-run again.

***Testing on Validation Data:***

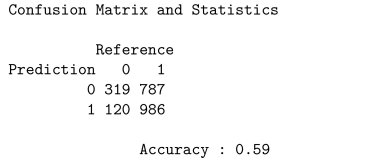
We deployed our GBM model on Validation data and calculated the Confusion Matrix:



As we had imbalance classes, we chose to focus on Precision over Accuracy. Here, precision was found to be as low as 0.5%. This was expected due to the huge difference in data distributions across both classes.

**Approach 2: Stratified Sampling**

We re-sampled the data in such a way that both classes have an equal number of data samples which resulted in a new dataset having 14,758 records. This dataset again was partitioned into Training and Validation with 80% training and 20% Validation samples.

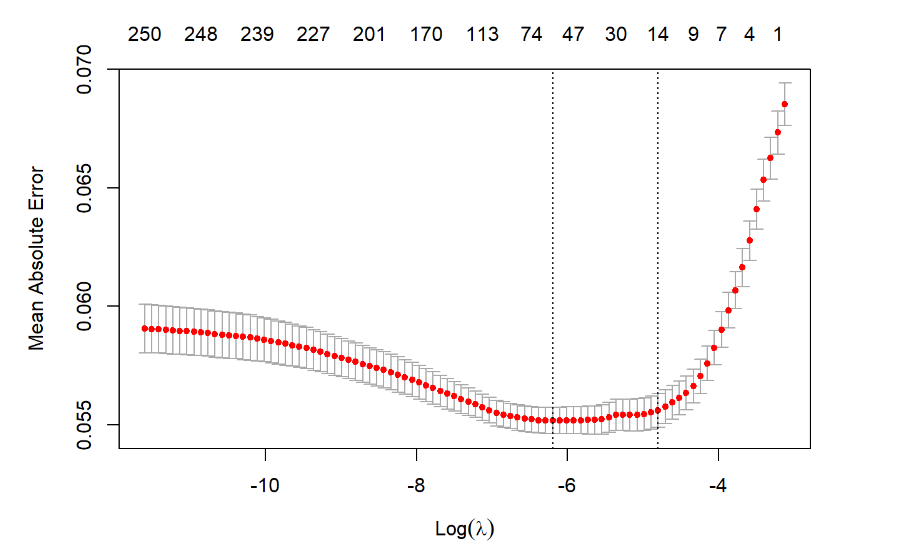
A new GBM model was run on the training set, and ConfusionMatrix was developed. 

Precision: 90.77%

**3.2 Regression Model:**

***Feature Selection:* Lasso Model**

A new Lasso model was developed to select features based on the LOSS column. Only 104 columns were selected for further analysis.



**Figure 6. Lambda vs MAE**

***Data Partition:***

We partitioned the data into Training and Validation with 70% data samples in Training and 30% samples in VALIDATION.

***Model Selection: Random Forest***

The random Forest model was implemented using the following Hyper-parameters:A graph showing the number of trees in bagged decision

Description automatically generated with medium confidence

***Testing the Model:***

RF model was deployed on Validation data and MAE was calculated.

MAE was observed to be 6%.

**Deploying the results of both Classification and Regression Models onto TEST dataset:**

After loading the TEST dataset, we implemented classification Model (GBM model) on TEST dataset to classify customers into default or non-default categories.

As a result, we got probabilities of customers being default. We chose a Classification Threshold of 0.3 as we wanted to identify every customer who had the slightest chance of being defaulted.

After identifying customer records which were defaulted, a subset of these records was created to further deploy Regression Model (RandomForest) on it.

On this subset of data which has only customers who would default, we deployed a RandomForest model to calculate the percentage of loss a bank/company can incur if loan is approved.

These probabilities were recorded into a CSV file and are submitted.

**Conclusion:**

Investors may successfully reduce investment risks and increase returns while making well-informed judgments on loan approvals by choosing significant factors utilizing feature selection approaches and machine learning algorithms like GBM and Random Forest. The findings of this initiative, offer useful information that financial institutions and investors may use to make data-driven choices in the lending sector.