Project – 1

**Company Loan Default Prediction**

Ananth Gopalakrishna Achar

PGP-DSBA 2020 - 2021

**Table of Contents**

**Contents**

**Executive Summary.................................................................................................................................2**

**Data Description ......................................................................................................................................2**

**Exploratory Data Analysis........................................................................................................................4**

**Descriptive Stats ......................................................................................................................................5**

**Check for missing values in the dataset:...............................................................................................7**

**Checking for Outliers..............................................................................................................................11**

**Correlation check....................................................................................................................................19**

**Splitting the data.....................................................................................................................................20**

**Calculating Variance Inflation Factor....................................................................................................21**

**Logistic Regression Modelling with Statsmodel library.....................................................................24**

**Model 2....................................................................................................................................................25**

**Validating the model on the train and choosing the optimum cut-off...........................................26**

**Model 3.....................................................................................................................................................28**

**Conclusion................................................................................................................................................30**

**Executive Summary:**

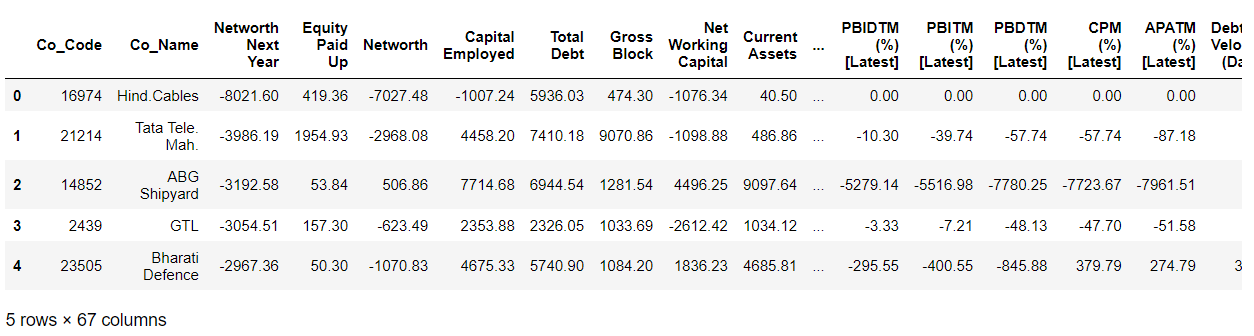
A bank’s loan data is given to us containing the list of variables taken from the Financial statements of the companies. We are expected to predict whether the company will default on their loan obligation or not. Attributes like credit worthiness, interest rates on the loan and financial stability of the company can be inferred from whether the company has defaulted or not.

**Data Description:**

|  |  |  |
| --- | --- | --- |
| **#** | **Field Name** | **Description** |
| 1 | Co\_Code | Company Code |
| 2 | Co\_Name | Company Name |
| 3 | Networth Next Year | Value of a company as on 2016 - Next Year(difference between the value of total assets and total liabilities) |
| 4 | Equity Paid Up | Amount that has been received by the company through the issue of shares to the shareholders |
| 5 | Networth | Value of a company as on 2015 - Current Year |
| 6 | Capital Employed | Total amount of capital used for the acquisition of profits by a company |
| 7 | Total Debt | The sum of money borrowed by the company and is due to be paid |
| 8 | Gross Block | Total value of all of the assets that a company owns |
| 9 | Net Working Capital | The difference between a company's current assets (cash, accounts receivable, inventories of raw materials and finished goods) and its current liabilities (accounts payable). |
| 10 | Current Assets | All the assets of a company that are expected to be sold or used as a result of standard business operations over the next year. |
| 11 | Current Liabilities and Provisions | Short-term financial obligations that are due within one year (includes amount that is set aside cover a future liability) |
| 12 | Total Assets/Liabilities | Ratio of total assets to liabailities of the company |
| 13 | Gross Sales | The grand total of sale transactions within the accounting period |
| 14 | Net Sales | Gross sales minus returns, allowances, and discounts |
| 15 | Other Income | Income realized from non-business activities (e.g. sale of long term asset) |
| 16 | Value Of Output | Product of physical output of goods and services produced by company and its market price |
| 17 | Cost of Production | Costs incurred by a business from manufacturing a product or providing a service |
| 18 | Selling Cost | Costs which are made to create the demand for the product (advertising expenditures, packaging and styling, salaries, commissions and travelling expenses of sales personnel, and the cost of shops and showrooms) |
| 19 | PBIDT | Profit Before Interest, Depreciation & Taxes |
| 20 | PBDT | Profit Before Depreciation and Tax |
| 21 | PBIT | Profit before interest and taxes |
| 22 | PBT | Profit before tax |
| 23 | PAT | Profit After Tax |
| 24 | Adjusted PAT | Adjusted profit is the best estimate of the true profit |
| 26 | CP | Commercial paper , a short-term debt instrument to meet short-term liabilities. |
| 27 | Revenue earnings in forex | Revenue earned in foreign currency |
| 28 | Revenue expenses in forex | Expenses due to foreign currency transactions |
| 29 | Capital expenses in forex | Long term investment in forex |
| 30 | Book Value (Unit Curr) | Net asset value |
| 31 | Book Value (Adj.) (Unit Curr) | Book value adjusted to reflect asset's true fair market value |
| 32 | Market Capitalisation | Product of the total number of a company's outstanding shares and the current market price of one share |
| 33 | CEPS (annualised) (Unit Curr) | Cash Earnings per Share, profitability ratio that measures the financial performance of a company by calculating cash flows on a per share basis |
| 34 | Cash Flow From Operating Activities | Use of cash from ongoing regular business activities |
| 35 | Cash Flow From Investing Activities | Cash used in the purchase of non-current assets–or long-term assets– that will deliver value in the future |
| 36 | Cash Flow From Financing Activities | Net flows of cash that are used to fund the company (transactions involving debt, equity, and dividends) |
| 37 | ROG-Net Worth (%) | Rate of Growth - Networth |
| 38 | ROG-Capital Employed (%) | Rate of Growth - Capital Employed |
| 39 | ROG-Gross Block (%) | Rate of Growth - Gross Block |
| 40 | ROG-Gross Sales (%) | Rate of Growth - Gross Sales |
| 41 | ROG-Net Sales (%) | Rate of Growth - Net Sales |
| 42 | ROG-Cost of Production (%) | Rate of Growth - Cost of Production |
| 43 | ROG-Total Assets (%) | Rate of Growth - Total Assets |
| 44 | ROG-PBIDT (%) | Rate of Growth- PBIDT |
| 45 | ROG-PBDT (%) | Rate of Growth- PBDT |
| 46 | ROG-PBIT (%) | Rate of Growth- PBIT |
| 47 | ROG-PBT (%) | Rate of Growth- PBT |
| 48 | ROG-PAT (%) | Rate of Growth- PAT |
| 49 | ROG-CP (%) | Rate of Growth- CP |
| 50 | ROG-Revenue earnings in forex (%) | Rate of Growth - Revenue earnings in forex |
| 51 | ROG-Revenue expenses in forex (%) | Rate of Growth - Revenue expenses in forex |
| 52 | ROG-Market Capitalisation (%) | Rate of Growth - Market Capitalisation |
| 53 | Current Ratio[Latest] | Liquidity ratio, company's ability to pay short-term obligations or those due within one year |
| 54 | Fixed Assets Ratio[Latest] | Solvency ratio, the capacity of a company to discharge its obligations towards long-term lenders indicating |
| 55 | Inventory Ratio[Latest] | Activity ratio, specifies the number of times the stock or inventory has been replaced and sold by the company |
| 56 | Debtors Ratio[Latest] | Measures how quickly cash debtors are paying back to the company |
| 57 | Total Asset Turnover Ratio[Latest] | The value of a company's revenues relative to the value of its assets |
| 58 | Interest Cover Ratio[Latest] | Determines how easily a company can pay interest on its outstanding debt |
| 59 | PBIDTM (%)[Latest] | Profit before Interest Depreciation and Tax Margin |
| 60 | PBITM (%)[Latest] | Profit Before Interest Tax Margin |
| 61 | PBDTM (%)[Latest] | Profit Before Depreciation Tax Margin |
| 62 | CPM (%)[Latest] | Cost per thousand (advertising cost) |
| 63 | APATM (%)[Latest] | After tax profit margin |
| 64 | Debtors Velocity (Days) | Average days required for receiving the payments |
| 65 | Creditors Velocity (Days) | Average number of days company takes to pay suppliers |
| 66 | Inventory Velocity (Days) | Average number of days the company needs to turn its inventory into sales |
| 67 | Value of Output/Total Assets | Ratio of Value of Output (market value) to Total Assets |
| 68 | Value of Output/Gross Block | Ratio of Value of Output (market value) to Gross Block |

