**Business Report on**

***FRA Project (Milestone-1)***

***Submitted to***



**Great Learning Olympus**

***Prepared by***

**Mrinmoy Majumdar**

**June-D Batch**

**Post Graduate Program in Data Science & Business Analytics**

**From**

****

UT Austin

**May, 2022**

TABLE OF CONTENT

**1.**1 Outlier Treatment………………………………………………………….6

**1.2 Missing Value Treatment**………………………………………………….7

**1.3 Transform Target variable into 0 and 1…………………………………….**8

###### **1.4 Univariate & Bivariate analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building**…9

**1.5 Train Test Split**….........................................................................................12

1.6. Build Logistic Regression Model (using stats model library) on most important variables on Train Dataset and choose the optimum cut off. Also showcase your model building approach………………………………………12

1.7. Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model ……………………………………...30

**LIST OF FIGURES**

|  |  |
| --- | --- |
| **Images** | **Page No** |
| **Information on dataset** | **5** |
| **Shape of dataset** | **5** |
| **Outlier detection boxplot** | **6** |
| **Missing value per column** | **7** |
| **Default value distribution** | **8** |
| **Histogram plot** | **9** |
| **Count plot** | **10** |
| **Bar plot** | **11** |
| **Strip plot** | **11** |
| **Correlation plot** | **11** |
| **Record count** | **12** |
| **Results of iteration** | **14** |
| **Distribution plot** | **27** |
| **Adjusted R-squared** | **27** |
| **Prediction on samples** | **28** |
| **Confusion matrix** | **28** |

**LIST OF TABLES**

|  |  |
| --- | --- |
| Table 1 | 4 |
| Table 2 | 5 |
| Table 3 | 6 |
| Table 4 | 13 |
| Table 5 | 14 |
| Table 6 | 15 |
| Table 7 | 16 |
| Table 8 | 17 |
| Table 9 | 18 |
| Table 10 | 19 |
| Table 11 | 20 |
| Table 12 | 21 |
| Table 13 | 22 |
| Table 14 | 23 |
| Table 15 | 24 |
| Table 16 | 25 |
| Table 17 | 26 |

**Problem Statement:**

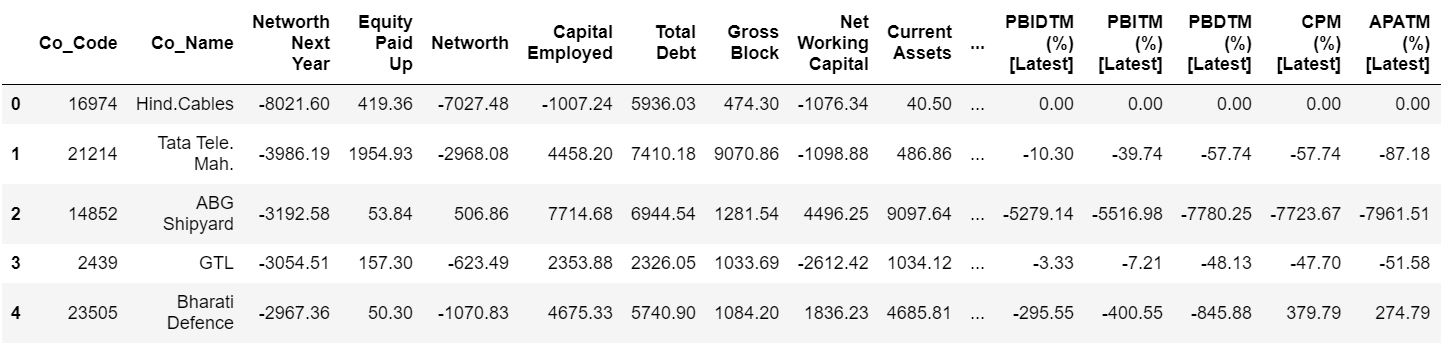
Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any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

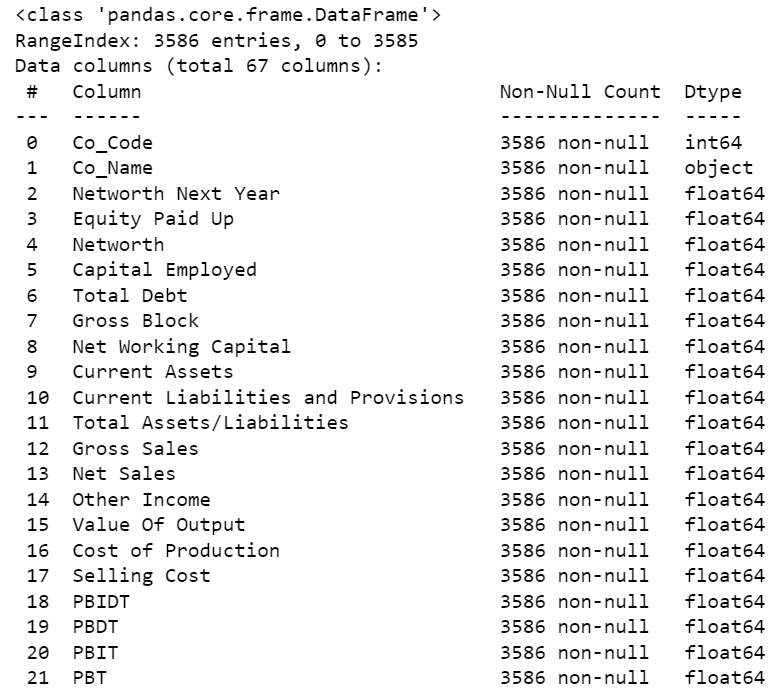
Data that is available includes information from the financial statement of the companies for the previous year (2015). Also, information about the Net worth of the company in the following year (2016) is provided which can be used to drive the labelled field.

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

Displaying company data:

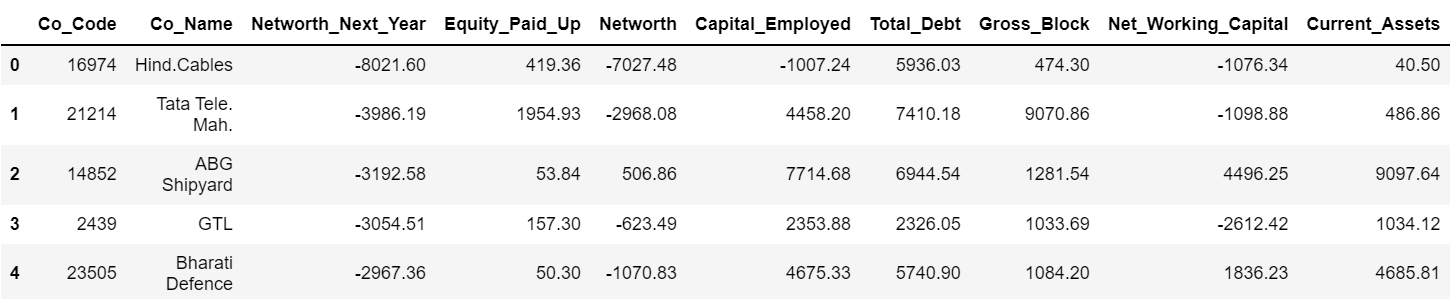


**Table 1: Top 5 rows of Company data Frame**



**Figure 1: Information on dataset**

We can observe that many column names containing spaces, commas, brackets and percentages were replaced and messy column names were fixed.

