S.VISHAL PGDSBA JULY 2021

finance and risk analytics BUsiness Report

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **SNO** | **TITLE** | **PAGE NO** |
| **A** | **LIST OF FIGURES** | **2** |
| **1** | **PROBLEM STATEMENT** | **3** |
| **2** | **EXPLORATORY DATA ANALYSIS** | **4** |
| **3** | **UNIVARIATE ANALYSIS** | **8** |
| **4** | **BIVARIATE ANALYSIS** | **15** |
| **5** | **MISSING VALUE TREATMENT** | **21** |
| **6** | **OUTLIER TREATMENT** | **22** |
| **7** | **LOGISTIC REGRESSION MODEL** | **23** |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **SNO** | **TITLE** | **PAGE NO** |
| **1** | **PAIR PLOT 1** | **15** |
| **2** | **PAIR PLOT 2** | **16** |
| **3** | **PAIR PLOT 3** | **17** |
| **4** | **PAIR PLOT 4** | **18** |
| **5** | **HEAT MAP OF CORRELATION** | **19** |
| **6** | **OUTLIER TREARTMENT BOX PLOT** | **23** |
| **7** | **ROC FOR TRAINING DATA** | **31** |
| **8** | **ROC FOR TEST DATA** | **32** |

**Problem Statement**

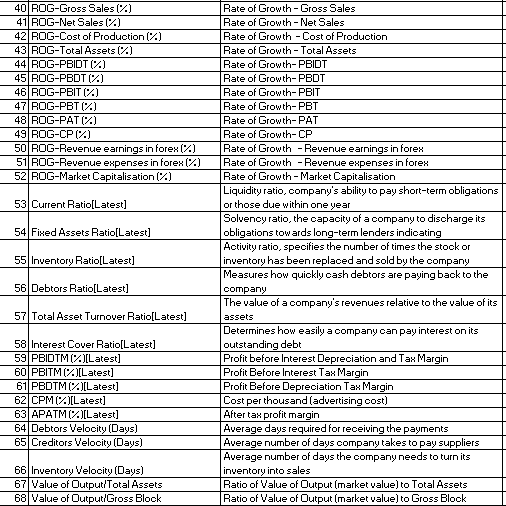
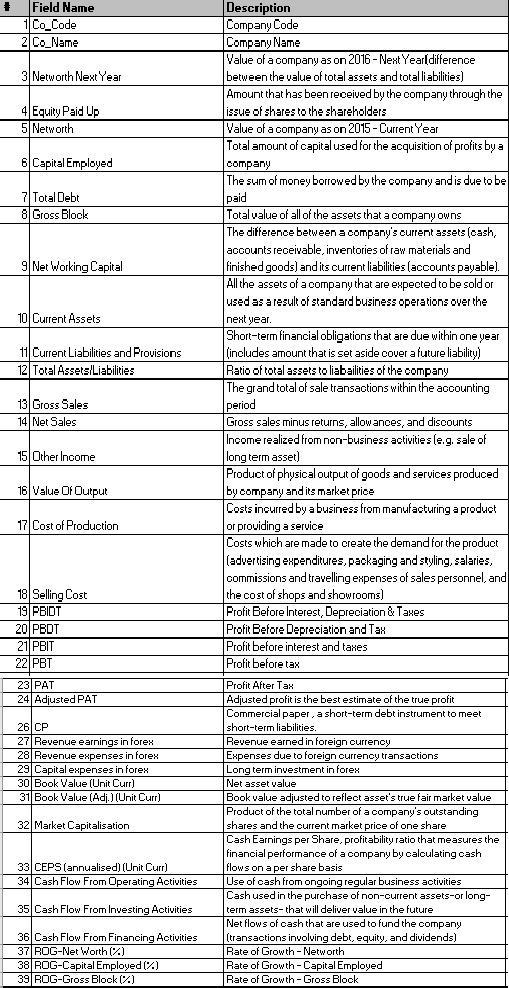
Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any 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 labeled field.

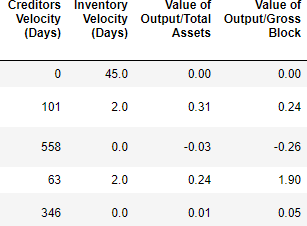
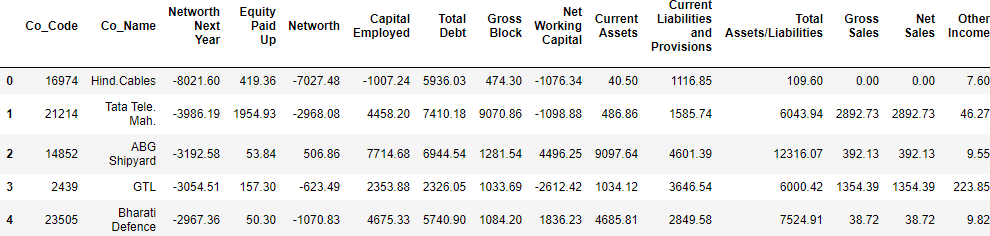
We need to create a default variable that should take the value of 1 when net worth next year is negative & 0 when net worth next year is positive.

**Data Set: Company\_Data2015-1.xlsx,**

**Data Dictionary: Credit Default Data Dictionary.xlsx**

**Exploratory Data Analysis (EDA)**

A quick glimpse of the data is shown below.



