Murali B

FRA PROJECT MILESTONE-1

**Contents Page numbers**

**Table**

**1.1 Head of the data 3  
1.2 Tail of the data 4  
1.4 Describe the data 4  
  
Output.**

**1.3 Shape of the data 4  
1.5 Column names details 5  
1.6 Renaming column names 6  
1.7 Duplicate values 6  
1.8 Datatypes details 7  
1.11 Missing values 10  
1.12 After fixing missing values 11   
1.40 Train test split data 27  
1.41 Logistic Regression 27  
1.42 Logit Regression results 28  
1.43 Prediction on Train set 29  
1.44 Prediction on train set 29  
1.45Threshold value 29  
1.54 Classification report 35  
1.56 Classification report 36**

**Figure:**

**1.9 Checking for outliers 8  
1.10 After fixing outliers 9  
1.13 Univariate Analysis 12  
Bivariate Analysis 1.14 to 1.37 13-24  
1.38 Heatmap 25  
1.39 Multivariate Analysis 26  
1.46 Probability cut off 0.2 30  
1.47 Probability cut off 0.2 31  
1.48 Probability cut off 0.2 31  
1.49 Probability cut off 0.2 32  
1.50 Probability cut off 0.2 32  
1.51 Probability cut off 0.2 33  
1.52 AUC value and RUC curve 34  
1.53 Probability cut off 0.2 34  
1.55 Probability cut off 0.2 35  
1.56 AUC value and RUC curve 36**

**Problem Statement**

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 can handle 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'

**Hints**:

Test Train Split -   Split the data into Train and Test dataset in a ratio of 67:33 and use random state =42. Model Building is to be done on Train Dataset and Model Validation is to be done on Test Dataset.

**EDA:**   
  
**Head of the data:**

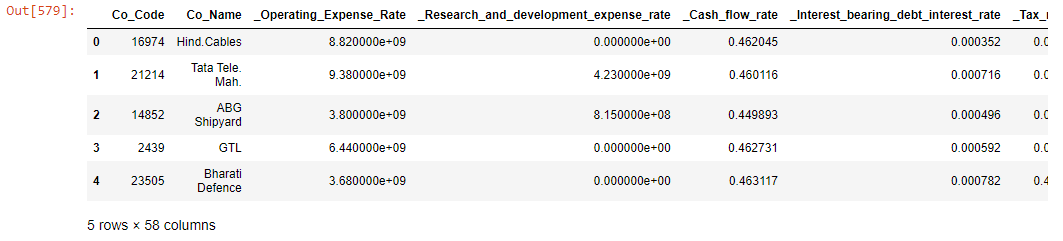


Table 1.1   
  
Top 5 rows of the data

**Tail of the data:**

A screenshot of a computer

Description automatically generated with low confidence

Table 1.2   
  
Last 5 rows of the data.

**Shape of the data:**



Output 1.3   
  
There are 2058 entries with 58 observations.

**Describe the data:**

A screenshot of a computer

Description automatically generated with low confidence

Table 1.4   
  
It describes the mean, standard deviation, minimum value, 25% 50% 75% , count and max values of the dataset.

**Column names:**

A screenshot of a computer screen

Description automatically generated with low confidence

Output 1.5   
  
Column names details

**Renaming the column names:**   
  
As we can see that the column names have ‘\_’ so we have removed the ‘\_’ symbol from column name to set a meaningful column name.   
  
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.6   
  
Renaming column names.

**Checking for duplicate values:**



Output 1.7   
  
We can see no duplicate values present in the dataset.

**Datatypes details:**

A screenshot of a computer program

Description automatically generated with medium confidence

Output 1.8   
  
We can see 53 columns are float type and 4 columns are integer and 1 column is of object type.

1.1 **Outlier Treatment**

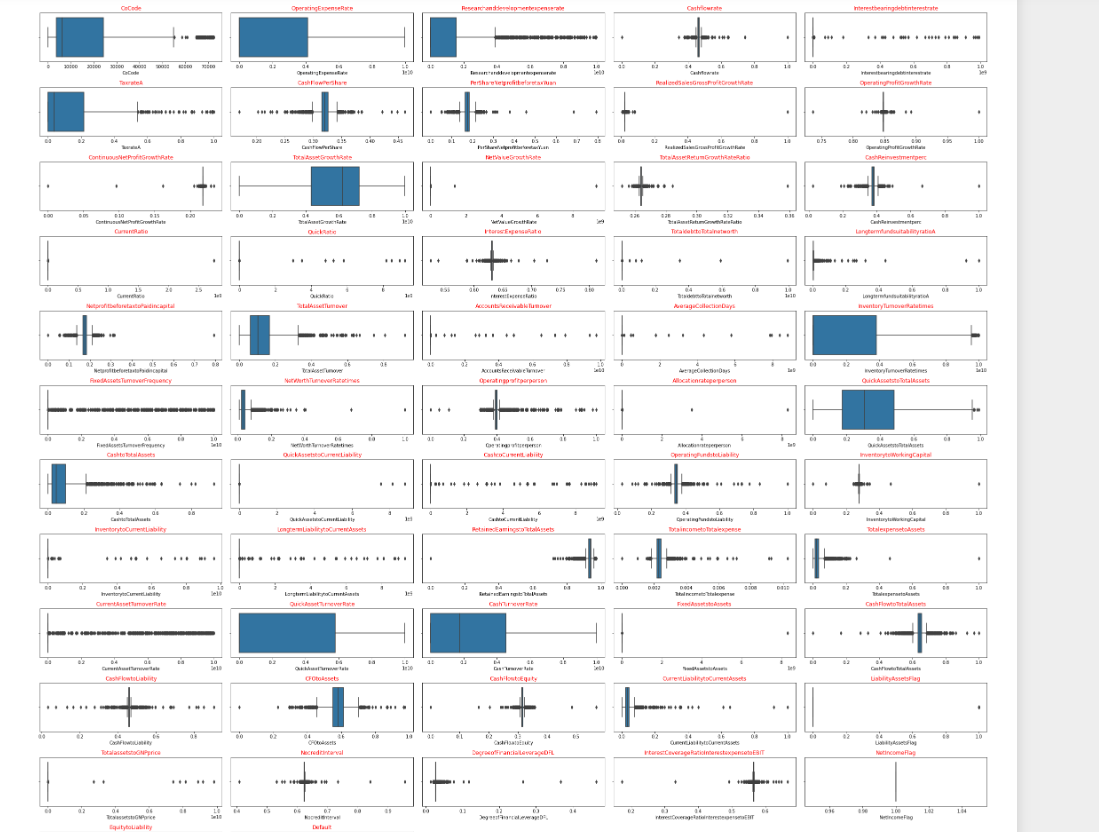


Fig 1.9   
  
We can see many outliers present in the dataset. Let’s fix it.

**After fixing outliers:**

A picture containing text, screenshot, crossword puzzle, square

Description automatically generated

Fig 1.10   
  
Now we can see outliers are fixed and no more outliers are present in the dataset.

1.2 **Missing Value Treatment**

A screenshot of a computer screen

Description automatically generated with low confidence  
  
Output 1.11

We can see 167 missing values in CashFlowperShare , 21 in TotaldebttoTotalnetworth , 96 in QuickAssetstoCurrentLiability and 14 in CurrentLiabilitytoCurrentAssets.

**After fixing Missing values:**

We have imputed the missing values with Median.



Output 1.12

Now we can see there is no missing values present in the dataset.

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)

**Univariate Analysis:**

A picture containing diagram, text, line, plot

Description automatically generated

A picture containing diagram, line, plot, plan

Description automatically generatedA screenshot of a graph

Description automatically generated with low confidence

Fig 1.13   
  
We can see that in the given dataset, many of them have skewed distribution to the left side and only few have right skewness.

**Bi-variate Analysis:**

A picture containing rectangle, diagram, screenshot, line

Description automatically generated  
Fig 1.14

We can see there is a slight increase in the number of defaulters with respect to CoCode.

A picture containing diagram, screenshot, line, rectangle

Description automatically generated  
Fig 1.15

Here we can see the number of non-defaulters is more than defaulters in OperatingExpenseRate.

A picture containing diagram, text, rectangle, screenshot

Description automatically generated  
Fig 1.16

The number of defaulters is more in ResearchanddevelopmentExpenserate.

A picture containing diagram, line

Description automatically generated  
Fig 1.17

The number of defaulters and non-defaulters both are same in Cashflowrate

A picture containing text, line, screenshot, rectangle

Description automatically generated  
Fig 1.18   
  
Here, also the number remains the same for both defaulter and non-defaulters in Interestbearingdebtinterestrate

A picture containing diagram, rectangle, line, screenshot

Description automatically generated  
Fig 1.19

The number of no-defaulters are more than the defaulters in TaxrateA

**Bi-variate Analysis:**

A screenshot of a graph

Description automatically generated with low confidence  
Fig 1.20   
  
We can observe that there are no outliers present in OperatingExpenseRate but the distribution of data is left skewed.

A screenshot of a graph

Description automatically generated with low confidence  
Fig 1.21  
  
We can see a greater number of Outliers present in Researchanddevelopmentexpenserate and again the data observations seem to be left skewed.

A picture containing diagram, line, plot, parallel

Description automatically generated  
Fig 1.22

Here we see outliers are present but the data but there is no uniform in the distribution of the data with respect to CashFlowrate.

A picture containing text, diagram, screenshot, line

Description automatically generated  
Fig 1.23

As the data are observed to be more left skewed and many outliers present in the TaxrateA.

A picture containing text, screenshot, diagram, line

Description automatically generated  
Fig 1.24

Unlike the Cashflowrate, we can see that again in CashFlowpershare the data distribution is not properly segregated, and more outliers are present on both the sides.

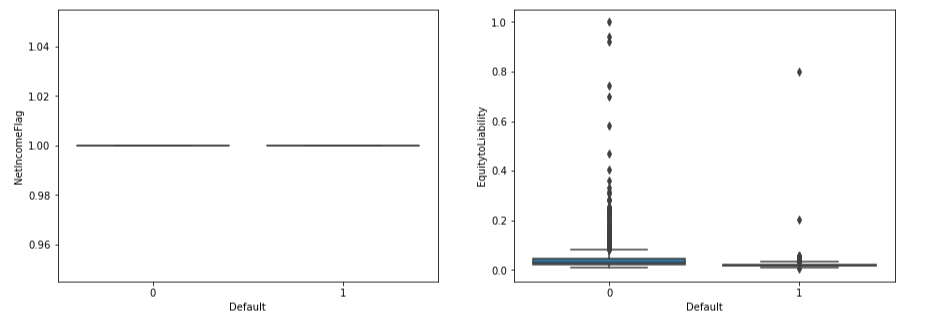
A screenshot of a graph

Description automatically generated with low confidence  
Fig 1.25

We can see that the data is left skewed and most of the outliers are present on the left side of the data in Persharenetprofitbeforetax

**Bi-variate Analysis:**

We shall do further analysis with respect to default and other variables.

  
Fig 1.26   
  
We can see non defaulters are more in EquitytoLiability when compared to NetIncomeFlag

A picture containing diagram, line, rectangle, parallel

Description automatically generated  
Fig 1.27

Here we see in LiabilityAssetsFlag, the number of defaulters shows same as NocreditInterval.

A picture containing diagram, line, technical drawing, parallel

Description automatically generated  
Fig 1.28

So, from the above plot, we see there is a slight change in the number of defaulters in CFOtoAssets when compared to CashFlowtoEquity.

A picture containing diagram, line, rectangle, plan

Description automatically generated  
Fig 1.29

Here, we see defaulters are more in CashFlowtoLiability but it’s less in CashTurnoverRate.

**Bi-variate Analysis:**

A picture containing text, screenshot

Description automatically generated  
Fig 1.30

We see that the data points are more scattered on the left side with respect to Researchanddevelopmentexpenserate and OperatingExpenseRate

A screen shot of a graph

Description automatically generated with low confidence  
Fig 1.31

There is no proper uniformity in the data distribution with regards to Cashflowrate and Interestbearingdebtinterestrate.

A picture containing text, screenshot, plot, line

Description automatically generated  
Fig 1.32

We see there in inward reflection of data from the left side to the right as the value increases with respect to TaxrateA and CashFlowPerShare.

A screen shot of a graph

Description automatically generated with low confidence  
Fig 1.33   
  
We can see there is no minimal representation of datapoints, and the points are not scattered much with regards to PerShareNetprofitbeforetaxYuan and RealizedSalesGrossProfitGrowthRate.   
A picture containing text, screenshot, line, diagram

Description automatically generated  
Fig 1.34   
  
We see there is less datapoints scattered around each other with regards to OperatingProfitGrowthRate and ContinuousNetProfitGrowthRate.   
A picture containing text, screenshot, line, rectangle

Description automatically generated  
Fig 1.35

Here, we see there is a under linear straight distribution of datapoints among TotalAssetGrowthRate and NetValueGrowthRate.

A screen shot of a graph

Description automatically generated with low confidence  
Fig 1.36

We see more datapoints are scattered onto the left side with regards to TotalAssetReturnGrowthRateRatio and CashReinvestmentperc.   
A picture containing text, screenshot, line, rectangle

Description automatically generated  
Fig 1.37

There is no proper relationship or datapoints found when tried between CurrentRatio and QuickRatio.

**Correlation Heatmap:**

A screenshot of a computer screen

Description automatically generated with low confidence

Fig 1.38

We can see that most of the variables are not properly correlated. We can see slightly negative correlation between few variables. Heatmap is mainly used to derive the correlation between two variables and many variables have correlation values as 1 which means they impose high collinearity when compared to other variables which leads to multi collinearity.

**Multi-variate Analysis:**

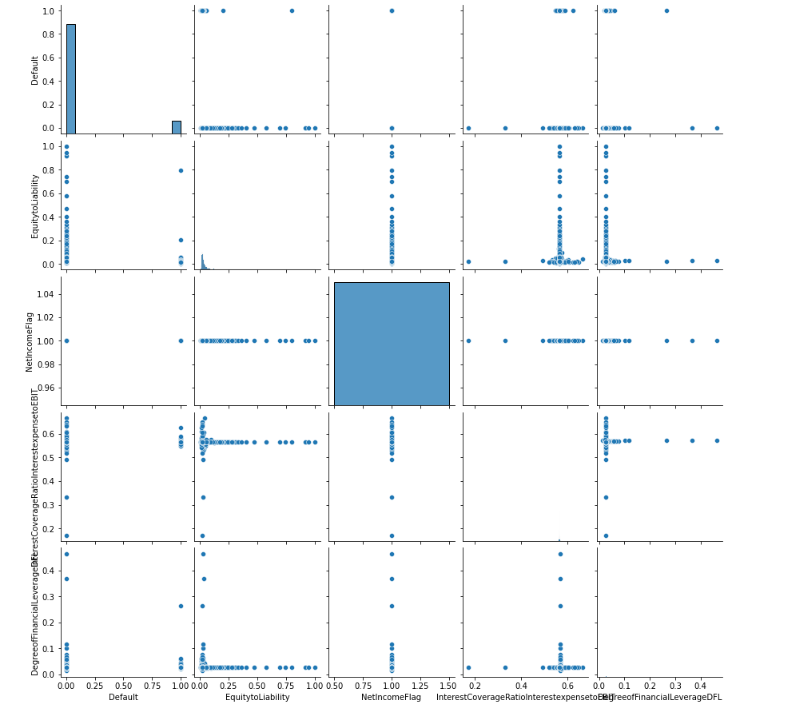


Fig 1.39

Pairplots are used for variables which have high correlation between other variables and the bivariate analysis includes pairplot and heatmap of correlation matrix. Here variables used are Default, EquitytoLiability,NetIncomeFlag,InterestCoverageRatioInterestexpensetoEBIT, DegreeofFinancialLeverageDFL for further analysis and we see relationship exists as datapoints do exists when plotted through pairplot against different variables.

1.4 **Train Test Split**

We have split the data into train and test dataset in the ratio as 67:33 and we used random\_state value as 42.

  
Output 1.40

Shape of the train and test dataset.

So, 1378 observations are for train set and 680 observations are for train set as the training set contains 67% and test data contains 33% of the train data.

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

A black text on a white background

Description automatically generated with low confidence  
Output 1.41

The Logit function of the model value shows as 0.290184 with iterations as 9.

**Logit Regression results**

A screenshot of a data

Description automatically generated with low confidence  
Output 1.42

So, we can see that that we have processed the logit regression model and number of observations includes 2058, co-efficient value of EquitytoLiability and TaxrateA and CFOtoAssets shows as negative but co-efficient values of Intercept and NetIncomeFlag and NetValueGrowthRate shows as positive values. We see that P-value of NetValueGrowthRate shows as 0.4 and remaining other variables used in this model shows as 0.

**Predicted values of the Train set:**

A screenshot of a computer

Description automatically generated with medium confidence  
Output 1.43

**Predicted values of Test set:**

A screenshot of a computer

Description automatically generated with medium confidence  
Output 1.44

**Optimal threshold value:**

  
Output 1.45

The model threshold value shows as 0.172 and is obtained by maximizing the difference between true positivity rate and false positivity rate.

We will be making use of statsmodels for logistic regression and sklearn.model\_selection import train\_test\_split to split the train and test dataset and sklearn.metrics to perform confusion matrix and to show the ROC curve and classification report. A negative coefficient will tell us that as if the independent variable increased then the dependent variable tends to decrease and in the same manner if there is a positive coefficient then if the independent variable increases so the dependent variable also increases.

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

**With probability cut off of 0.2:**

A screenshot of a graph

Description automatically generated with low confidence  
Fig 1.46

So, from the above heatmap, we can see those true negative shows as 519, false negative shows as 88 and false negative shows as 22 and true positive shows as 51.

**Probability cut off of 0.4:**

A picture containing text, screenshot, diagram, rectangle

Description automatically generated  
Fig 1.47

We can see those true negative shows as 605, false negative shows as 2 and false negative shows as 70 and true positive shows as 3.

**Probability cut off of 0.6:**

A picture containing text, screenshot, diagram, rectangle

Description automatically generated  
Fig 1.48

We can see those true negative shows as 607, false negative shows as 0 and false negative shows as 72 and true positive shows as 1.

**Probability cut off of 0.8:**

A picture containing text, screenshot, diagram, rectangle

Description automatically generated  
Fig 1.49

We can see those true negative shows as 607, false negative shows as 0 and false negative shows as 72 and true positive shows as 1.

**Probability cut off of 0.3:**

A picture containing screenshot, text, design

Description automatically generated  
Fig 1.50

We can see those true negative shows as 599, false negative shows as 8 and false negative shows as 63 and true positive shows as 10.

**Probability cut off of 0.5:**

A picture containing screenshot, text, diagram, colorfulness

Description automatically generated  
Fig 1.51

We can see those true negative shows as 607, false negative shows as 0 and false negative shows as 71 and true positive shows as 2.

**AUC value and RUC curve:** A picture containing text, line, diagram, plot

Description automatically generated  
Fig 1.52

**Probability cut off of 0.7:**

A picture containing text, screenshot, diagram, rectangle

Description automatically generated  
Fig 1.53

We can see those true negative shows as 607, false negative shows as 0 and false negative shows as 72 and true positive shows as 1.

**Classification report:**

A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.54

From the classification report, we see that 89% was made by the model with respect to correct predictions against the total prediction. We see 0.014% as defaulters which is a minimal number.

**Probability cut off of 0.9:**

A picture containing text, screenshot, diagram, rectangle

Description automatically generated  
Fig 1.55

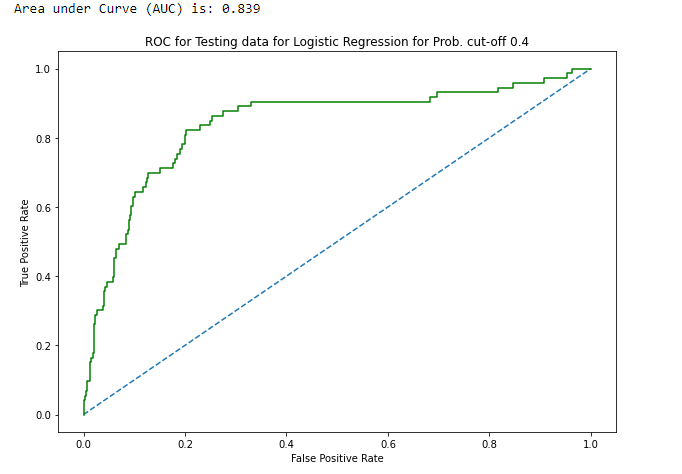
We can see those true negative shows as 607, false negative shows as 0 and false negative shows as 72 and true positive shows as 1.

**Classification report:** A picture containing text, screenshot, font, number

Description automatically generated

Output 1.56

**AUC value and RUC curve:**



Output 1.57

From the above graph, we see that the model AUC of train set shows as 0.839.

So, from the analysis, we can conclude that if there is positive coefficient then higher the value of the variable will in turn lead to higher default and negative coefficient works the other way round. Mainly credit report analysis is used to determine the credit worth of the customer. So based on the classification model we can classify the defaulters in various categories. We can improve the performance of the data model on the accuracy and recall values since there is no overfitting issues and the results derived are poor in numbers.