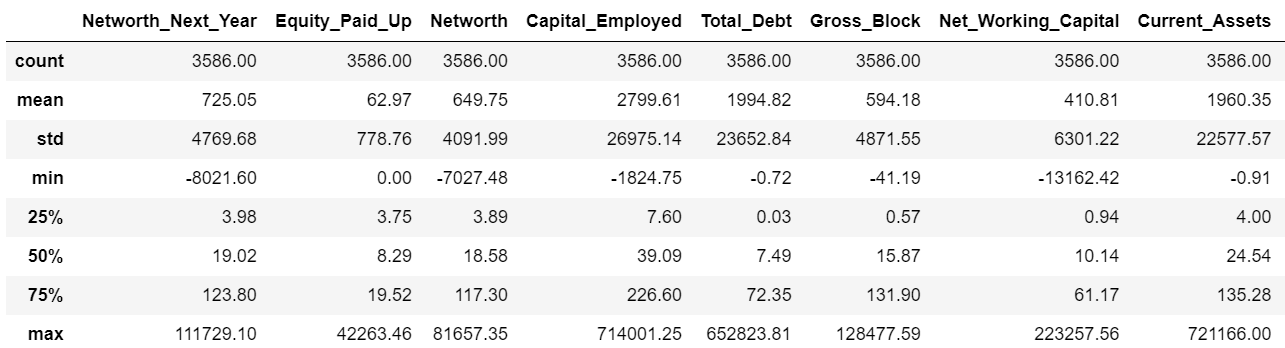


**Table 2: Dataset with renamed columns**



**Figure 2: Shape of dataset**

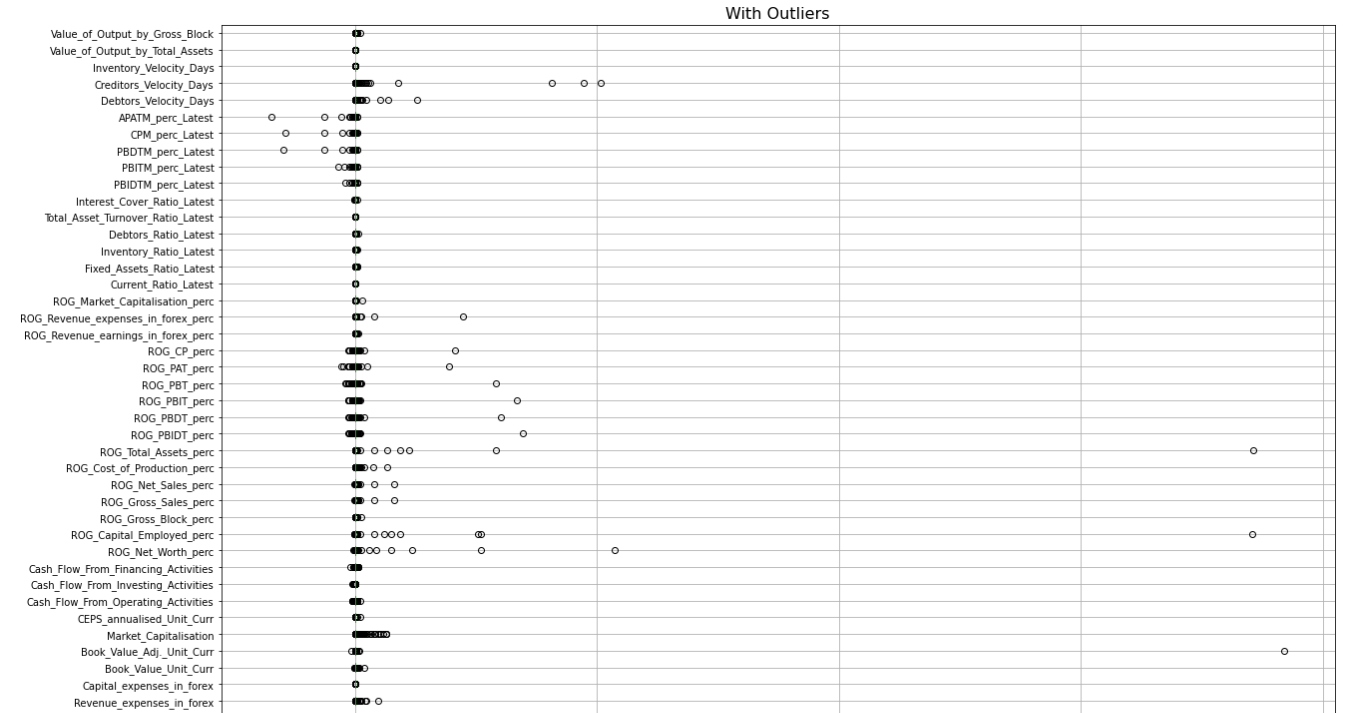
There are no duplicate records in the dataset. Datatypes of the variables are integers (3), object (1) and floating-point numbers (63). Company code and company name columns is not required for model building and therefore dropped.



**Table 3: Description of dataset**

The measure of descriptive statistics suggests that majority of the features are not normally distributed which may be due to presence of outliers and scaling will be required to handle indifferences in the data.

**Q1.1 Outlier Treatment**



**Figure 3: Outlier detection with boxplot**

From the above boxplot, it is clear that most of the variables have outliers in the dataset and need to be treated.

We are detecting which values are outside the upper limit (UL) and lower limit (LL), and we try to impute null values for the values above and below these limits.

Outliers are identified based on the inter-quartile distance as follows:

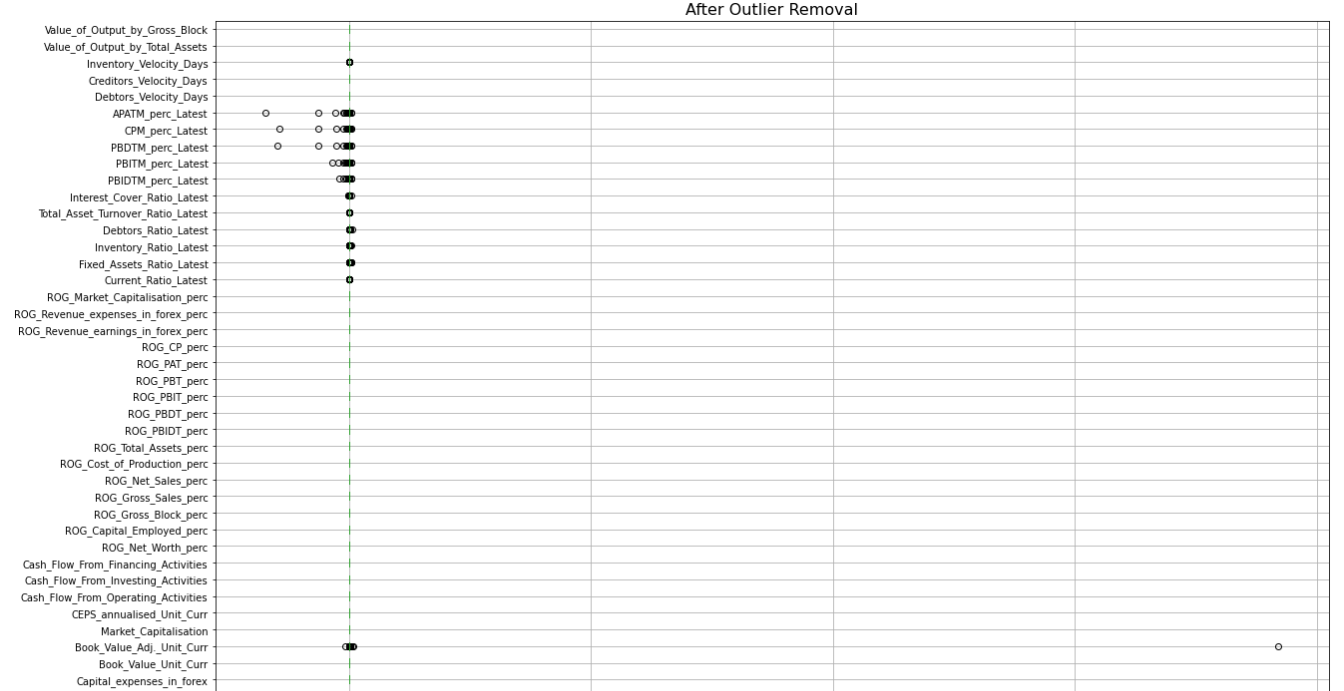
Q1 – 25th Percentile

Q3 – 75th Percentile

IQR = Q3 – Q1

Lower whisker = Value < 1.5 \* IQR

Upper whisker = Value > 1.5 \* IQR

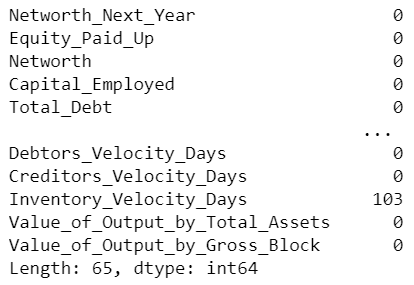


**Figure 4: Outlier treatment**

Outliers are still visible for few columns of the dataset and will be treated individually in a sequential order.

**Q1.2. Missing Value Treatment**

There are a total 118 missing values in the entire dataset.

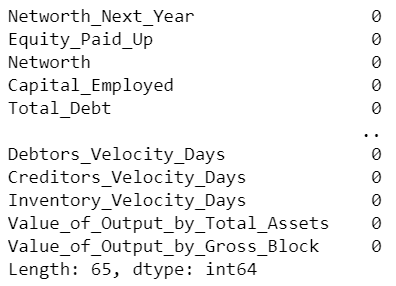


**Figure 5: Missing values per column**

### 

**Figure 6: Missing value location**

Median is the best measure of central tendency to treat missing values since the variables contain outliers.

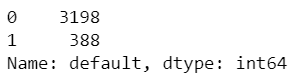


**Figure 7: Missing value check post treatment**

There are no missing values present after imputing the null values with median.

**Q1.3. Transform Target variable into 0 and 1.**

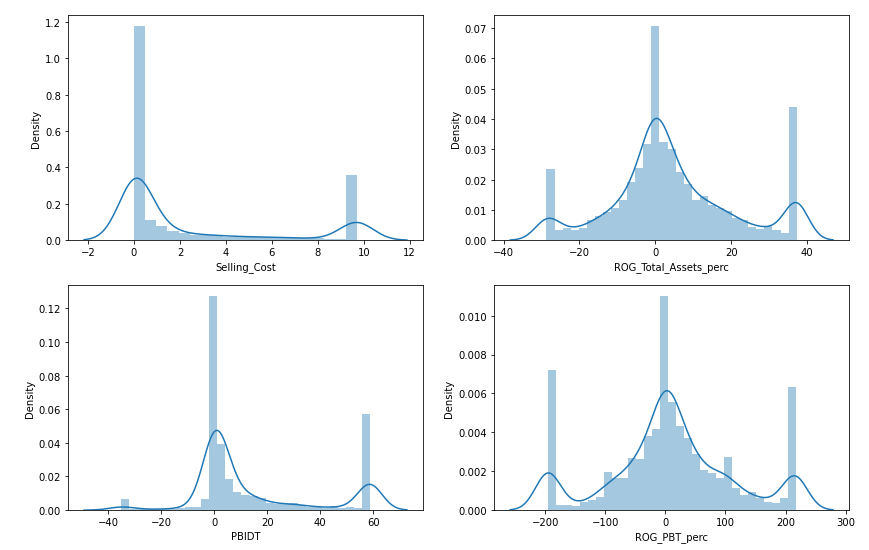
##### We need to create a default variable that should take the value of 1 when net worth next year is negative & 0 when net worth next year is positive.

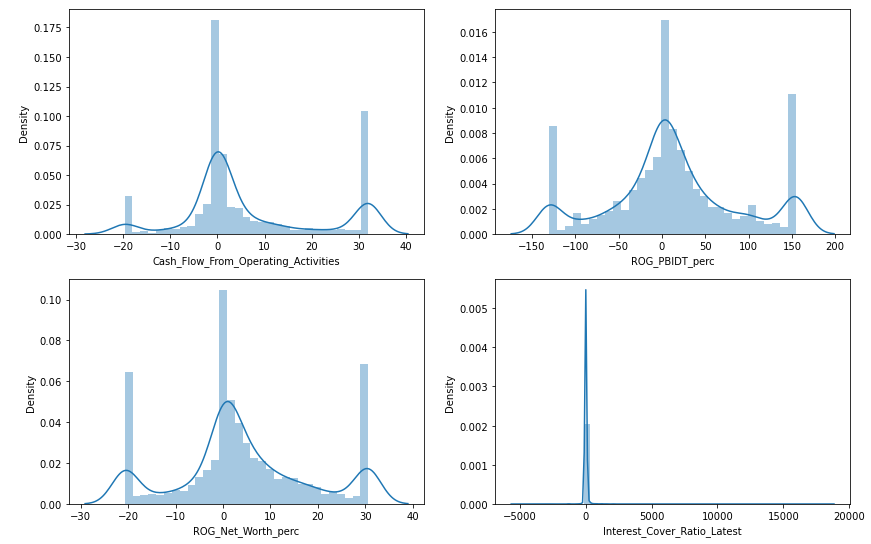


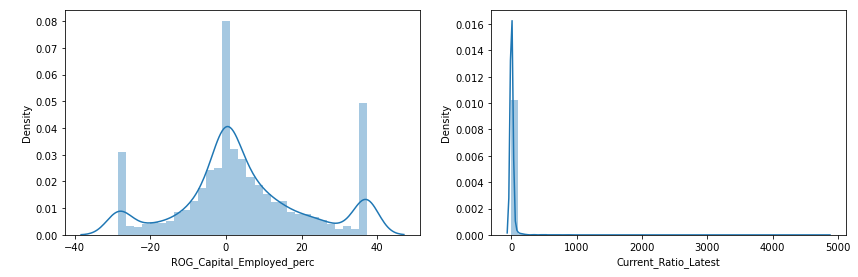
**Figure 8: Default values distribution**

For default=1 and default=0 percentage distributions are 11% and 89% respectively which shows an imbalanced dataset and requires to be fixed using techniques like SMOTE.