**Exploratory Data Analysis:**

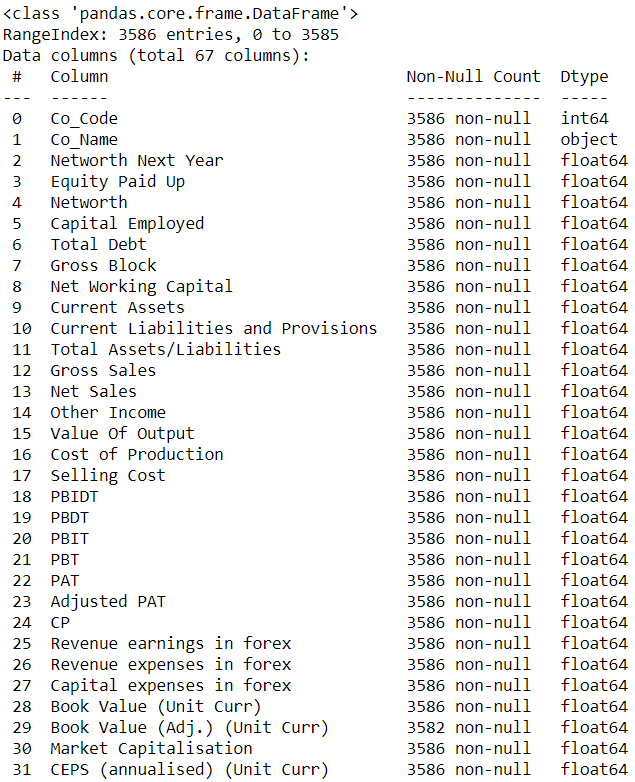
The head of the data looks like this, many columns are not shown here due to space constraint:

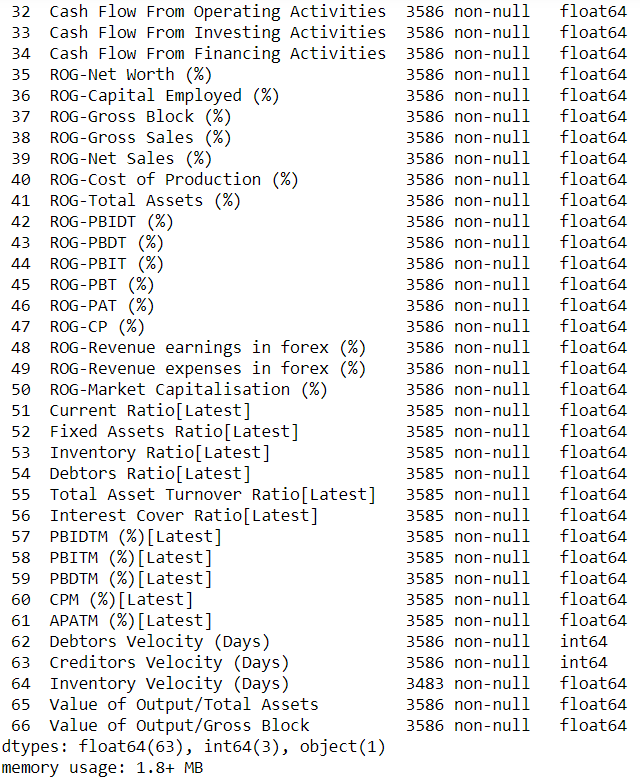


The shape of the data is as follows:



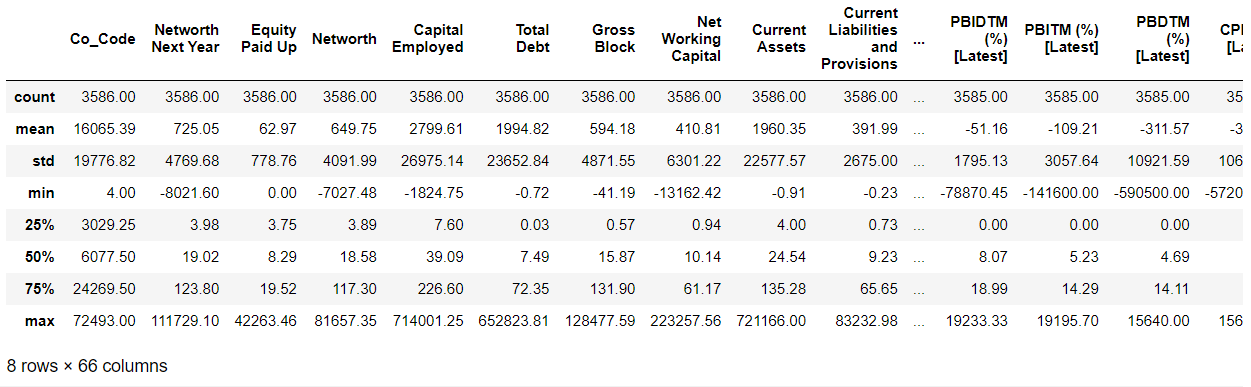
We then look at the info of the data which contains the name of each column, the number of non null entries and the data type:





It can be inferred from the above output that there are some missing values and all the variables have a numerical data type of Float or Int except for the column: Co\_Name which is the name of the company.

**Descriptive stats:**



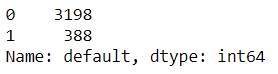
It can be seen that most of the data is left skewed.

We then fix the columns names to a more readable manner as follows:

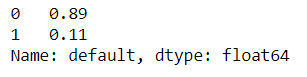


We then create a binary target variable: ‘default ‘ in such a way that if the Networth\_Next\_Year is greater than 0 then the company has not defaulted and if not, the company has defaulted with the representation shown as 0 and 1 respectively.

Let us then check the value counts of the target variable: default



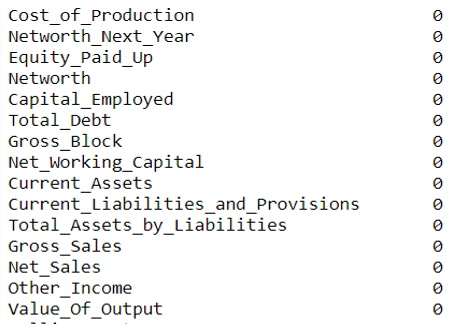
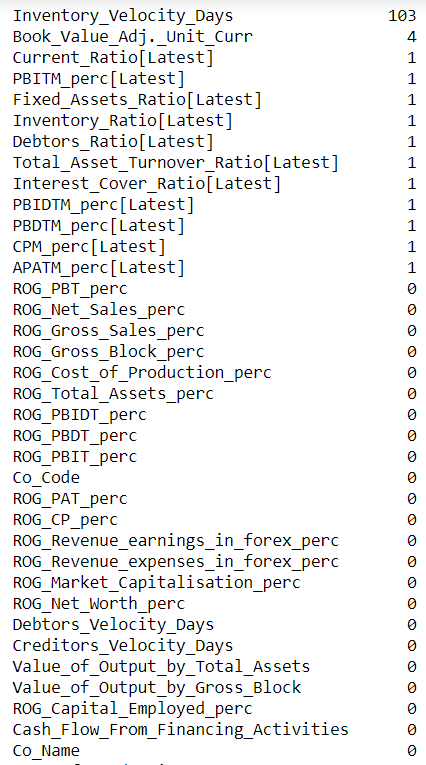
Looking at the proportion:

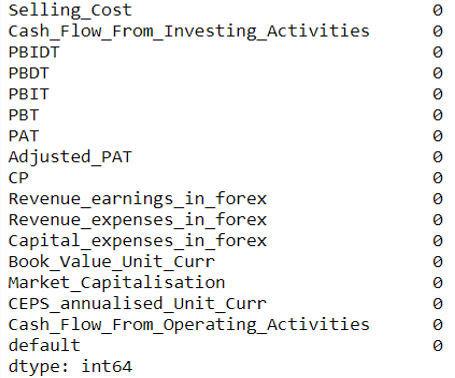


The data is highly disproportionate in terms of the number of defaults.

**Checking for Missing Values in the data:**

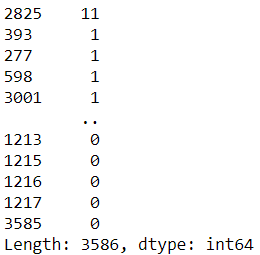
**Columnwise:**





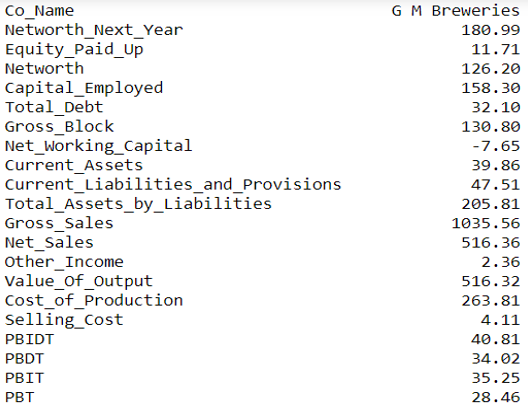
We have a total of 118 missing values.

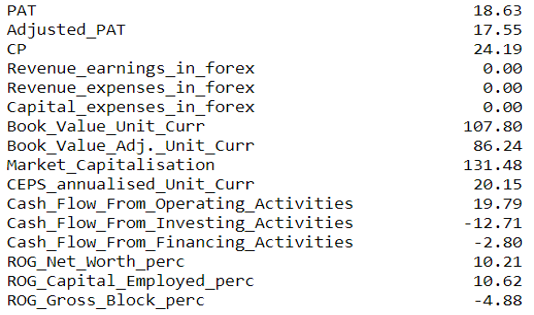
We then look at the missing values by each company:

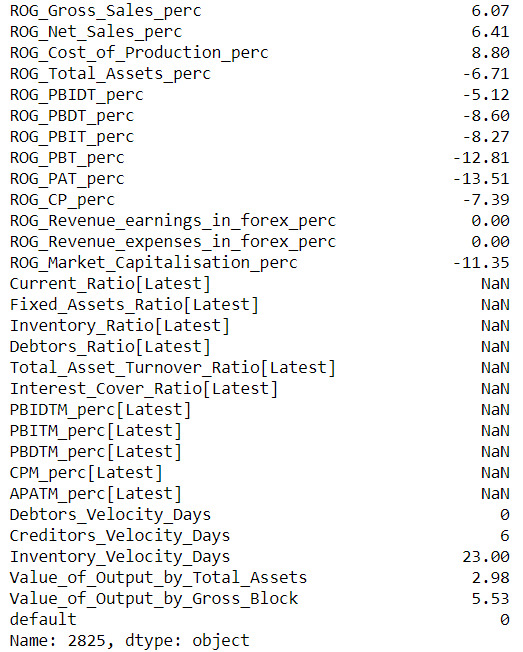


We then discover that there is a company that has 11 columns of missing values.

We then check the company and discover the following:

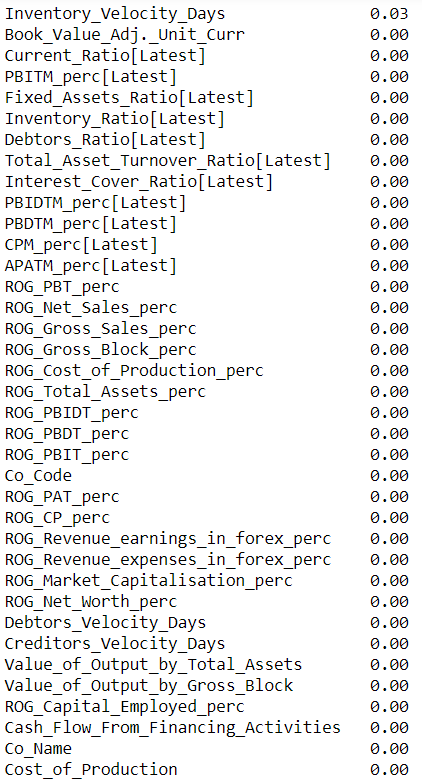


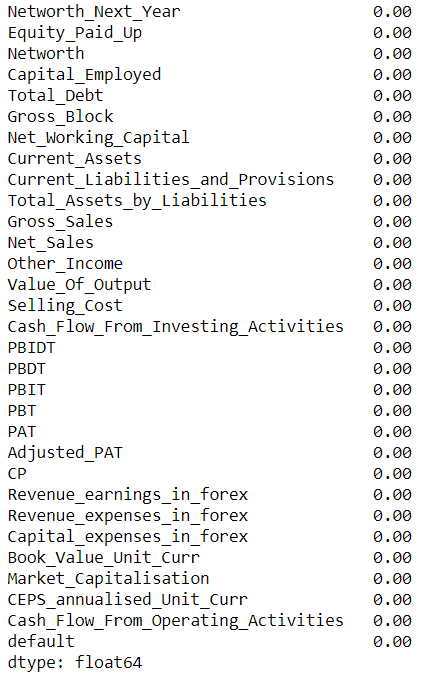




Even though the Networth\_Next\_Year is positive (i.e. this company has not defaulted), this company has not provided 11 columns of data. Ideally we would have to go back to the business requesting data but in this case we shall go ahead and impute the same to the best of our knowledge.

We then look at the proportion of missing values per column:



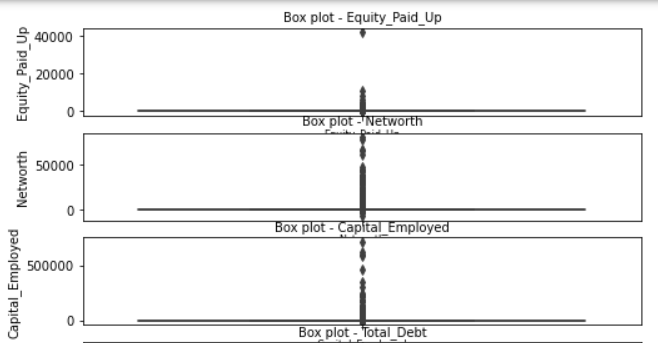
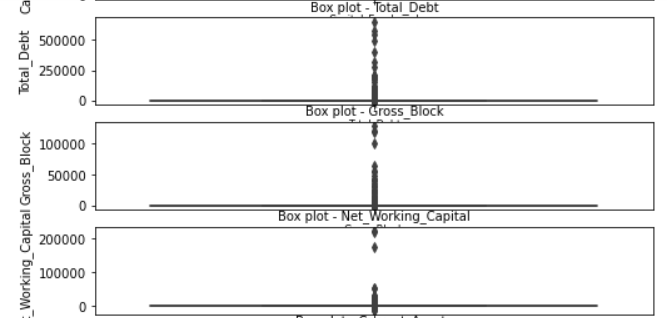
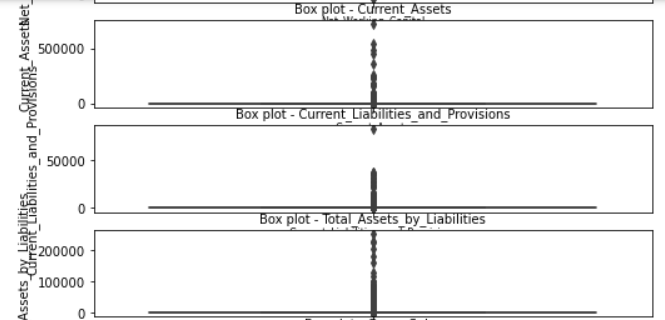


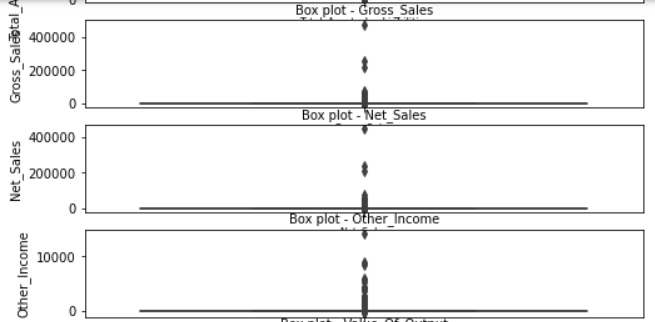
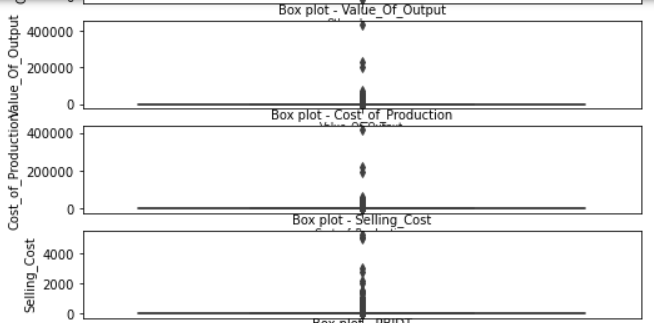
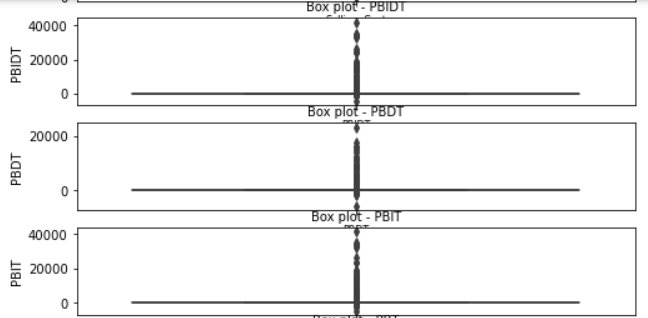
Only 3% of the data in the Inventory Velocity days is missing, so we will impute the missing values in this column using the SimpleImputer.

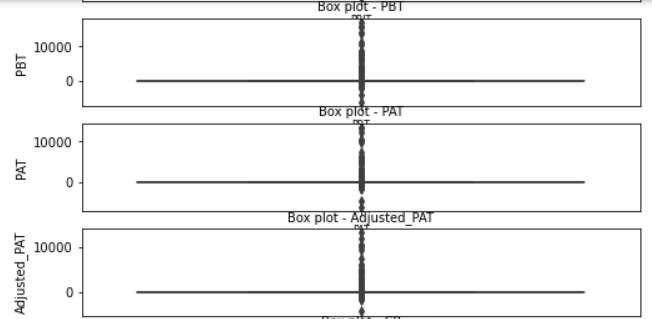
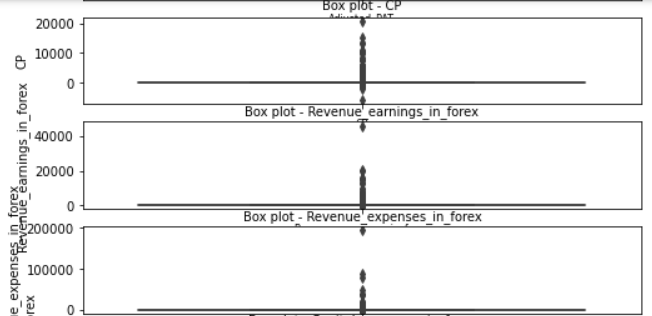
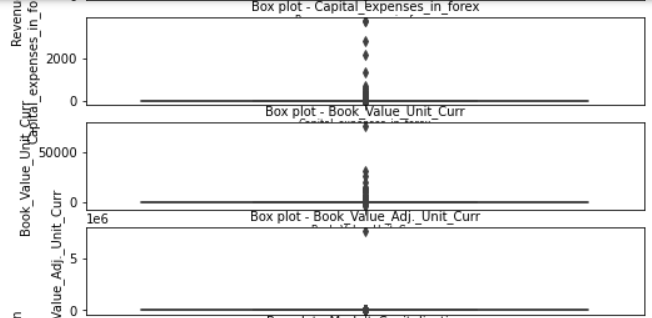
Removing the columns: Company Code, Company Name and Value of a company as on 2016 - Next Year(difference between the value of total assets and total liabilities) as it is not useful for our analysis.

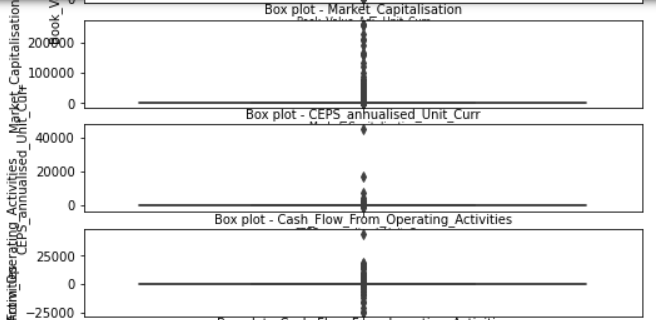
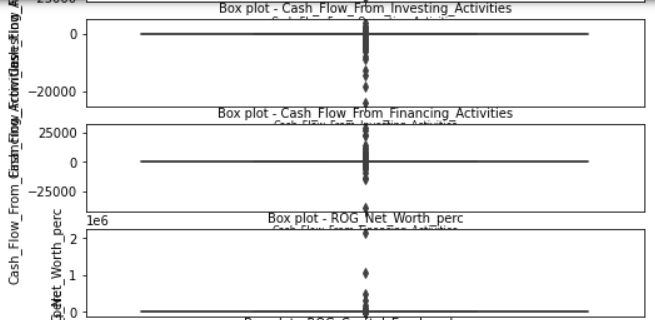
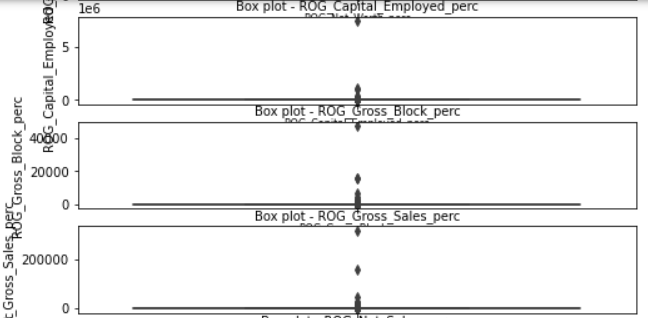
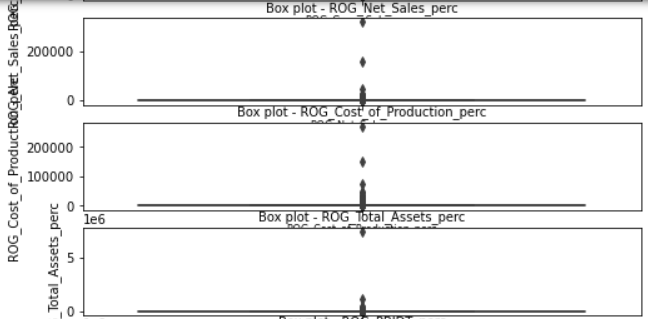
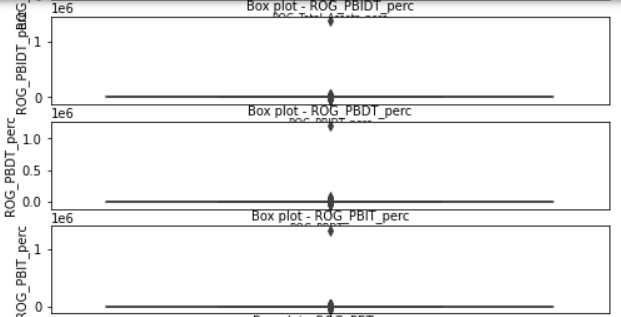
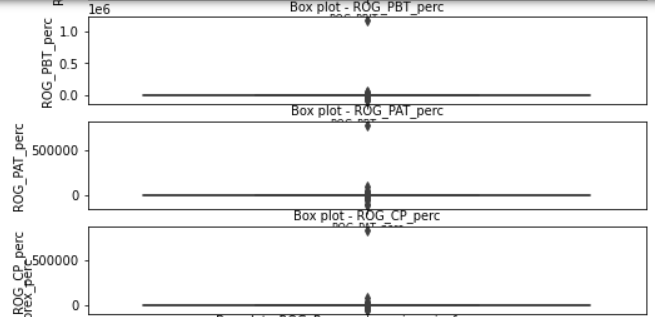
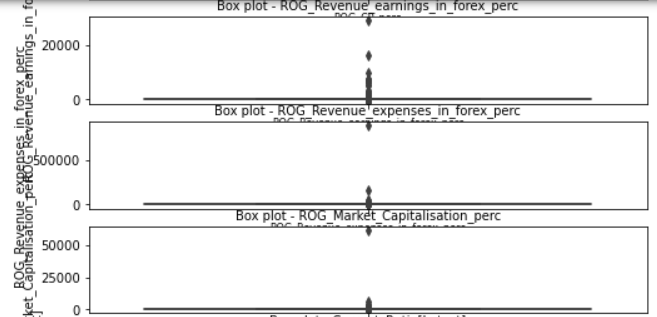
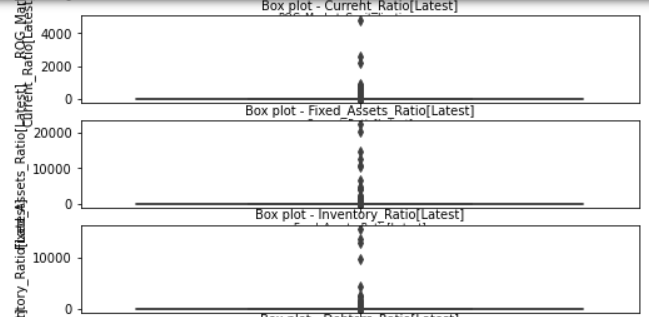
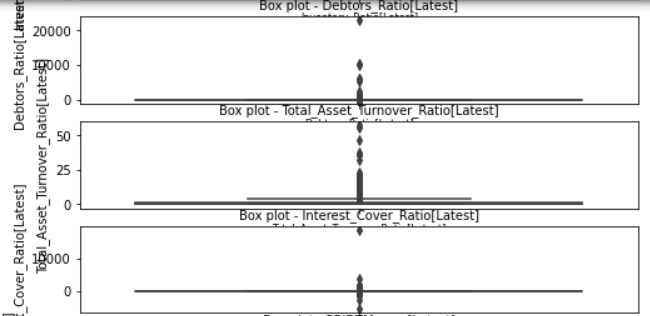
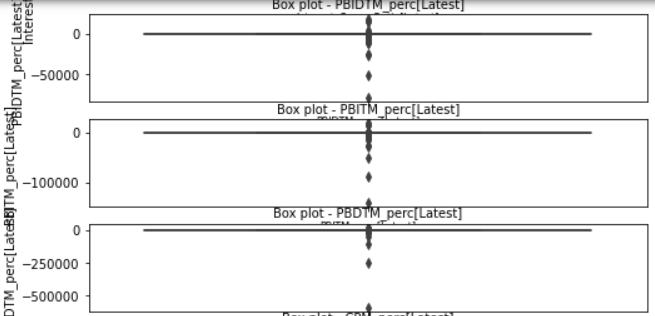
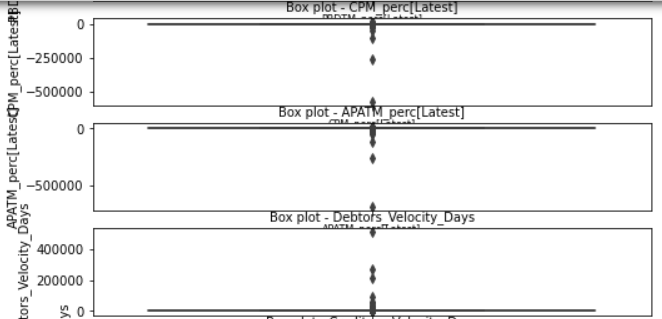
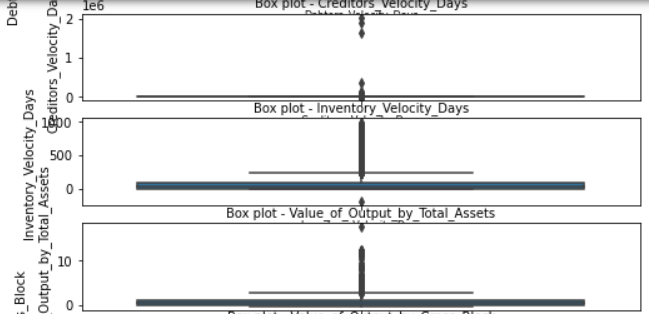
**Checking Outliers in the data per column:**

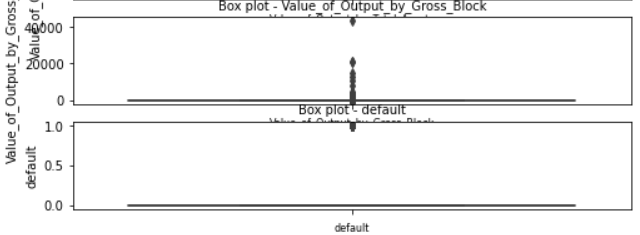
We will create boxplots for each variable and check for the presence of outliers in the data:

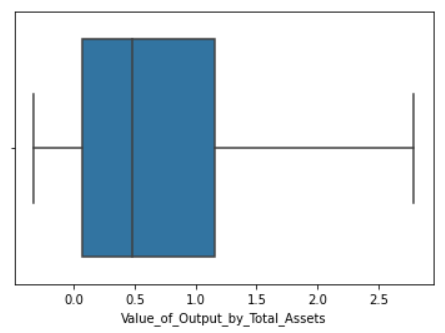
  



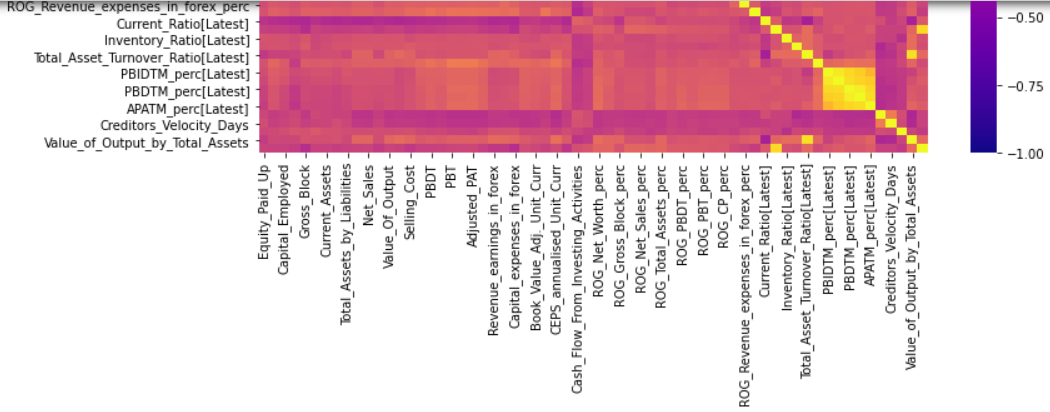
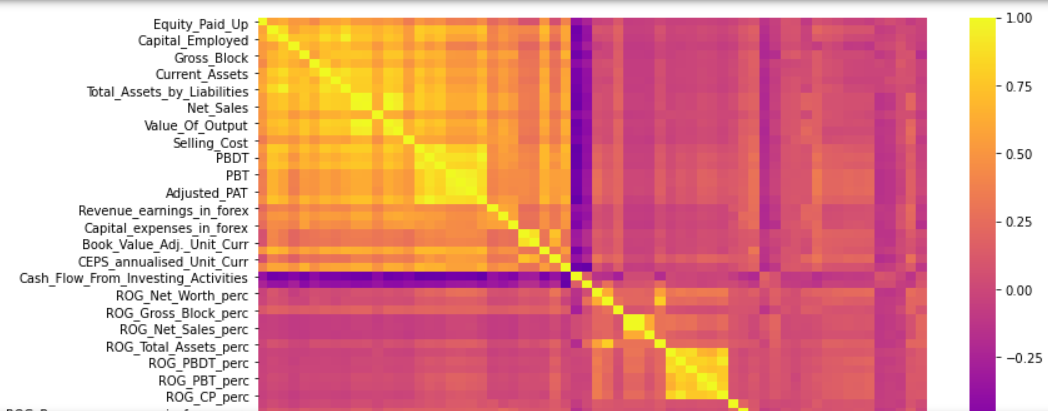
It can be seen that almost all the variables have outliers present in them We shall go ahead and treat the same using the capping method that corresponds to 1.5 times the interquartile length from the first and the third quartiles.

Then we make sure that the outliers have been capped by illustrating a boxplot for a variable where outliers were observed previously: Value\_of\_Output\_by\_Total\_Assets



**Correlations between independant variables:**

A heatmap between the independent variables is shown as follows:



* High positive correlations are observed between a major portion of the dependant variables: Equity\_Paid\_Up, Capital\_Employed, Gross\_Block, Current\_Assets, Total\_Assets\_by\_Liabilities, Net\_Sales, Value\_Of\_Output, Selling\_Cost, PBDT, PBT, Adjusted\_PAT, Revenue\_earnings\_in\_forex, Capital\_expenses\_in\_forex, Book\_Value\_Adj.\_Unit\_Curr, CEPS\_annualised\_Unit\_Curr.
* High negative correlations are observed between the variable: Cash\_Flow\_From\_Investing\_Activities and the variables specified above.

**Splitting the data:**

The data is then split using the train\_test\_split function by Scikit learn using the following parameters:

Train data size = 67%

Test data size = 33%

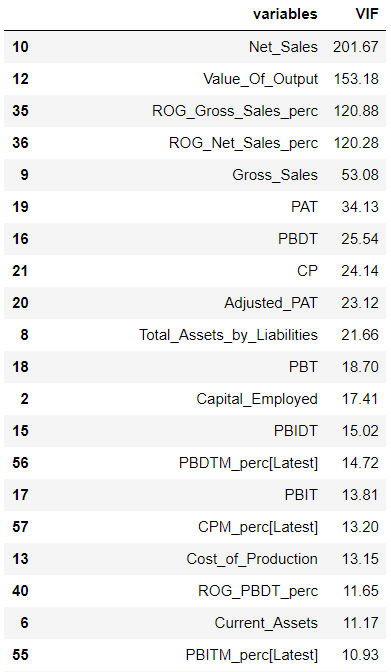
Random state = 42

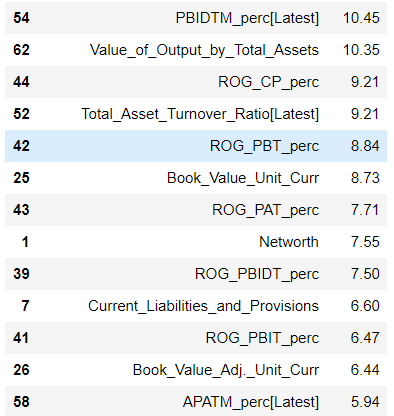
We also make use of the parameter, ‘stratify’, because there is imbalance in the data for the target variable: default. Just to recollect, we only had 11% of positive cases in the target variable where the default actually occurred. Well, from a business standpoint, having a low number of positives in the target variable is actually a good sign.

**Calculating the Variance Inflation Factor to determine the variables used for prediction:**

Variance Inflation Factor (VIF) is used to determine the degree of multicollinearity between the dependant variables and the outcome variable. We are to use variables that have a low multicollinearity with the target variable so that their predicting power can be increased.

We sort the independent variables in the descending order of VIF:





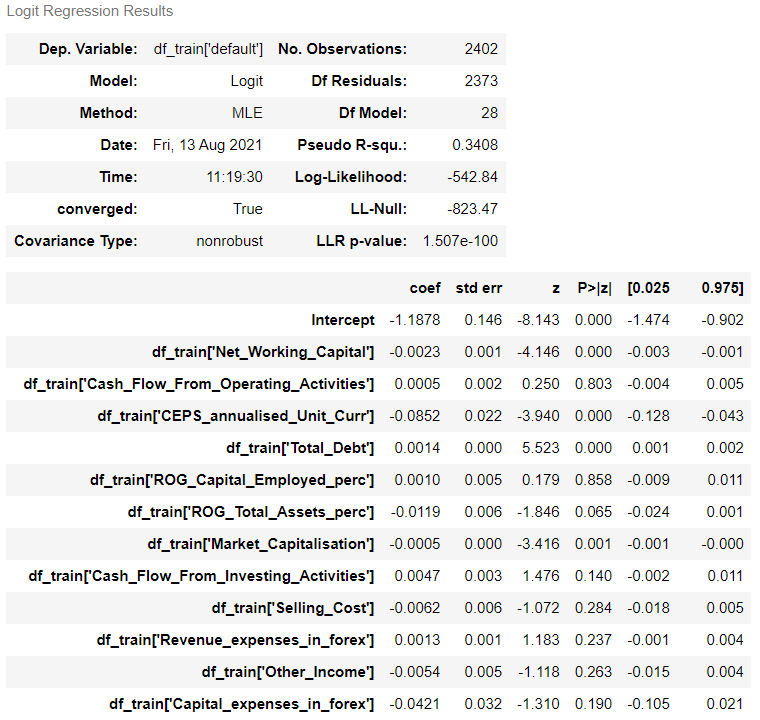


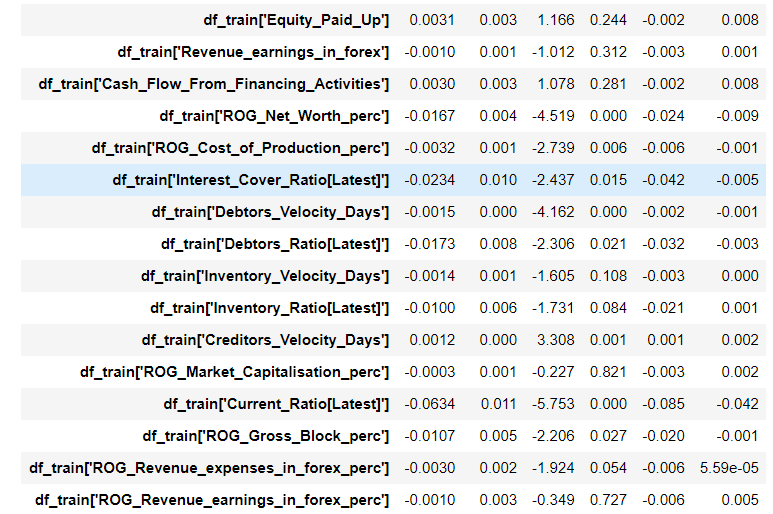


* