# FINANCE AND RISK ANALYTICS PROJECT REPORT

**Prepared By: Barkha Agarwal**

## PROBLEM STATEMENT-1

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

Data that is available includes information from the financial statement of the companies for the previous year.

Explanation of data fields available in Data Dictionary, **'Credit Default Data Dictionary.xlsx'**

Perform the following in given order:

1.1 Outlier Treatment

1.2 Missing Value Treatment

1.3 Univariate (4 marks) & Bivariate (6marks) analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building)

1.4 Train Test Split

1.5 Build Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff. Also showcase your model building approach

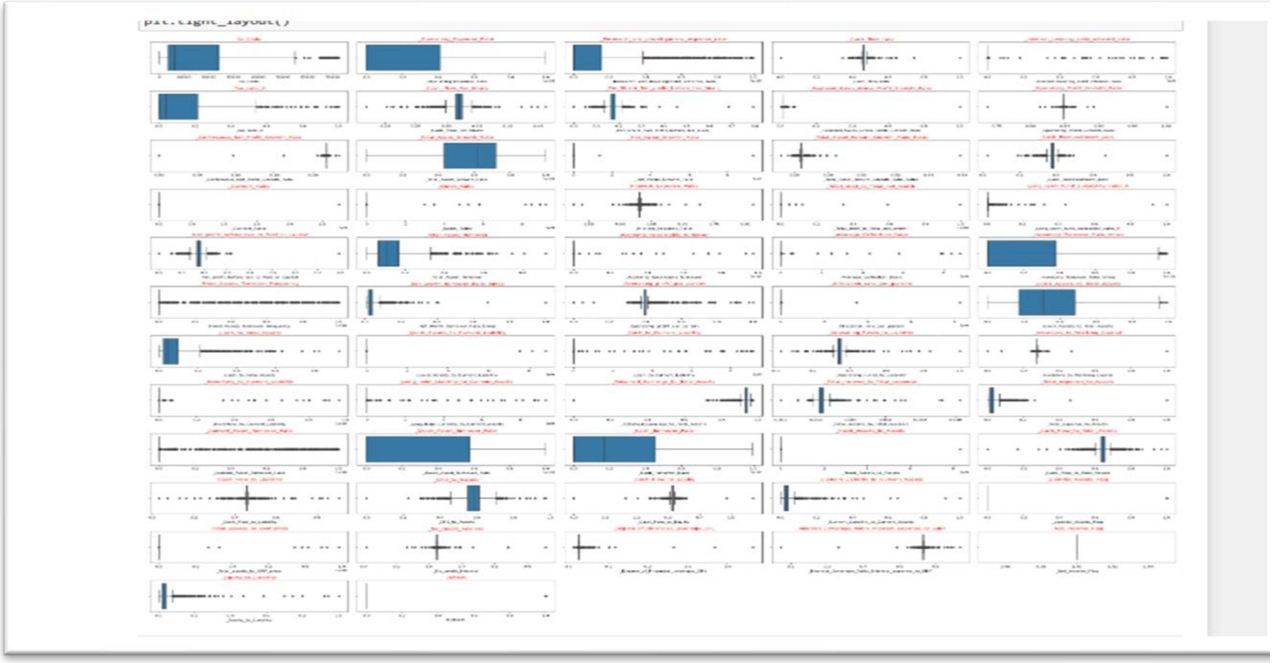
1.6 Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model

# INTRODUCTION TO PROBLEM STATEMENT

The purpose of this report is to analyze the credit risk dataset and build a logistic regression model to predict credit defaults. We will use the provided data dictionary to understand the variables and perform various data preprocessing and analysis steps. The report will include outlier treatment, missing value treatment, univariate and bivariate analysis, model building, and model validation.

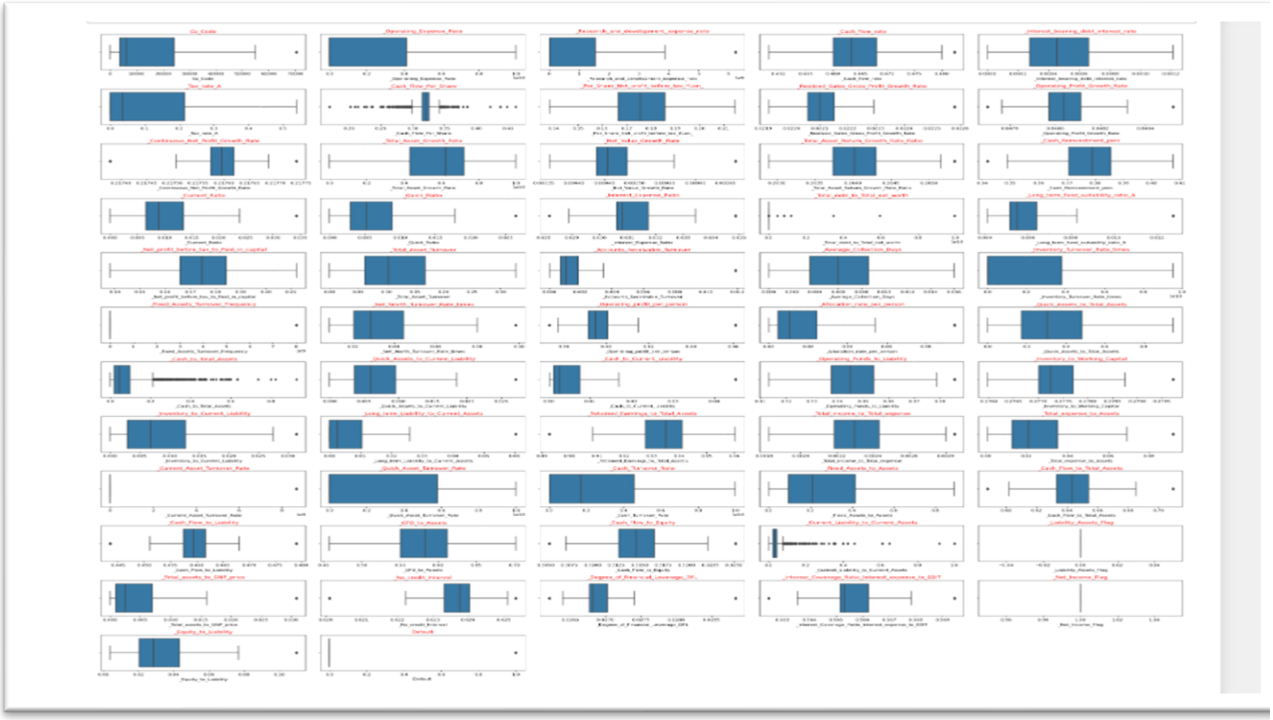
## 1.1 Outlier Treatment

Outliers can have a significant impact on model performance, so it is essential to identify and handle them appropriately. We will use appropriate statistical techniques, such as interquartile range (IQR), to identify and handle outliers in numerical variables.



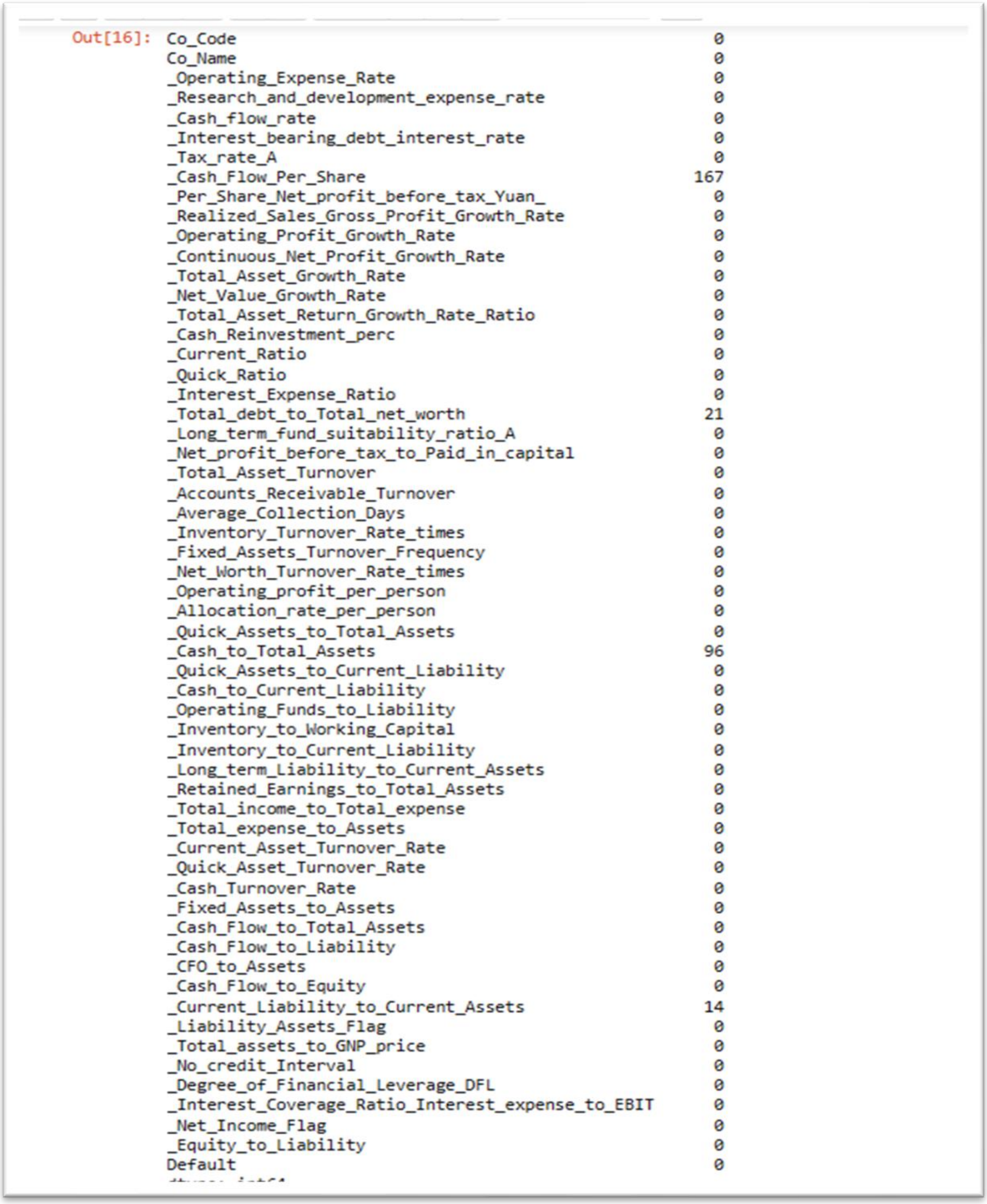
```
plt.tight_layout()
```

We can see that there are many outliers present in the dataset. So, we will treat these outliers using
interquartile range (IQR).



## 1.2 Missing Value Treatment

Missing values can also affect model performance and the reliability of the analysis. We will examine the dataset for missing values and employ suitable techniques to handle them. Depending on the extent of missingness and the nature of the variables, methods such as mean/median imputation or using predictive models can be applied.



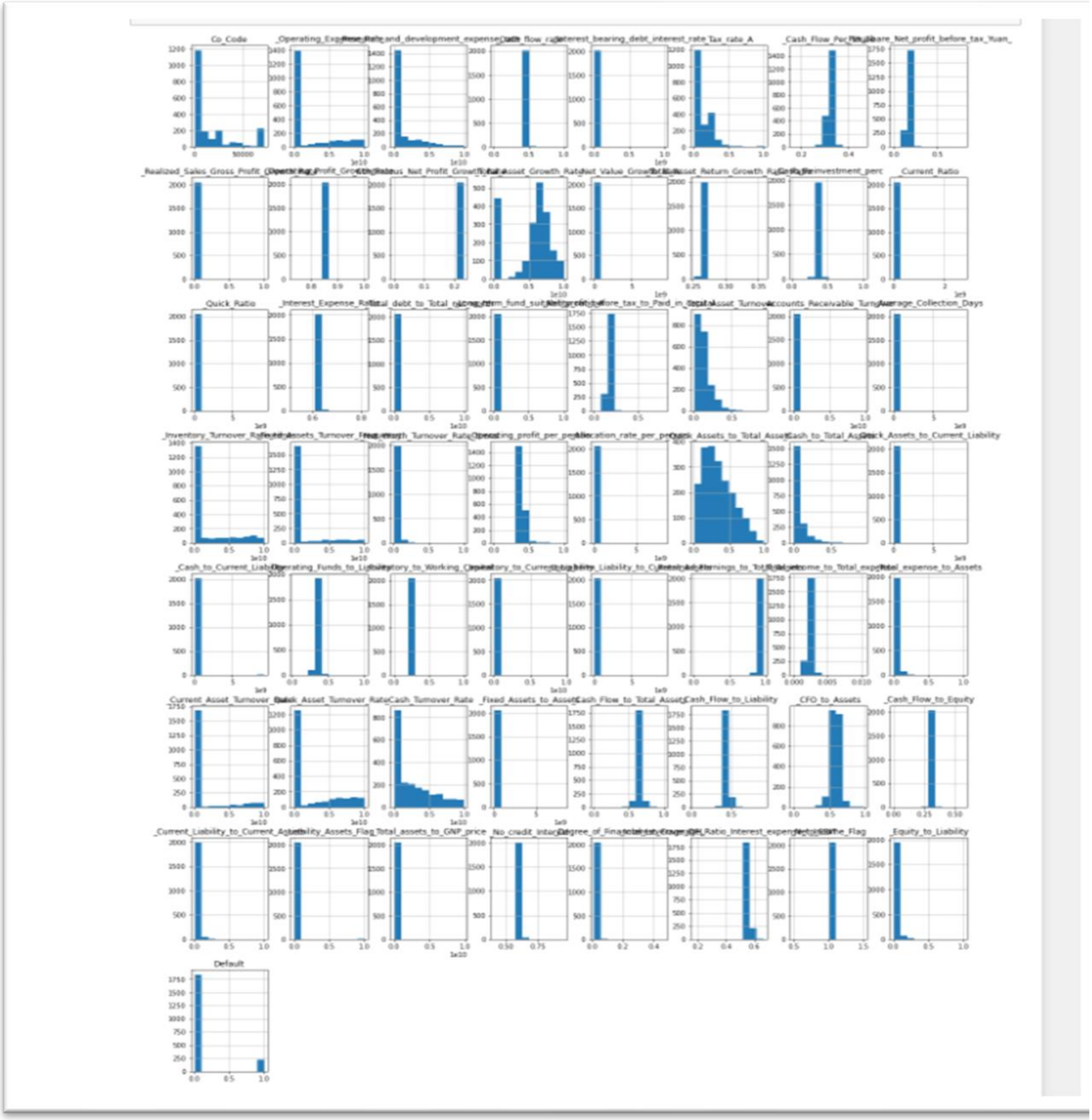
```
Out[16]: Co_Code                                          0
         Co_Name                                          0
         _Operating_Expense_Rate                          0
         _Research_and_development_expense_rate           0
         _Cash_flow_rate                                  0
         _Interest_bearing_debt_interest_rate             0
         _Tax_rate_A                                      0
         _Cash_Flow_Per_Share                           167
         _Per_Share_Net_profit_before_tax_Yuan_           0
         _Realized_Sales_Gross_Profit_Growth_Rate         0
         _Operating_Profit_Growth_Rate                    0
         _Continuous_Net_Profit_Growth_Rate               0
         _Total_Asset_Growth_Rate                         0
         _Net_Value_Growth_Rate                           0
         _Total_Asset_Return_Growth_Rate_Ratio            0
         _Cash_Reinvestment_perc                          0
         _Current_Ratio                                   0
         _Quick_Ratio                                     0
         _Interest_Expense_Ratio                          0
         _Total_debt_to_Total_net_worth                  21
         _Long_term_fund_suitability_ratio_A              0
         _Net_profit_before_tax_to_Paid_in_capital        0
         _Total_Asset_Turnover                            0
         _Accounts_Receivable_Turnover                    0
         _Average_Collection_Days                         0
         _Inventory_Turnover_Rate_times                   0
         _Fixed_Assets_Turnover_Frequency                 0
         _Net_Worth_Turnover_Rate_times                   0
         _Operating_profit_per_person                     0
         _Allocation_rate_per_person                      0
         _Quick_Assets_to_Total_Assets                    0
         _Cash_to_Total_Assets                           96
         _Quick_Assets_to_Current_Liability               0
         _Cash_to_Current_Liability                       0
         _Operating_Funds_to_Liability                    0
         _Inventory_to_Working_Capital                    0
         _Inventory_to_Current_Liability                  0
         _Long_term_Liability_to_Current_Assets           0
         _Retained_Earnings_to_Total_Assets               0
         _Total_income_to_Total_expense                   0
         _Total_expense_to_Assets                         0
         _Current_Asset_Turnover_Rate                     0
         _Quick_Asset_Turnover_Rate                       0
         _Cash_Turnover_Rate                              0
         _Fixed_Assets_to_Assets                          0
         _Cash_Flow_to_Total_Assets                       0
         _Cash_Flow_to_Liability                          0
         _CFO_to_Assets                                   0
         _Cash_Flow_to_Equity                             0
         _Current_Liability_to_Current_Assets            14
         _Liability_Assets_Flag                           0
         _Total_assets_to_GNP_price                       0
         _No_credit_Interval                              0
         _Degree_of_Financial_Leverage_DFL                0
         _Interest_Coverage_Ratio_Interest_expense_to_EBIT 0
         _Net_Income_Flag                                 0
         _Equity_to_Liability                             0
         Default                                          0
         dtype: int64
```

Here, we can see that there are missing values present in the dataset in variables _Cash_Flow_Per_Share, _Total_debt_to_Total_net_worth, _Cash_to_Total_Assets and _Current_Liability_to_Current_Assets. We will teat the missing values using median.

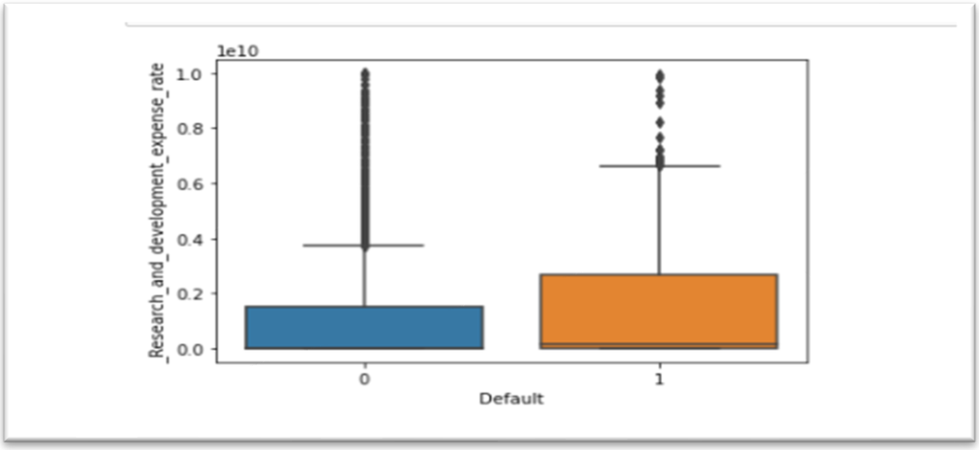
### 1.3 Univariate (4 marks) & Bivariate (6marks) analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building).

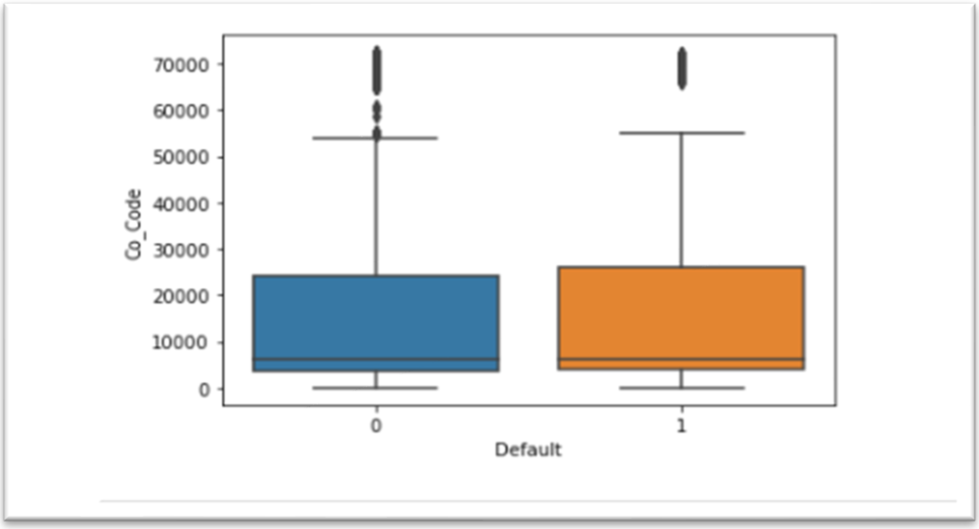
<u>Univariate Analysis:</u>

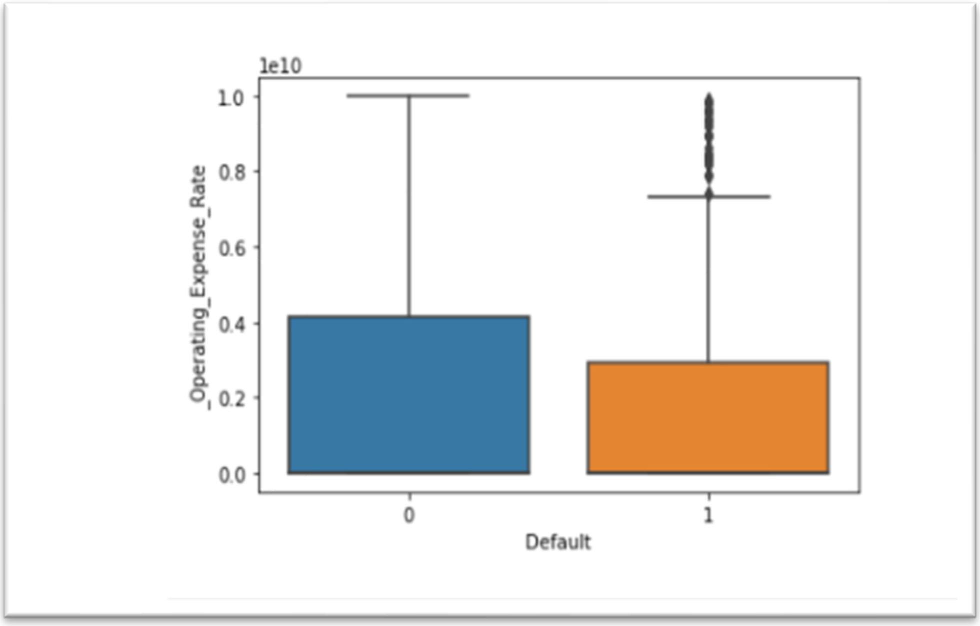
We will analyze each variable individually to understand its distribution, central tendency, and spread. We will generate appropriate visualizations such as histograms and box plots to present the findings.

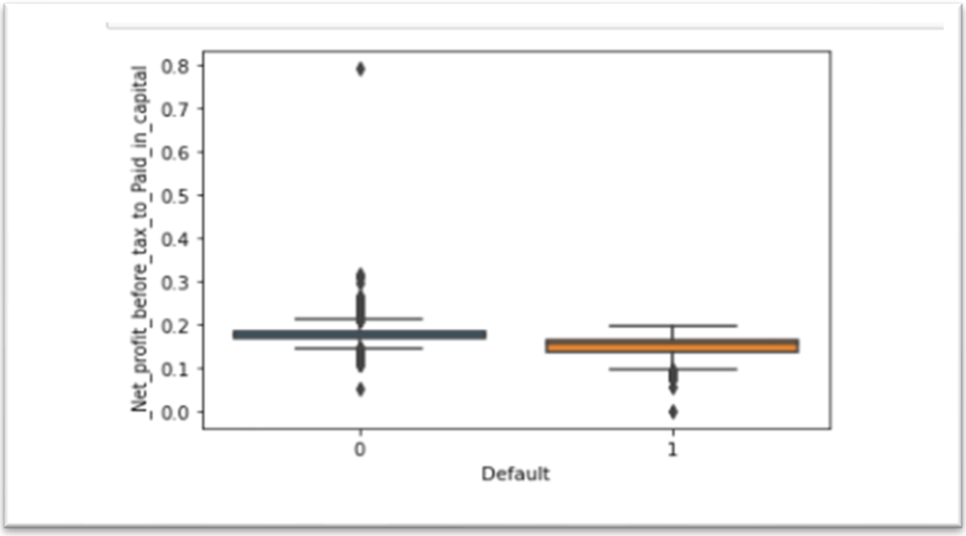

Co_Code
_Operating_Expense_Rate
_Research_and_development_expense_rate
_Cash_flow_rate
_Interest_bearing_debt_interest_rate
_Tax_rate_A
_Cash_Flow_Per_Share
_Per_Share_Net_profit_before_tax_Yuan_
_Realized_Sales_Gross_Profit_Growth_Rate
_Operating_Profit_Growth_Rate
_Continuous_Net_Profit_Growth_Rate
_Total_Asset_Growth_Rate
_Net_Value_Growth_Rate
_Total_Asset_Return_Growth_Rate_Ratio
_Cash_Reinvestment_perc
_Current_Ratio
_Quick_Ratio
_Interest_Expense_Ratio
_Total_debt_to_Total_net_worth
_Long_term_fund_suitability_ratio_A
_Net_profit_before_tax_to_Paid_in_capital
_Total_Asset_Turnover
_Accounts_Receivable_Turnover
_Average_Collection_Days
_Inventory_Turnover_Rate_times
_Fixed_Assets_Turnover_Frequency
_Net_Worth_Turnover_Rate
_Operating_profit_per_person
_Allocation_rate_per_person
_Quick_Assets_to_Total_Assets
_Cash_to_Total_Assets
_Quick_Assets_to_Current_Liability
_Cash_to_Current_Liability
_Operating_Funds_to_Liability
_Inventory_to_Working_Capital
_Inventory_to_Current_Liability
_Long_term_Liability_to_Current_Assets
_Retained_Earnings_to_Total_Assets
_Total_income_to_Total_expense
_Total_expense_to_Assets
_Current_Asset_Turnover_Rate
_Quick_Asset_Turnover_Rate
_Cash_Turnover_Rate
_Fixed_Assets_to_Assets
_Cash_Flow_to_Total_Assets
_Cash_Flow_to_Liability
_CFO_to_Assets
_Cash_Flow_to_Equity
_Current_Liability_to_Current_Assets
_Liability_Assets_Flag
_Total_assets_to_GNP_price
_No_credit_Interval
_Degree_of_Financial_Leverage_DFL
_Interest_Coverage_Ratio_Interest_expense_to_EBIT
_Net_Income_Flag
_Equity_to_Liability
Default

Boxplot for Default vs other variables.



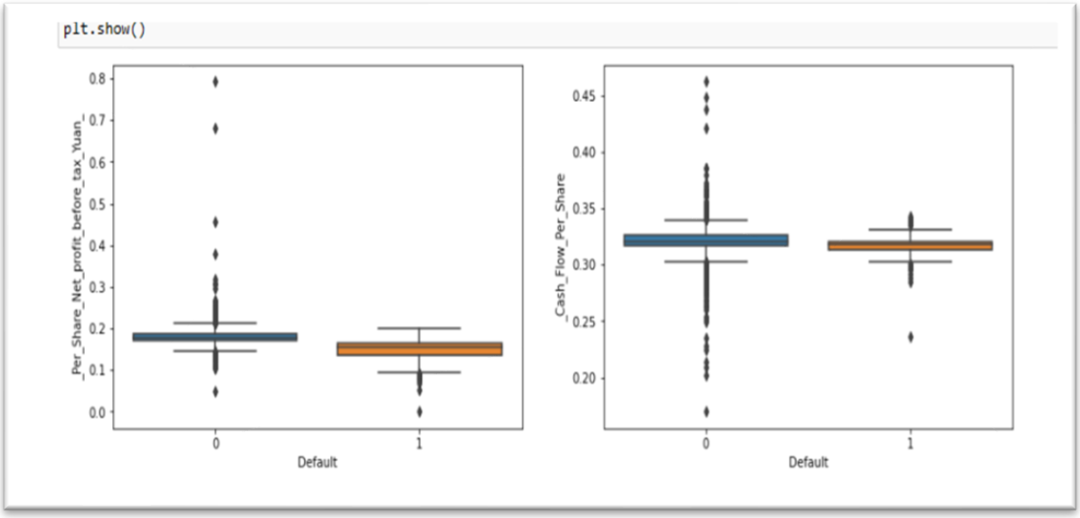




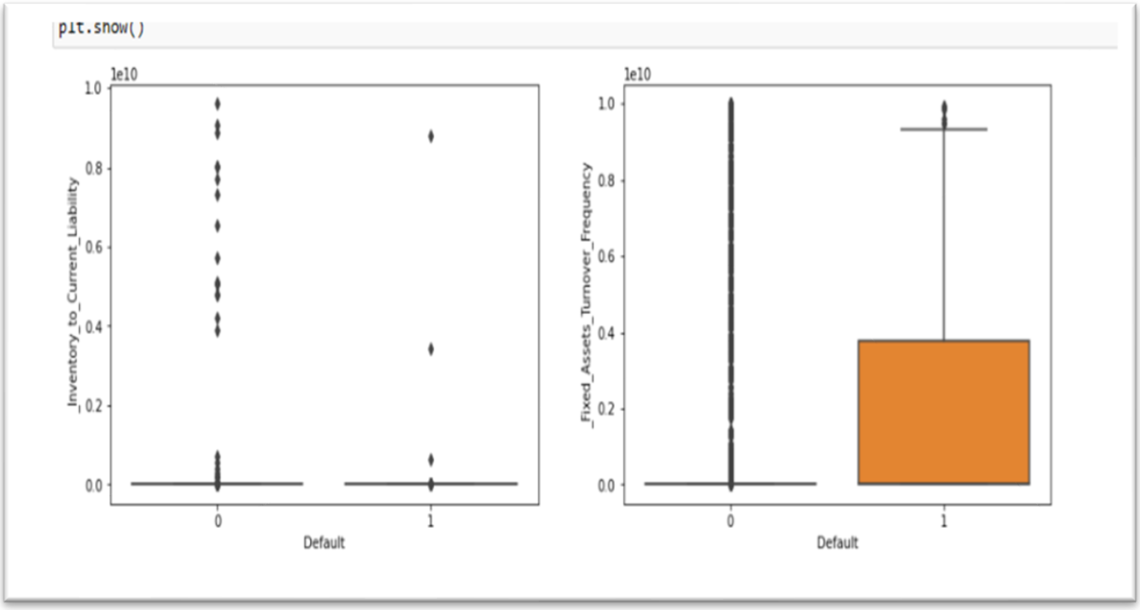


<u>Bivariate Analysis:</u>

Bivariate analysis examines the relationship between two variables. We will explore the relationships between the target variable (Default) and other significant variables identified during model building. This analysis will help us understand the influence of different factors on Defaults. We will utilize techniques such as correlation analysis to gain insights.



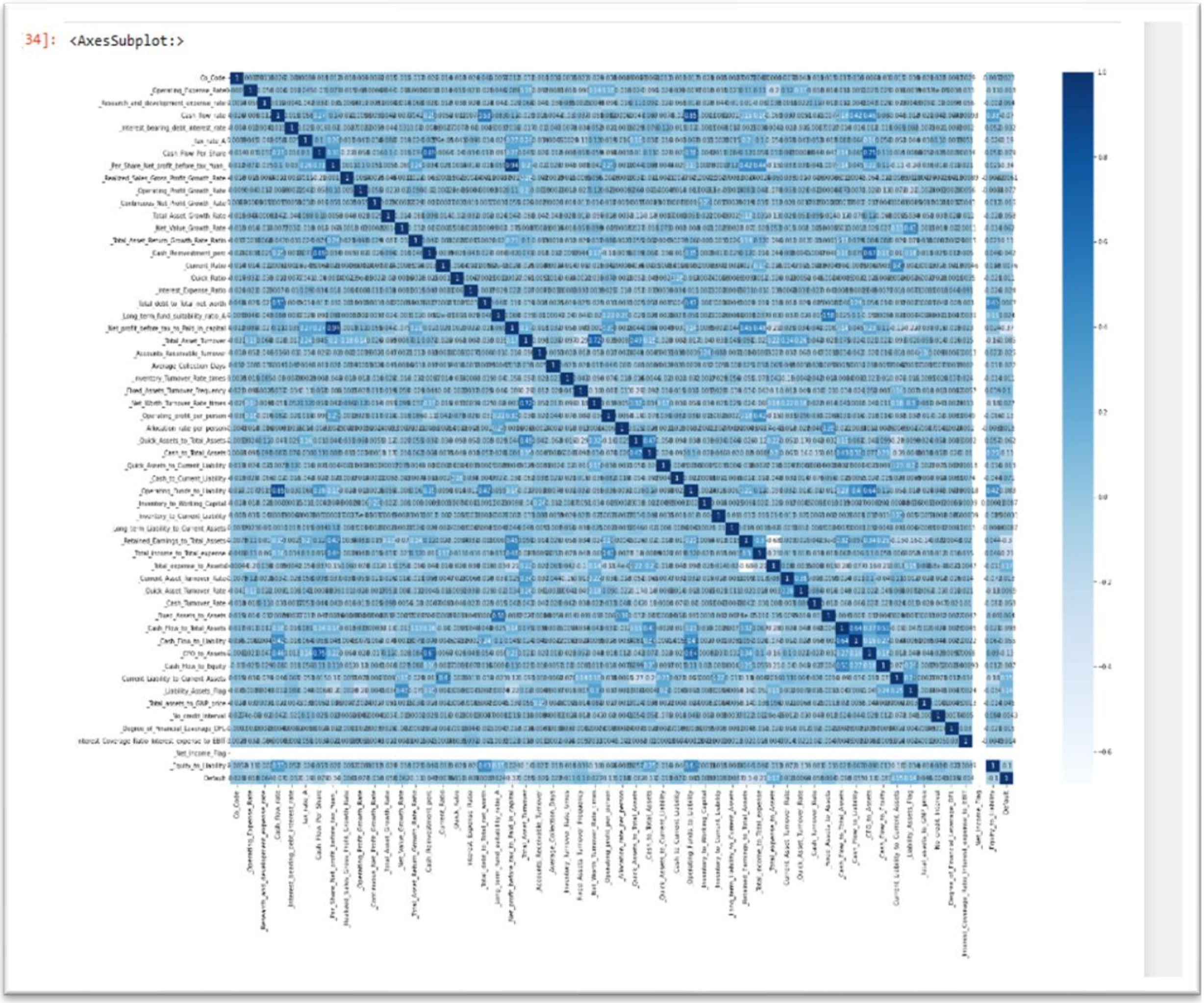
```
plt.show()
```



```
plt.show()
```

1e10
1.0
0.8
0.6
0.4
0.2
0.0
_Inventory_to_Current_Liability
0
1
Default
1e10
1.0
0.8
0.6
0.4
0.2
0.0
_Fixed_Assets_Turnover_Frequency
0
1
Default

Heatmap of the variables

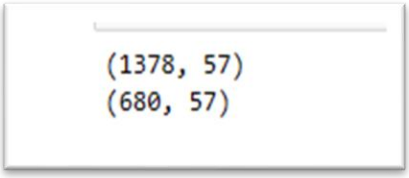


```
34]: <AxesSubplot:>
```

Co_Code
_Operating_Expense_Rate
_Research_and_development_expense_rate
_Cash_flow_rate
_Interest_bearing_debt_interest_rate
_Tax_rate_A
_Cash_Flow_Per_Share
_Per_Share_Net_profit_before_tax_Yuan_
_Realized_Sales_Gross_Profit_Growth_Rate
_Operating_Profit_Growth_Rate
_Continuous_Net_Profit_Growth_Rate
_Total_Asset_Growth_Rate
_Net_Value_Growth_Rate
_Total_Asset_Return_Growth_Rate_Ratio
_Cash_Reinvestment_perc
_Current_Ratio
_Quick_Ratio
_Interest_Expense_Ratio
_Total_debt_to_Total_net_worth
_Long_term_fund_suitability_ratio_A
_Net_profit_before_tax_to_Paid_in_capital
_Total_Asset_Turnover
_Accounts_Receivable_Turnover
_Average_Collection_Days
_Inventory_Turnover_Rate_times
_Fixed_Assets_Turnover_Frequency
_Net_Worth_Turnover_Rate_times
_Operating_profit_per_person
_Allocation_rate_per_person
_Quick_Assets_to_Total_Assets
_Cash_to_Total_Assets
_Quick_Assets_to_Current_Liability
_Cash_to_Current_Liability
_Operating_Funds_to_Liability
_Inventory_to_Working_Capital
_Inventory_to_Current_Liability
_Long_term_Liability_to_Current_Assets
_Retained_Earnings_to_Total_Assets
_Total_income_to_Total_expense
_Total_expense_to_Assets
_Current_Asset_Turnover_Rate
_Quick_Asset_Turnover_Rate
_Cash_Turnover_Rate
_Fixed_Assets_to_Assets
_Cash_Flow_to_Total_Assets
_Cash_Flow_to_Liability
_CFO_to_Assets
_Cash_Flow_to_Equity
_Current_Liability_to_Current_Assets
_Liability_Assets_Flag
_Total_assets_to_GNP_price
_No_credit_Interval
_Degree_of_Financial_Leverage_DFL
_Interest_Coverage_Ratio_Interest_expense_to_EBIT
_Net_Income_Flag
_Equity_to_Liability
Default
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6

### 1.4 Train Test Split

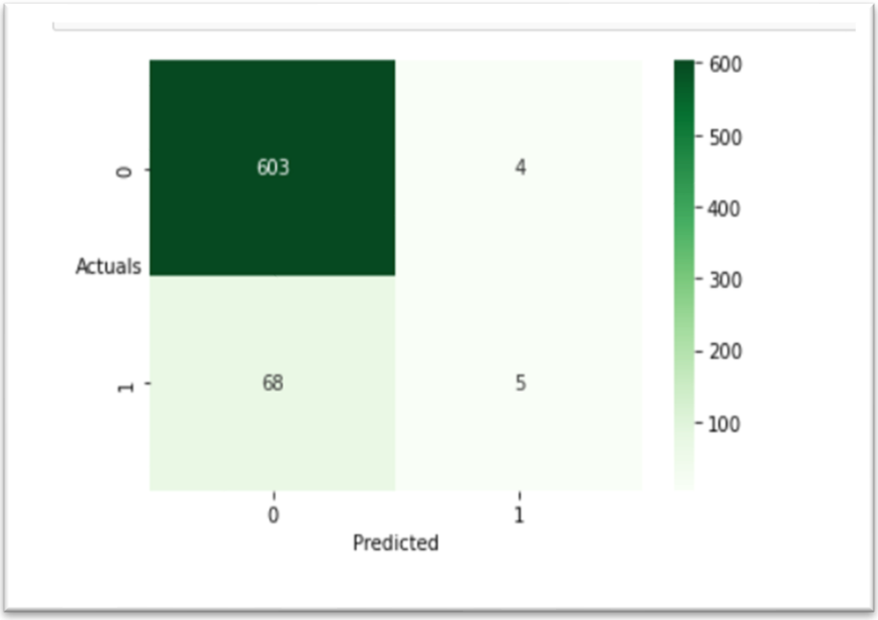
To evaluate the model's performance, we need to split the dataset into a training set and a test set. The recommended ratio for the split is 67:33, and we will use a random_state of 42 to ensure reproducibility.



```
(1378, 57)
(680, 57)
```

### 1.5 Build Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff. Also showcase your model building approach.

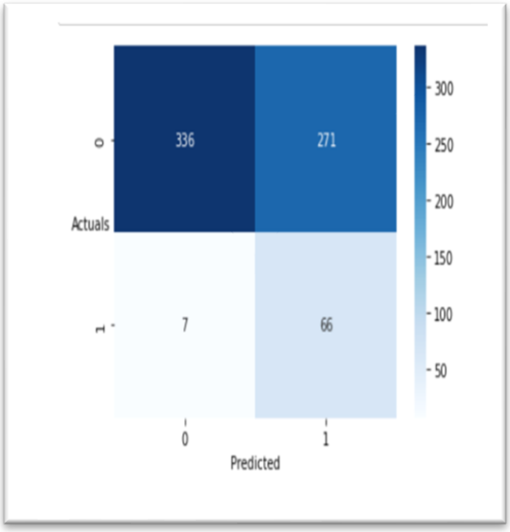
Logistic regression is a suitable model for predicting binary outcomes like credit defaults. We will build a logistic regression model on the training dataset using the statsmodels library. The model will be constructed using the most important variables identified during the analysis. Optimum Cutoff to determine the optimal cutoff probability for classifying credit defaults, we will evaluate the model's performance metrics such as accuracy, precision. Based on these metrics, we will select the cutoff that maximizes the model's predictive performance.

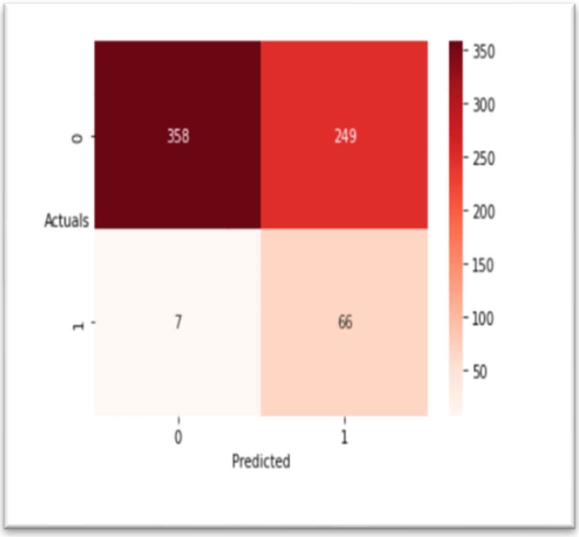


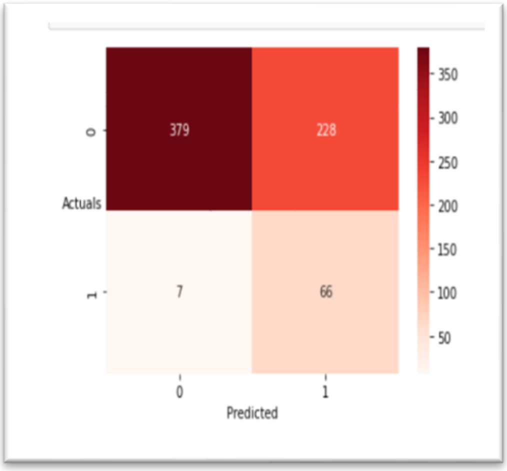
The optimum threshold is 0.1912

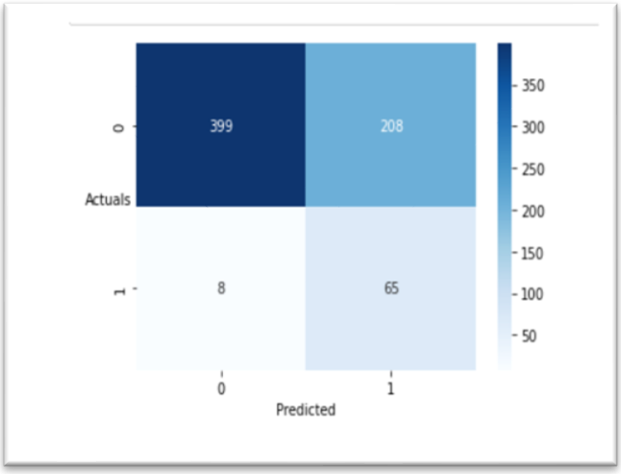
### 1.6 Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model

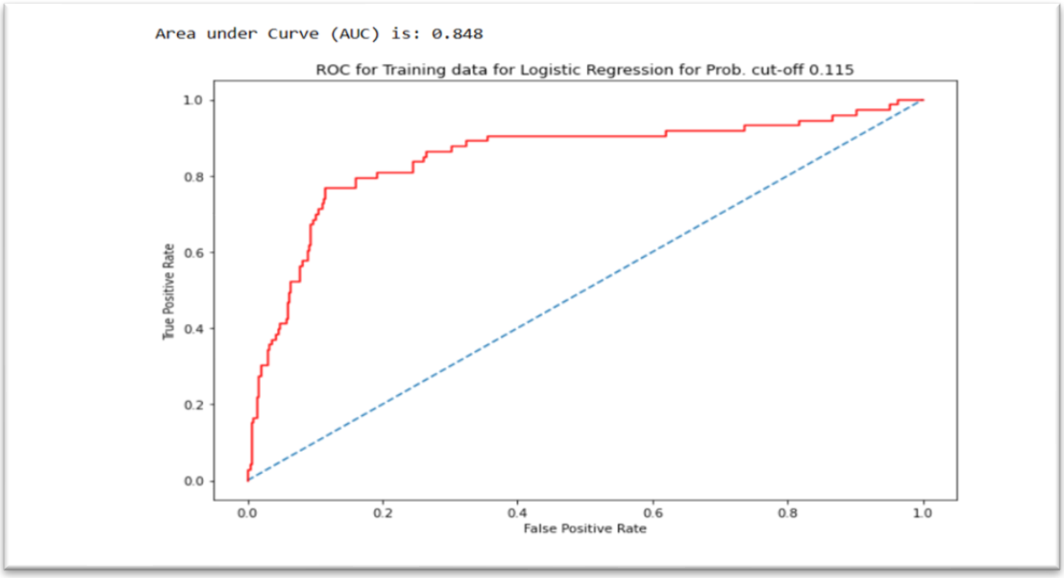
Once the model is built and the cutoff is chosen, we will validate the model using the test dataset. We will calculate the performance metrics mentioned earlier and interpret the results. These metrics will help us assess the model's ability to predict credit defaults accurately.







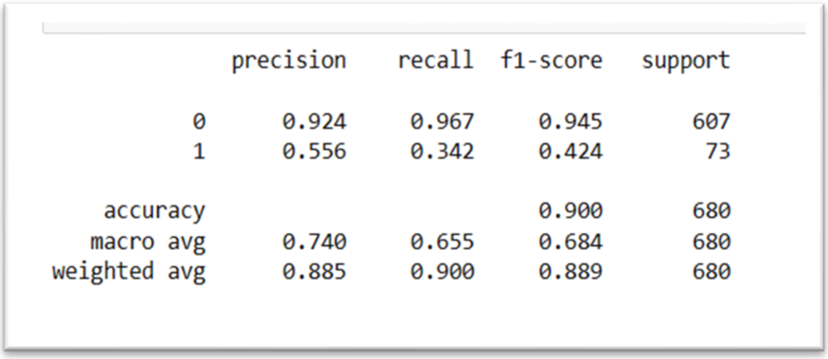




```
Area under Curve (AUC) is: 0.848
```

In the context of credit risk analysis, where the goal is to predict default or non-default status, an AUC of 0.848 suggests that the model has reasonably good predictive capability. The higher the AUC, the better the model's ability to distinguish between positive and negative instances. An AUC of 1.0 represents a perfect classifier, while an AUC of 0.5 indicates a random or non-discriminatory classifier.

Therefore, an AUC of 0.848 indicates that the model has a good level of discriminatory power in separating default and non-default instances, although further interpretation or analysis may be necessary to determine the model's effectiveness in practical applications.



```
              precision    recall  f1-score   support

           0      0.924     0.967     0.945       607
           1      0.556     0.342     0.424        73

    accuracy                          0.900       680
   macro avg      0.740     0.655     0.684       680
weighted avg      0.885     0.900     0.889       680
```

- Precision: Precision measures the proportion of correctly predicted positive instances (in this case, default instances) out of all instances predicted as positive. A precision of 0.924 for class 0 (non-default) indicates that 92.4% of the instances predicted as non-default were correctly classified. For class 1 (default), the precision is 0.556, indicating that 55.6% of the instances predicted as default were correctly classified.
- Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive instances (default instances) that were correctly predicted. A recall of 0.967 for class 0 indicates that 96.7% of the actual non-default instances were correctly classified. For class 1, the recall is 0.342, indicating that only 34.2% of the actual default instances were correctly classified.

- F1-score: The F1-score is the harmonic mean of precision and recall. It provides a balanced measure of the model's accuracy. An F1-score of 0.945 for class 0 indicates a good balance between precision and recall for non-default instances. For class 1, the F1-score is 0.424, indicating a lower balance between precision and recall for default instances.
- Support: Support represents the number of instances in each class.
- Accuracy: The overall accuracy of the model is 0.900, which indicates that 90.0% of the instances were correctly classified.
- Macro Average: The macro average calculates the average performance across both classes, giving equal weight to each class. The macro average precision is 0.740, recall is 0.655, and F1-score is 0.684.
- Weighted Average: The weighted average calculates the average performance across both classes, considering the number of instances in each class as weights. The weighted average precision is 0.885, recall is 0.900, and F1-score is 0.889.

In summary, the model performs relatively well in predicting non-default instances (class 0), with high precision, recall, and F1-score. However, it has a lower performance in predicting default instances (class 1), with lower precision, recall, and F1-score.

## CONCLUSION

This report aims to provide a comprehensive analysis of the credit risk dataset, including outlier treatment, missing value treatment, univariate and bivariate analysis, model building, and model validation. By following these steps, we can gain insights into the factors influencing credit defaults and develop a reliable logistic regression model for predicting future defaults. The findings and interpretations from this analysis will assist investors and businesses in making informed decisions regarding credit risk management.