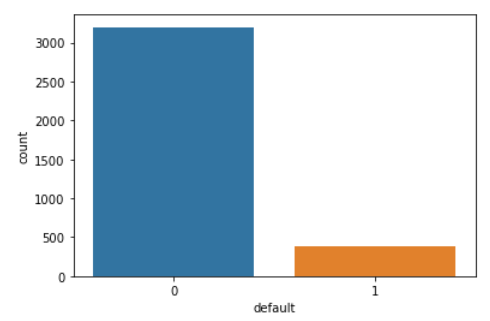
**Q1.4. Univariate & Bivariate analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building).**







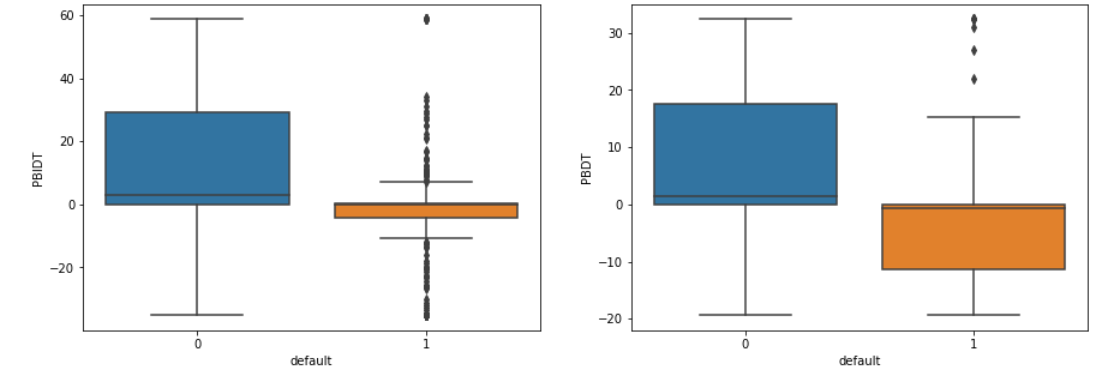
**Figure 9: Histogram plot of numerical features**



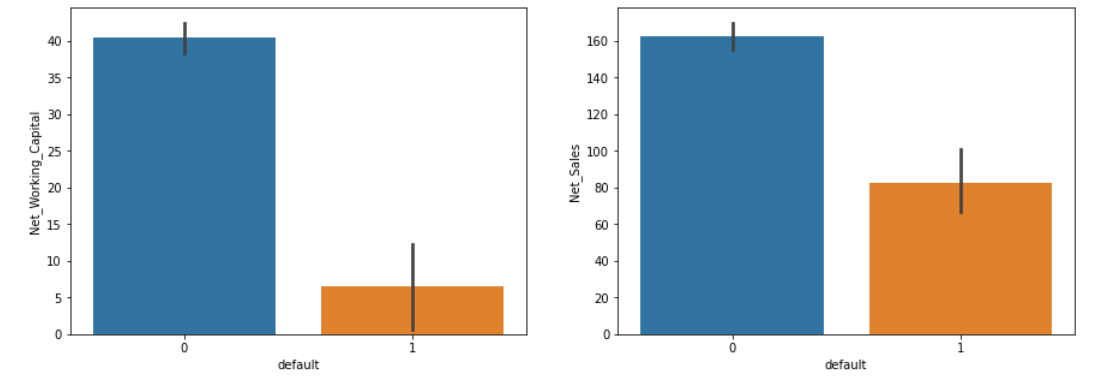
**Figure 10: Count plot of categorical column**

Inference-

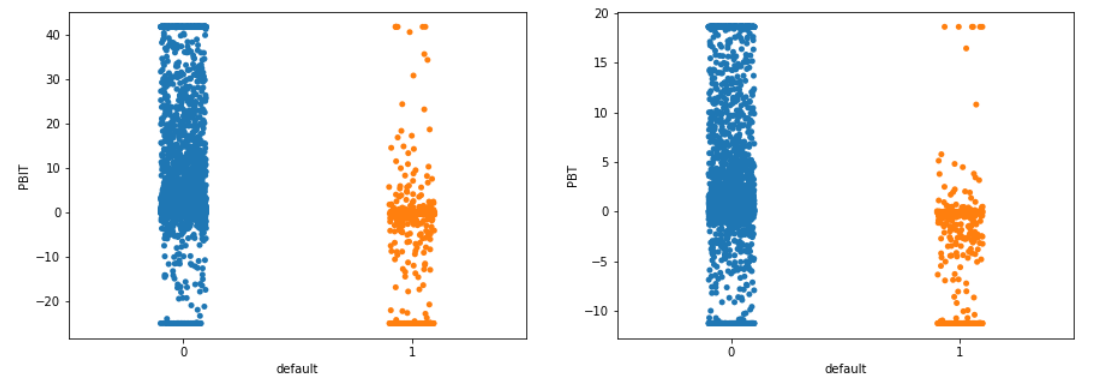
* Most of the numerical variables have multimodal distribution.
* The categorical univariate distribution indicates that there are a greater number of non-defaulters than defaulters.



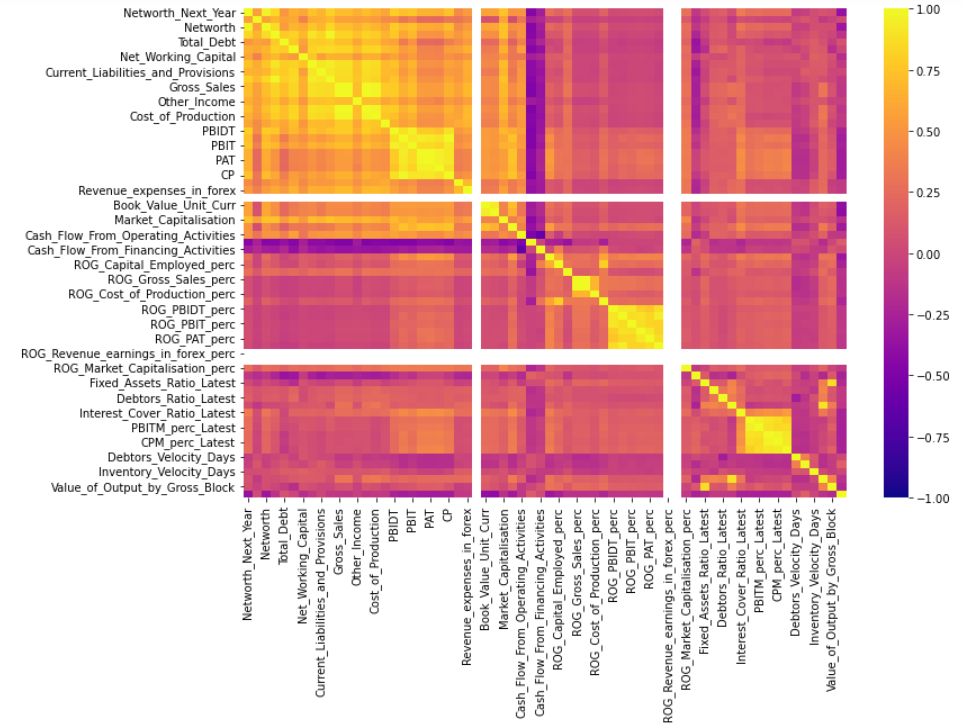
**Figure 11: Bivariate boxplot**



**Figure 12: Bivariate bar plot**



**Figure 13: Bivariate strip plot**



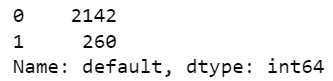
**Figure 14: Correlation plot**

Inference-

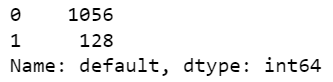
* PBIDT and PBDT have different median values for defaulters and non-defaulters.
* Sales and working capital have been higher for non-defaulters.
* Net worth is highly correlated to Total Debt and Net Working Capital.
* Variables like Net worth, Capital Employed, PBIDT, Current Assets, PBDT are highly correlated with Net worth Next Year.

**Q1.5. Train Test 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.



**Figure 15: Record count in training dataset**



**Figure 16: Record count in testing dataset**

**Q 1.6. Build Logistic Regression Model (using stats model library) on most important variables on Train Dataset and choose the optimum cut off. Also showcase your model building approach**

The equation of the Logistic Regression by which we predict the corresponding probabilities and then go on predict a discrete target variable is –

y = 1/1+𝑒−𝑧

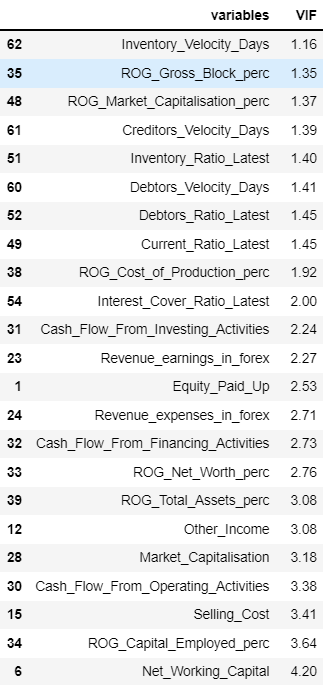
Note: z = 𝛽0 +∑𝑛𝑖=1(𝛽𝑖𝑋1)

Stats model library is imported. Splitting arrays or matrices into random train and test subsets. Model will be fitted on train set and predictions will be made on the test set.

Model Building Approach -

Model 1:

* The multicollinearity present within the dataset is addressed by calculating VIF values for each of the features and excluding those with VIF>5 from the model building exercise.
* Creating logistic regression equation & storing it. [model = SM. logit (formula=’Dependent Variable ~ Σ𝐼𝑛𝑑𝑒𝑝𝑒𝑛𝑑𝑒𝑛𝑡 𝑉𝑎𝑟𝑖𝑎𝑏𝑙𝑒𝑠 (𝑘)’ data = ‘Data Frame containing the required values’). fit ()]
* The model is built with the rest of the columns iteratively using the concept of null hypothesis that there is no relationship between dependent and independent variables. p-values <0.05 are restored for upcoming loops and remaining are excluded sequentially.
* The steps are continued till there are no variables with p-value>0.05.



**Table 4: VIF values < 5**

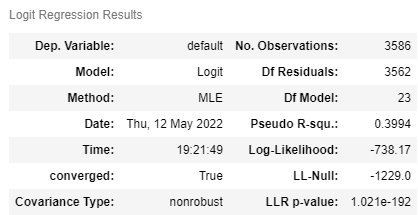
Loop 1:

Formula based on VIF < 5 -

default ~ Inventory Velocity Days + ROG Gross Block perc + ROG Market Capitalisation perc + Creditors Velocity Days + Inventory Ratio Latest + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Cash Flow From Investing Activities + Revenue earnings in forex + Equity Paid Up + Revenue expenses in forex + Cash Flow From Financing Activities + ROG Net Worth perc + ROG Total Assets perc + Other Income + Market Capitalisation + Cash Flow From Operating Activities + Selling Cost + ROG Capital Employed perc + Net Working Capital



**Figure 17: Results of 1st loop**





**Table 5: Summary of loop 1**

ROG Market Capitalisation perc has highest p-value among all other variables and is removed in the next iteration.

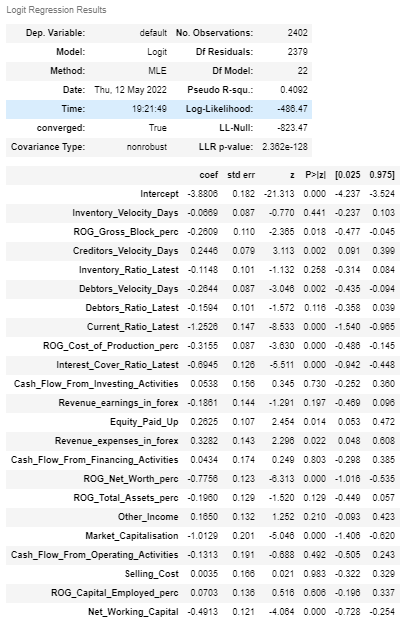
Loop 2:

Formula –

default ~ Inventory Velocity Days + ROG Gross Block perc + Creditors Velocity Days + Inventory Ratio Latest + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Cash Flow From Investing Activities + Revenue earnings in forex + Equity Paid Up + Revenue expenses in forex + Cash Flow From Financing Activities + ROG Net Worth perc + ROG Total Assets perc + Other Income + Market Capitalisation + Cash Flow From Operating Activities + Selling Cost + ROG Capital Employed perc + Net Working Capital



**Figure 18: Results of 2nd loop**



**Table 6: Summary of loop 2**

Selling Cost is removed in the next cycle.

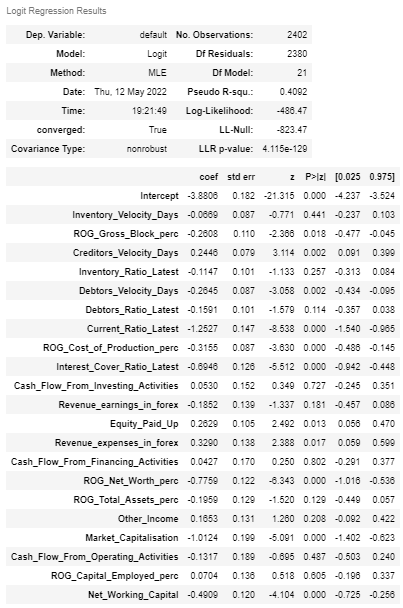
Loop 3:

Formula –

default ~ Inventory Velocity Days + ROG Gross Block perc + Creditors Velocity Days + Inventory Ratio Latest + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Cash Flow From Investing Activities + Revenue earnings in forex + Equity Paid Up + Revenue expenses in forex + Cash Flow From Financing Activities + ROG Net Worth perc + ROG Total Assets perc + Other Income + Market Capitalisation + Cash Flow From Operating Activities + ROG Capital Employed perc + Net Working Capital



**Figure 19: Results of 3rd loop**



**Table 7: Summary of loop 3**

Cash Flow from Financing Activities is removed from the current iteration to proceed with the next steps.

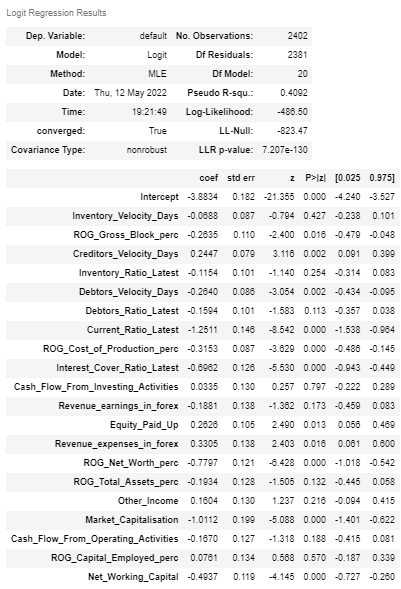
Loop 4:

Formula –

default ~ Inventory Velocity Days + ROG Gross Block perc + Creditors Velocity Days + Inventory Ratio Latest + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Cash Flow From Investing Activities + Revenue earnings in forex + Equity Paid Up + Revenue expenses in forex + ROG Net Worth perc + ROG Total Assets perc + Other Income + Market Capitalisation + Cash Flow From Operating Activities + ROG Capital Employed perc + Net Working Capital



**Figure 20: Results of 4th loop**



**Table 8: Summary of loop 4**

Cash Flow from Investing Activities is the next variable to be removed for the next step.

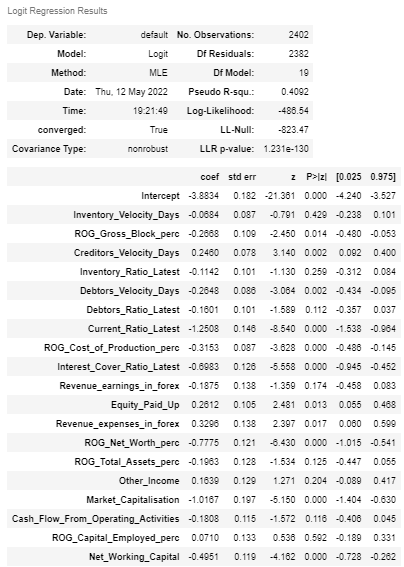
Loop 5:

Formula –

default ~ Inventory Velocity Days + ROG Gross Block perc + Creditors Velocity Days + Inventory Ratio Latest + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Revenue earnings in forex + Equity Paid Up + Revenue expenses in forex + ROG Net Worth perc + ROG Total Assets perc + Other Income + Market Capitalisation + Cash Flow from Operating Activities + ROG Capital Employed perc + Net Working Capital



**Figure 21: Results of 5th loop**



**Table 9: Summary of loop 5**

ROG Capital Employed perc is the next variable to be removed.

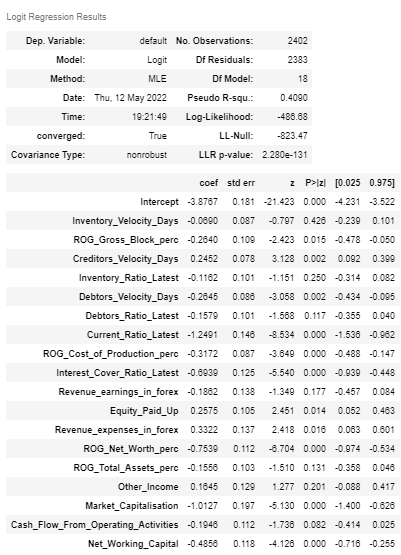
Loop 6:

Formula –

default ~ Inventory Velocity Days + ROG Gross Block perc + Creditors Velocity Days + Inventory Ratio Latest + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Revenue earnings in forex + Equity Paid Up + Revenue expenses in forex + ROG Net Worth perc + ROG Total Assets perc + Other Income + Market Capitalisation + Cash Flow from Operating Activities + Net Working Capital



**Figure 22: Results of 6th loop**



**Table 10: Summary of loop 6**

Inventory Velocity Days should be removed from the following steps.

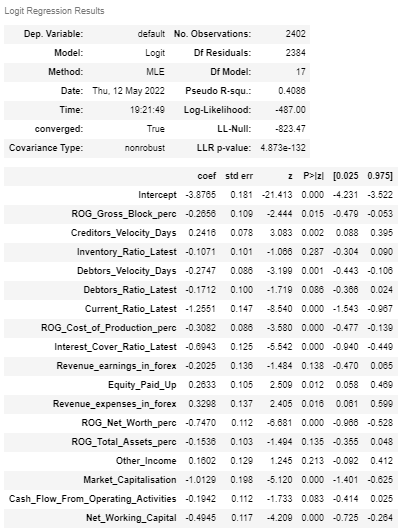
Loop 7:

Formula –

default ~ ROG Gross Block perc + Creditors Velocity Days + Inventory Ratio Latest + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Revenue earnings in forex + Equity Paid Up + Revenue expenses in forex + ROG Net Worth perc + ROG Total Assets perc + Other Income + Market Capitalisation + Cash Flow from Operating Activities + Net Working Capital



**Figure 23: Results of 7th loop**



**Table 11: Summary of loop 7**

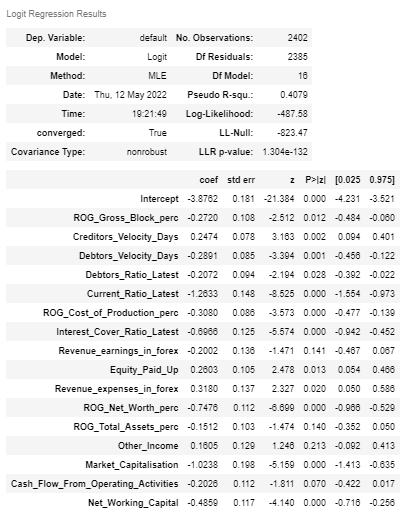
Inventory Ratio Latest is the current feature to be removed.

Loop 8:

Formula -  
default ~ ROG Gross Block perc + Creditors Velocity Days + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Revenue earnings in forex + Equity Paid Up + Revenue expenses in forex + ROG Net Worth perc + ROG Total Assets perc + Other Income + Market Capitalisation + Cash Flow from Operating Activities + Net Working Capital



**Figure 24: Results of 8th loop**



**Table 12: Summary of loop 8**

Other Income is the redundant variable for our analysis and hence removed.

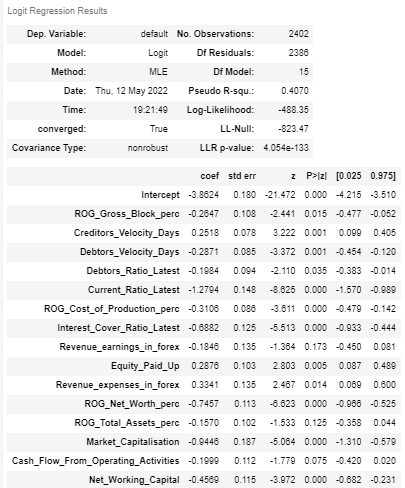
Loop 9:

Formula –

default ~ ROG Gross Block perc + Creditors Velocity Days + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Revenue earnings in forex + Equity Paid Up + Revenue expenses in forex + ROG Net Worth perc + ROG Total Assets perc + Market Capitalisation + Cash Flow from Operating Activities + Net Working Capital



**Figure 25: Results of 9th loop**



**Table 13: Summary of loop 9**

Revenue earnings in forex was removed to continue with further steps.

Loop 10:

Formula –

default ~ ROG Gross Block perc + Creditors Velocity Days + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Equity Paid Up + Revenue expenses in forex + ROG Net Worth perc + ROG Total Assets perc + Market Capitalisation + Cash Flow from Operating Activities + Net Working Capital



**Figure 26: Results of 10th loop**



**Table 14: Summary of loop 10**

ROG Total Assets perc is the next variable added to the list of redundant features.

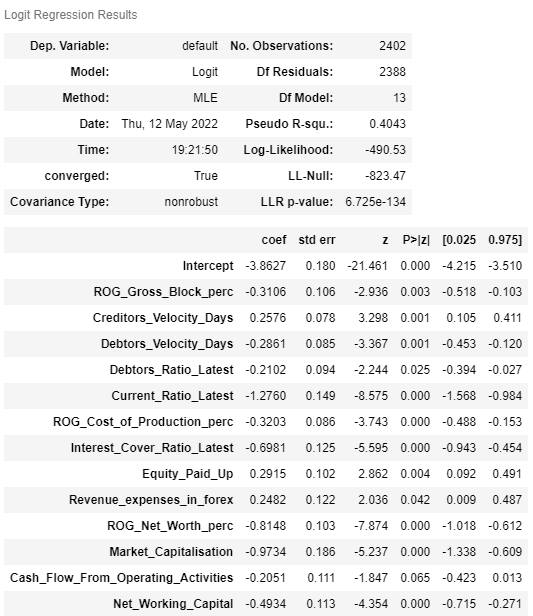
Loop 11:

Formula –

default ~ ROG Gross Block perc + Creditors Velocity Days + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Equity Paid Up + Revenue expenses in forex + ROG Net Worth perc + Market Capitalisation + Cash Flow from Operating Activities + Net Working Capital



**Figure 27: Results of 11th loop**



**Table 15: Summary of loop 11**

Cash Flow from Operating Activities will not be part of the next set of iterations.

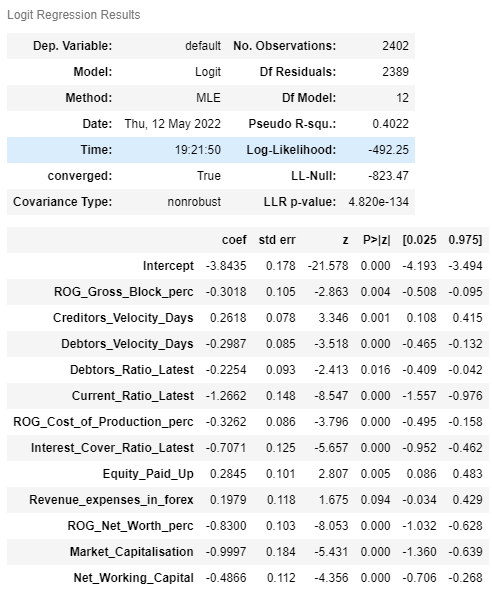
Loop 12:

Formula –

default ~ ROG Gross Block perc + Creditors Velocity Days + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Equity Paid Up + Revenue expenses in forex + ROG Net Worth perc + Market Capitalisation + Net Working Capital



**Figure 28: Results of 12th loop**



**Table 16: Summary of loop 12**

Revenue expenses in forex will not be used for the succeeding loop.

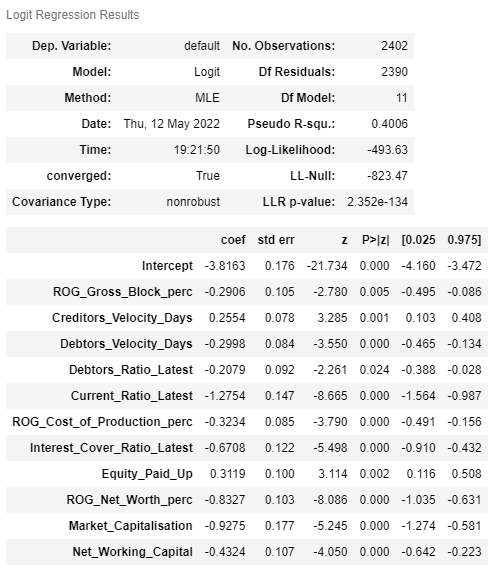
Loop 13:

Formula –

default ~ ROG Gross Block perc + Creditors Velocity Days + Debtors Velocity Days + Debtors Ratio Latest + Current Ratio Latest + ROG Cost of Production perc + Interest Cover Ratio Latest + Equity Paid Up + ROG Net Worth perc + Market Capitalisation + Net Working Capital

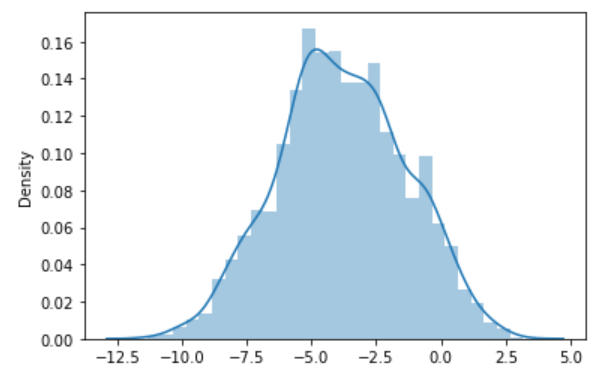


**Figure 29: Results of 13th loop**



**Table 17: Summary of loop 13**

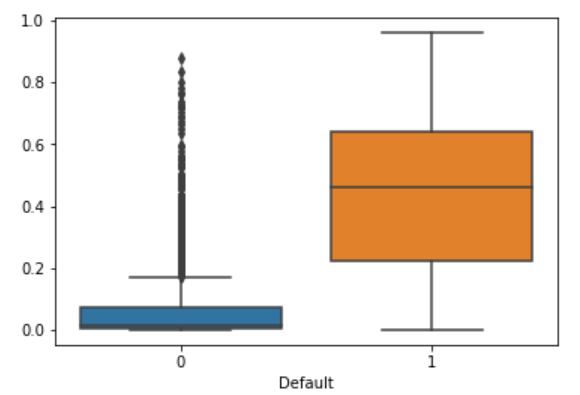
We can stop the iteration at this point as all variables have p-values < 0.05.



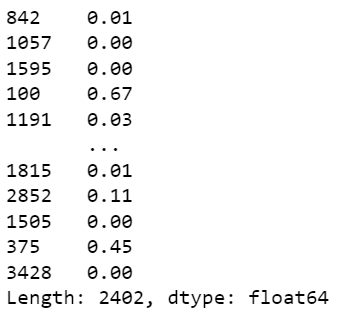
**Figure 30: Distribution plot of logit function values**



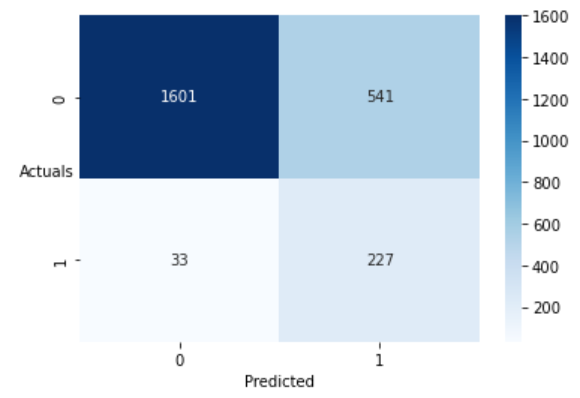
**Figure 31: Adjusted R-squared value of selected model**



**Figure 32: Boxplot to choose optimum cut-off**



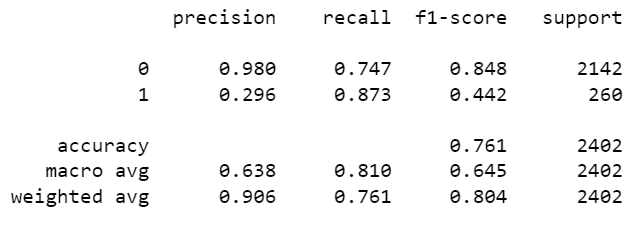
**Figure 33: Prediction on train set**



**Figure 34: Training data confusion matrix**

We see that adjusted R squared is now close to R squared, thus suggesting lesser insignificant variables in the model. We also notice that current model has no insignificant variables and can be used for prediction purposes.

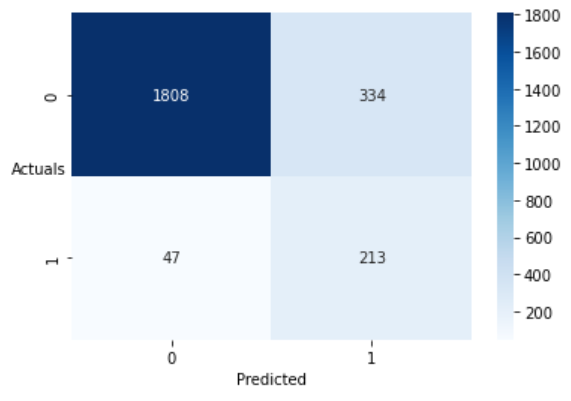
From the above boxplot, we need to decide on one such value of a cut-off which will give us the most reasonable descriptive power of the model. Let us take a cut-off of 0.05 and check.



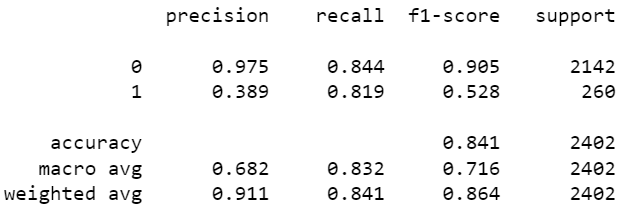
**Figure 35: Training data classification report**

As observed above, accuracy of the model i.e., %overall correct predictions are 76%. Sensitivity of the model is 87% i.e., 87% of those defaulted were correctly identified as defaulters by the model.

The optimal threshold is calculated as 0.13 based on roc scores.



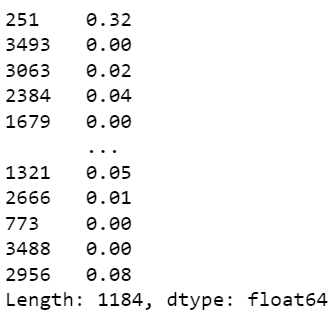
**Figure 36: Training data confusion matrix with revised threshold**



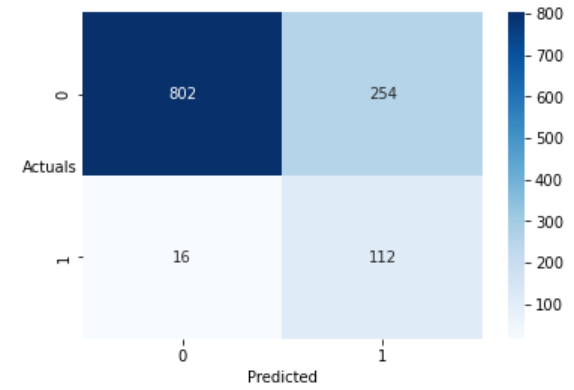
**Figure 37: Training data classification report with revised threshold**

Overall, 84% of correct predictions to total predictions were made by the model. 81% of those defaulted were correctly identified as defaulters by the model.

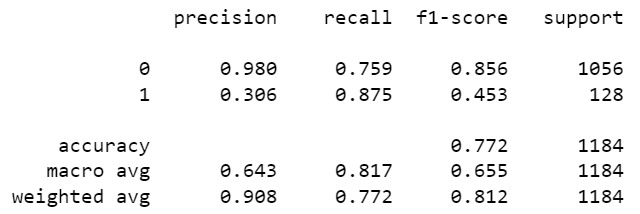
**Q1.7. Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model**



**Figure 38: Prediction on test set**

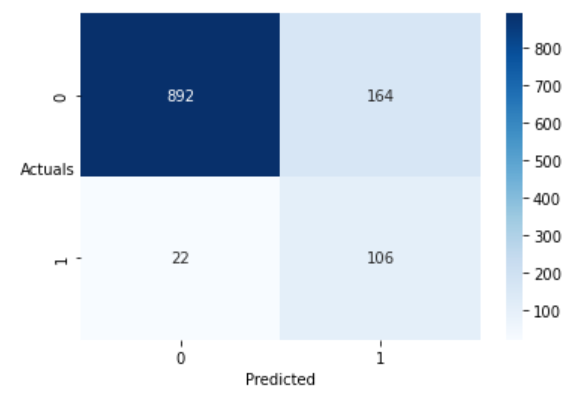


**Figure 39: Testing data confusion matrix**

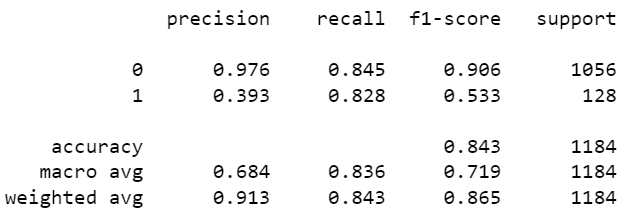


**Figure 40: Testing data classification report**

Overall, 77% of correct predictions to total predictions were made by the model. 87% of those defaulted were correctly identified as defaulters by the model.



**Figure 41: Testing data confusion matrix with revised threshold**



**Figure 42: Testing data classification report with revised threshold**

Interpretations of the model -

* Accuracy of the model i.e., %overall correct predictions is 77% & sensitivity of the model stands at 87% with threshold at 0.07.
* Accuracy of the model i.e., %overall correct predictions is 84% & sensitivity of the model stands at 82% with threshold at 0.13.
* We may choose cut-off of 0.13 as it gave higher model sensitivity & overall accuracy of the model in test dataset.
* Training and testing dataset performs equally well irrespective of threshold.
* We are able to achieve a descent recall value without overfitting. Considering the opportunities such as outliers, missing values and correlated features this is a fairly good model.
* It can be improved if we get better quality data where the features explaining the default are not missing to this extent. Of course, we can try other techniques which are not sensitive towards missing values and outliers.