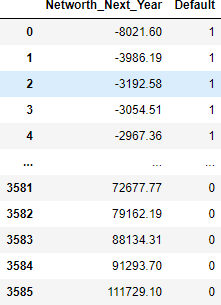
## Key points:

* The special characters in the variable names (Field names) have been replaced to get to the suggested variable names mentioned in data dictionary.
* There are 3586 rows and 67 columns (variables).
* All the variables are numeric type except one variable (Co\_Name) which is object type.
* For our analysis, Co\_Code and Co\_Name are dropped.
* There is no duplicate entry in the dataset.
* The problem statement requires to predict “default” status of the company where the “Networth Next Year” of the company is used to drive the “default” field. The “default” is 1 when “Networth Next Year” is negative and it is 0 when “Networth Next Year” is positive. The “Default” field is created and added to the dataset based on the condition mentioned above. Subsequently “Networth Next Year” is not considered further as it became redundant.
* There are missing values in 13 of the variables. Missing values will be treated with either mean or median values of corresponding variables.
* There are outliers in the dataset. It will be treated for our analysis.

**Target variable:**

As required, a transformed target variable “Default” is added to the dataset based on whether the variable “Networth Next Year” is positive or negative. “Default” will take value as 0 if “Networth Next Year” is positive, otherwise “Default” is 1.

The below picture captures the new variable “Default” (other variables are not displayed for clarity).



Also, the target variable “Default” is checked for counts.





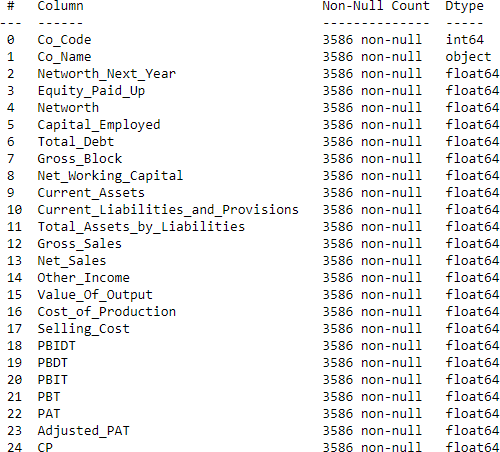
It is seen that almost 11% of the total entries in "Default" belong to category "1". The dataset has class imbalance issue.

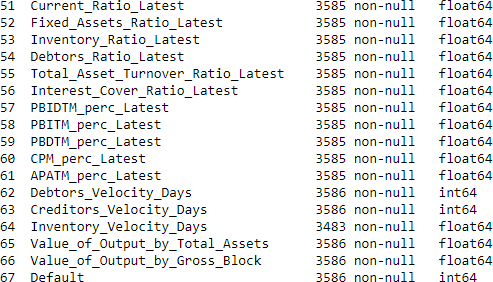
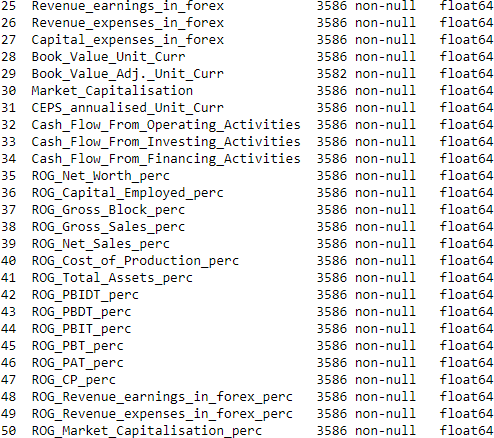
**Data type and Missing value:**

All the variables are of numeric types except Target variable “Default” and “Co\_Name”.

There are null values in 13 of the variables. These null values are imputed with median values as mean may not be correct one as the data variations are more and skewed.

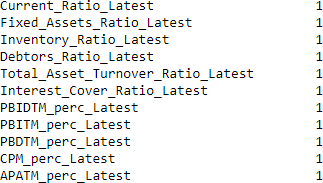
The following figure shows the overall data types and the variable with missing value

 Data inf



The following figure shows the missing value columns,

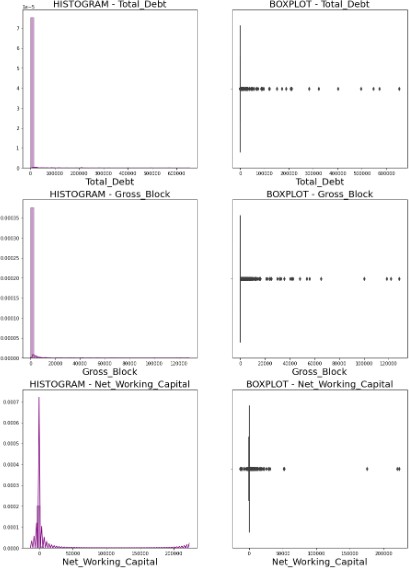
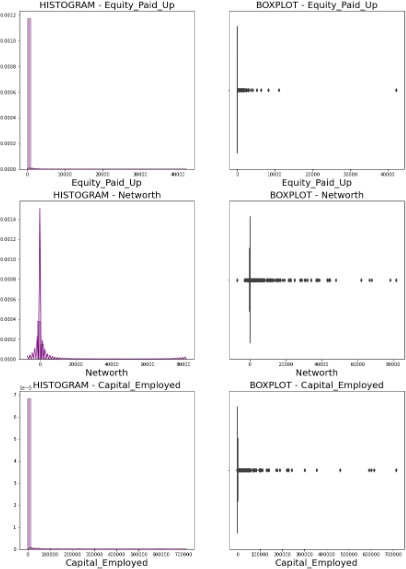






**Univariate analysis:**

Univariate analysis involving data distribution along with outlier detection (Boxplot) plots have been shown below. Due to large number of variables, the number of plots will be high.



Calendar

Description automatically generated

Calendar

Description automatically generated with medium confidenceCalendar

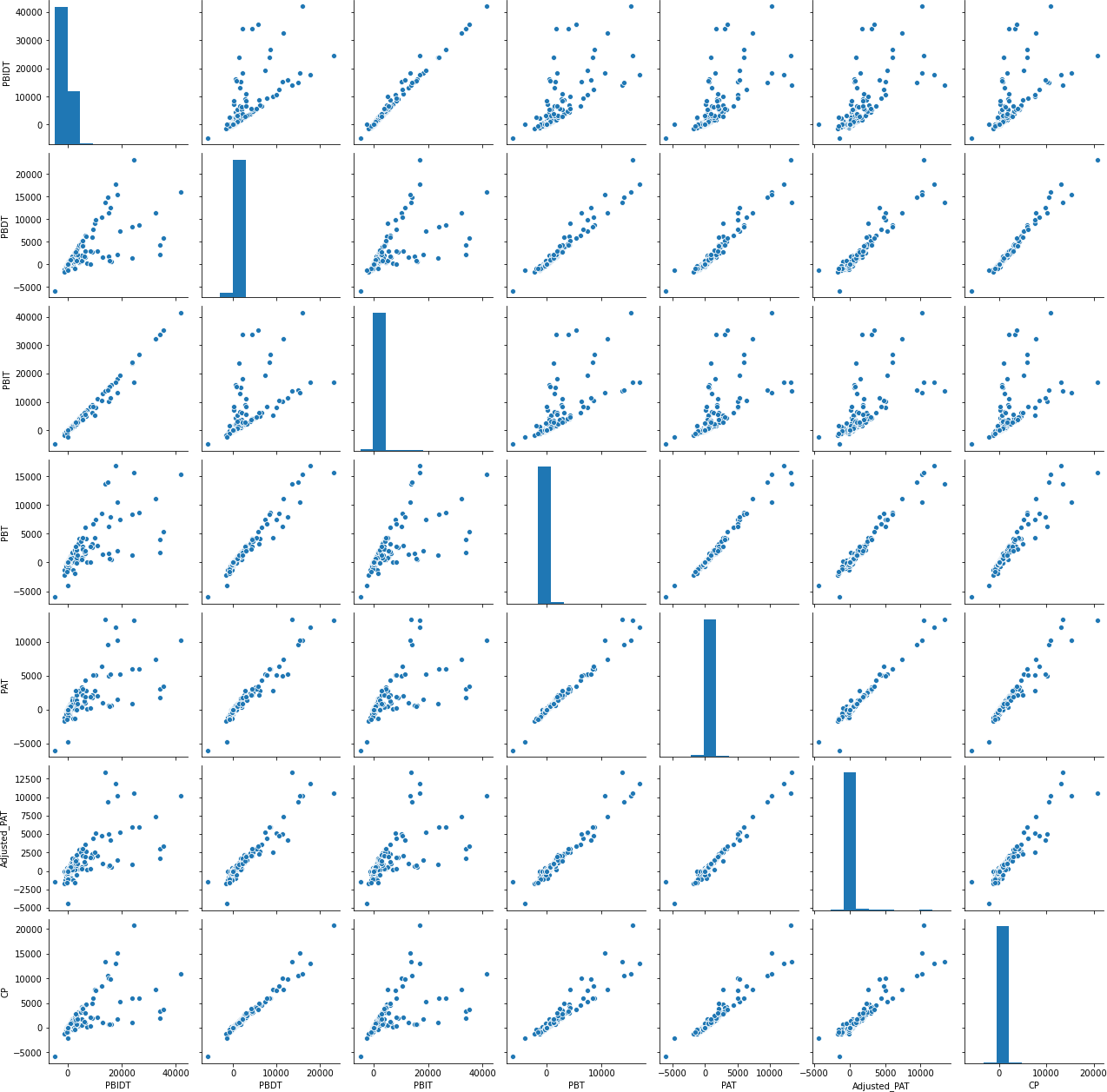
Description automatically generated

## Shape, calendar Description automatically generatedCalendar Description automatically generatedDiagram Description automatically generated with medium confidenceMost of the variables have skewed distribution. Also, all the variables have outliers. These outliers will be treated as we are going to apply Logistic regression to predict the outcome.

**Bi-variate Analysis:**

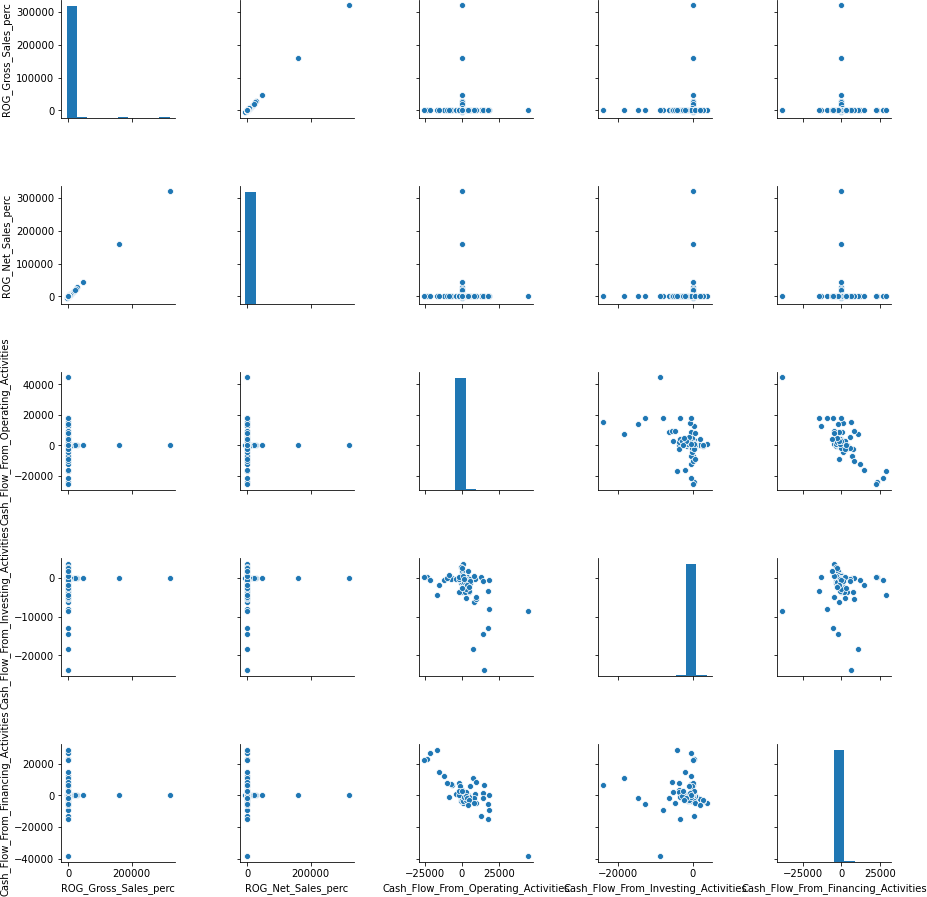
Bi-variate analysis includes pair plot and heatmap of correlation matrix. As the number of variables are high, the pair plot would not be so legible. For that reason, the pair plots are displayed for variables which are significant (derived using VIF score) in model prediction and which have significant correlations among each other.

**Pair plot - 1**



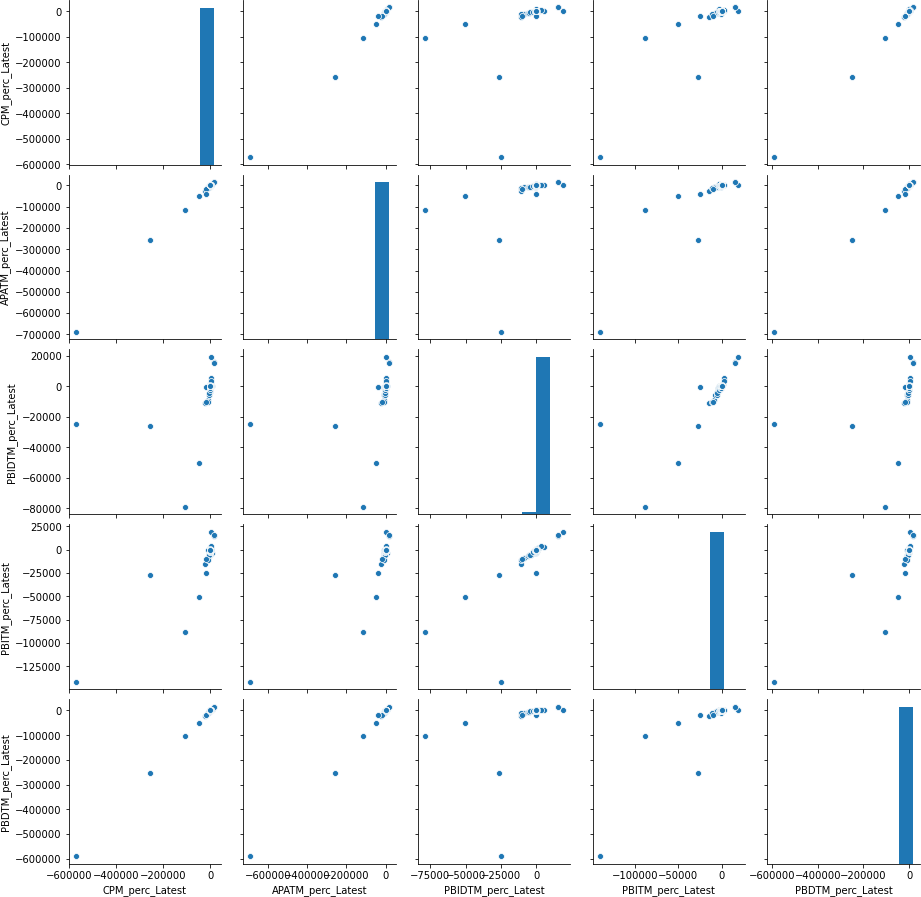
It is observed that there is high positive correlation between variables PBDT, PBIDT, PBIT, PBT, PAT, Adjusted PAT, and CP.

**Pair plot - 2**



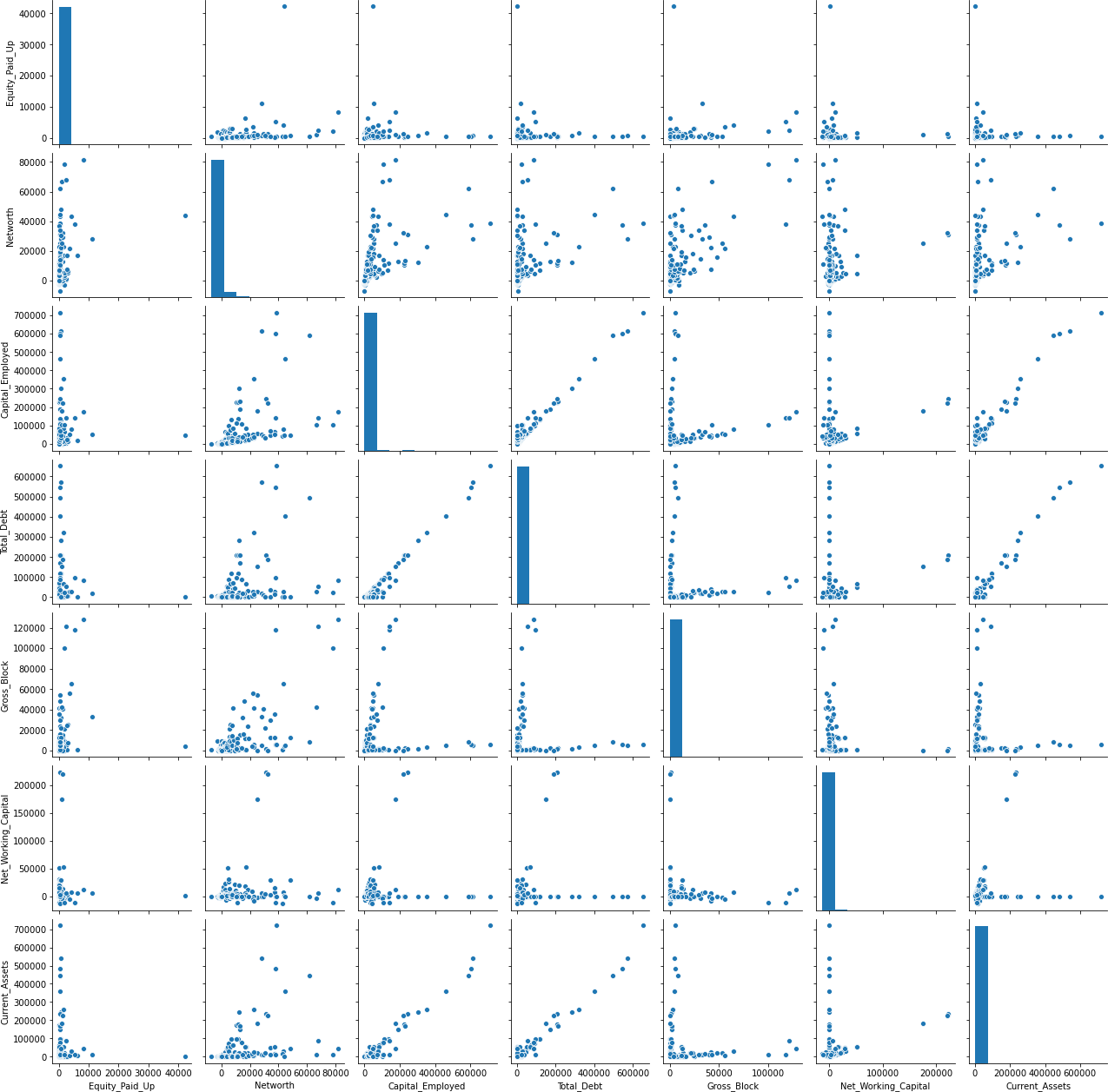
# It is observed that there is negative correlation between variables Cash\_Flow\_From\_Operating\_Activities and Cash\_Flow\_From\_Financing\_Activities.

**Pair plot - 3**



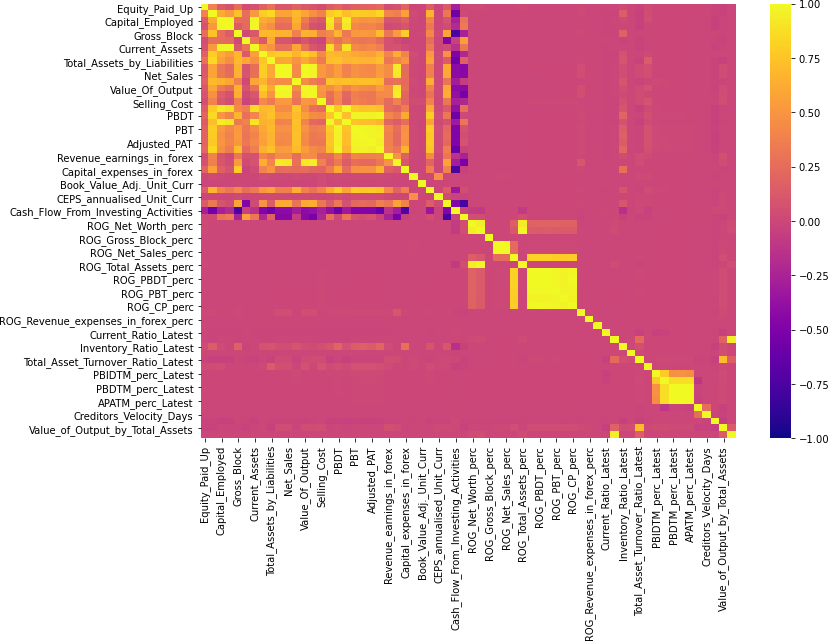
## It is observed that there is positive correlation between variables CPM\_perc\_Latest, APATM\_perc\_Latest, PBIDTM\_perc\_Latest, PBITM\_perc\_Latest, PBDTM\_perc\_Latest.

**Pair plot - 4**

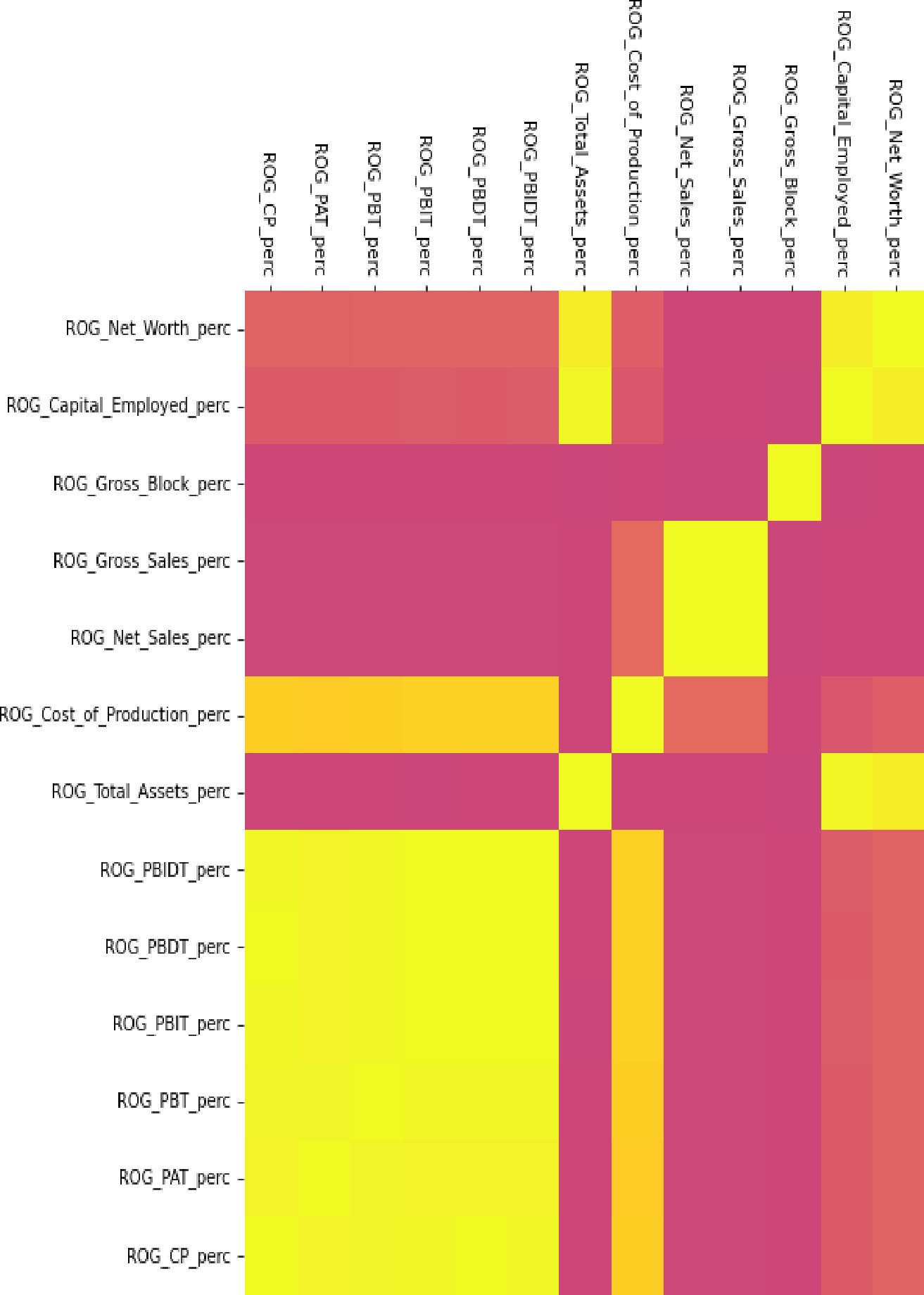


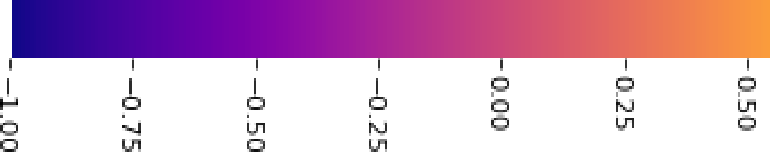
# It is observed that there is positive correlation between variables Capital\_Employed and Current\_Assets, Current\_Assets and Total\_debt, Total\_debt and Capital\_Employed.

**Heatmap of Correlation**



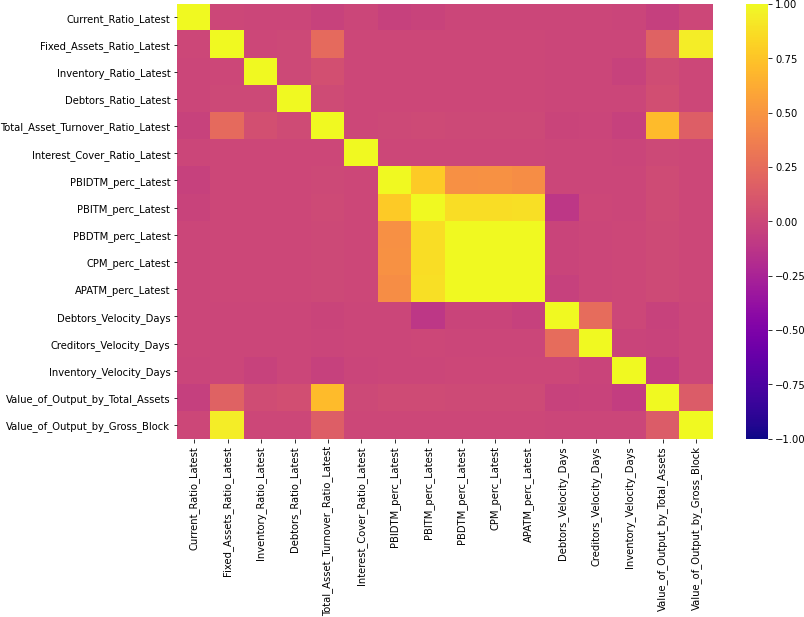
High positive and negative correlation between variables can be seen above. Majority of the variables are not correlated. Highly correlated variables are already captured in the pair plot. The above plot is dissected into smaller plots for more clarity in the subsequent pages of this report.







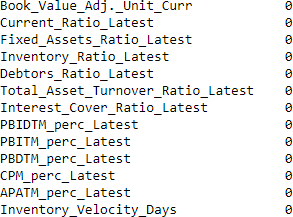


Many variables have correlation values close to 1 which denotes high collinearity among those variables.

**Missing value treatment:**

The missing values are treated with Simple Imputer Class. SimpleImputer is a scikit-learn class which is helpful in handling the missing data in the predictive model dataset. Here, median is used to fill up the missing value.

The following figure ensures that there is no missing values after treatment.



# Outlier treatment:

Outliers are present in all of the independent variables. For our dataset, we used IQR (Inter-Quartile Range) based calculation to treat the outliers. The following is the method,

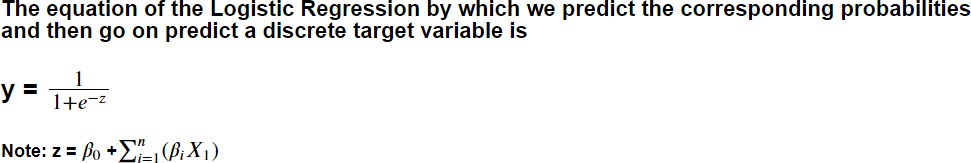
1. Arrange the data in ascending order
2. Calculate Q1 (the first Quarter)
3. Calculate Q3 (the third Quartile)
4. Find IQR = (Q3 - Q1)
5. Find the lower Range = Q1 - (1.5 \* IQR)
6. Find the upper Range = Q3 + (1.5 \* IQR)

Once the upper bound and lower bound range is calculated, we snap the values above upper range and values below lower range to upper and lower range values respectively.

It was observed that maximum of 45% of the total rows are outliers for a particular variable in the dataset. And the mean numbers of outliers above and below the specified band is around 18%.

## The following figure shows the boxplot of variables after outlier treatment. Chart, histogram Description automatically generated

# Logistic Regression Model (using statsmodel library)



## In our present case, we will be using stats models modules for logistic regression as required by client.

**Some of the libraries we will be using are as follows,**

* 1. **From sklearn.model\_selection train\_test\_split for splitting the train and test set.**
  2. **variance\_inflation\_factor module from statsmodels.stats.outliers\_influence**
  3. **metrics from sklearn**
  4. **roc\_auc\_score, roc\_curve, from sklearn.metrics**
  5. **classification\_report, confusion\_matrix, plot\_confusion\_matrix from sklearn.metrics**

Since there are larger number of variables present in the dataset and we observed that many of the variables are highly correlated, the problem of multicollinearity may occur. So, we identified those correlated variables through **VIF (variance inflation factor) calculation**. We did not consider the variables for model building whose VIF is greater than 5 (industry standard). The following variables are used for the preliminary model building after VIF calculation.

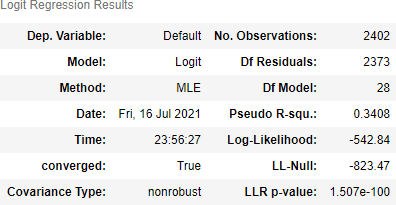


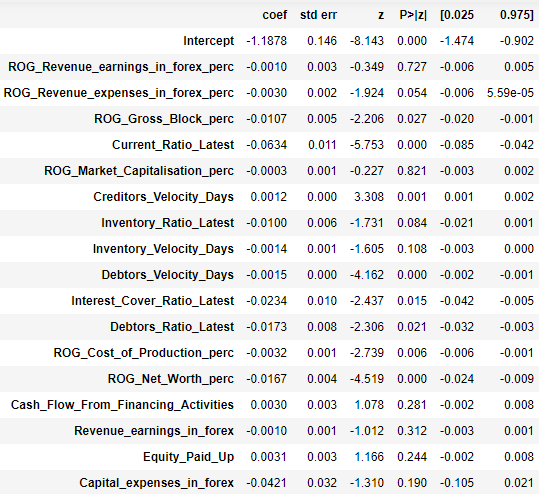
# Train test split:

The original data frame except the variables Co\_Code, Co\_Name, Networth\_Next\_Year is divided into dependent and independent variable type data frame. Then both independent and dependent variable data frame is split into 67:33 (train: test) ratio. One requirement for Stats model is that dependent and independent variables should be contained in same dataframe. So, concatenation was performed to combine dependent and independent variables arrays.



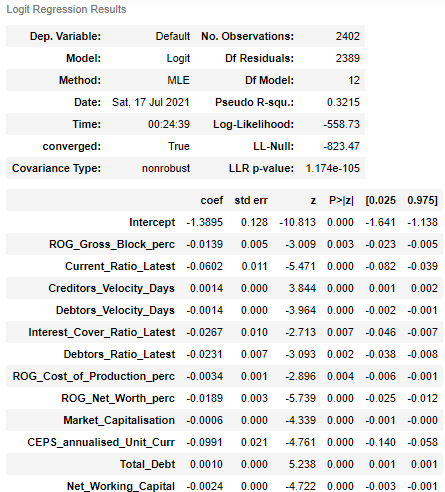
# Model building:

A preliminary logistic regression model is built on the **train set** with the variables whose VIF value is less than 5. The model output is shown below.



We checked the probability values for each independent variable and some of them are found to be > 0.05. So, at 95% confidence level, if p < 0.05, we can say that there is a relation between dependent and other independent variable. Alternately we can say that variables whose p > 0.05 donot have influence on the dependent variable. Therefore, ***a new model*** is prepared by discarding the variables whose p > 0.05.

**Model 2 summary (new model):**



The new model (model 2) has all the variables with p < 0.05. This model will be considered for **Test set prediction and performance evaluation**.

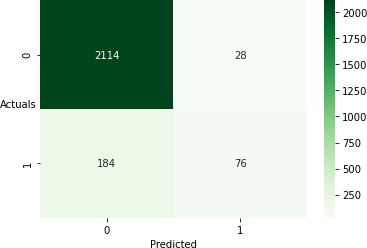
# Model Evaluation on the Training Data

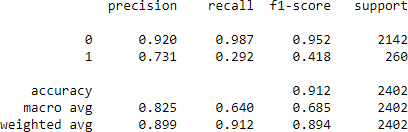
First, we will check the training set performance with predicted classes with **0.5 probability cut-off.**

Different matrices were used to check the model performance, namely,

1. Confusion matrix
2. Classification report (precision, recall, accuracy)
3. AUC-ROC curve

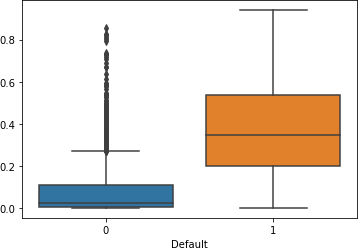
# Performance of 0.5 probability cut-off:



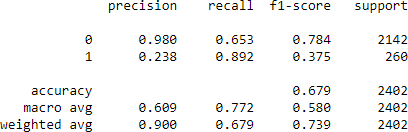
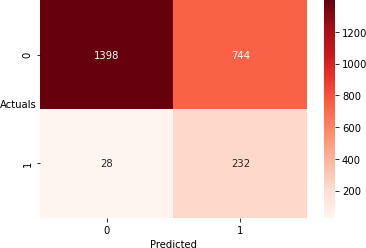


## Overall, 91% of correct predictions to total predictions were made by the model. 29% of those defaulted were correctly identified as defaulters by the model, which is not so good number.

**So, we will change the probability cut-off to 0.07 as from the boxplot it is clear that “Default” status 0 has very low probability median.**

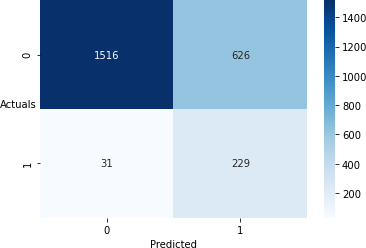


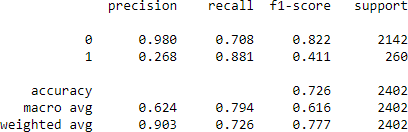
**Performance of 0.07 probability cut-off:**



Accuracy of the model i.e., %overall correct predictions has decreased from 91% to 68% but sensitivity of the model has increased from 29% to 89%, which is good for our prediction. But we will try with some more probability cut-off values.

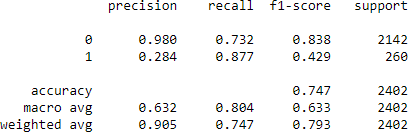
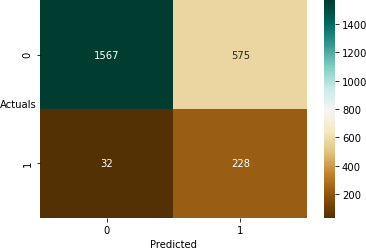
# Performance of 0.09 probability cut-off:





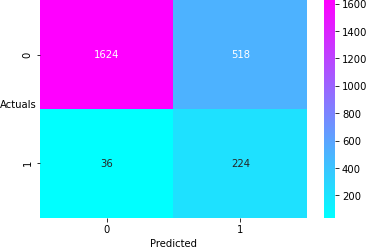
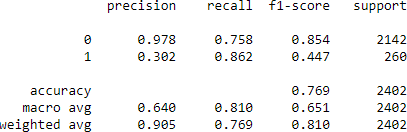
Accuracy of the model i.e., %overall correct predictions has increased from 71% to 73% but sensitivity of the model has decreased from 89% to 88%.

**Performance of 0.1 probability cut-off:**



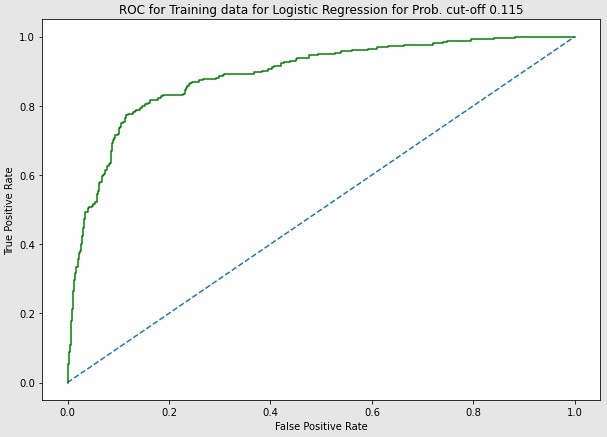
Accuracy of the model i.e., %overall correct predictions has increased from 73% to 75% but sensitivity of the model has not decreased (88%). But we will try with some more probability cut- off values.

# Performance of 0.115 probability cut-off:



Accuracy of the model i.e., %overall correct predictions has increased from 75% to 77% but sensitivity of the model has decreased slightly from 88% to 86%. We will keep this model (with p= 0.115 as cut-off) for further analysis as we are trying to maintain a balance between Accuracy and Recall.

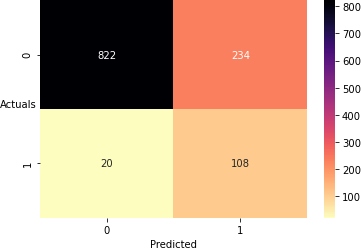
## AUC-ROC Curve:

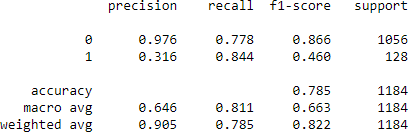


**The model AUC for training set is 0.888.**

**Model Evaluation on the Testing Data**

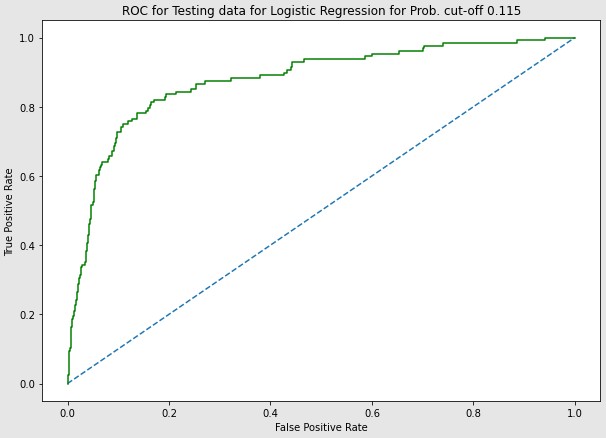
**The model (with p= 0.115 as cut-off) is checked to predict on the test set. The confusion matrix and classification report is discussed below.**





Accuracy of the model i.e., % overall correct prediction is 78% and sensitivity of the model is 84%. The model performs well on the test set also.

## The model AUC for testing set is 0.877.



While the model results between training and test sets are similar, indicating no under or overfitting issues, overall prediction of the model is weak. There is a scope of improvement on the accuracy and recall values by using techniques like re-sampling, cross validation etc.,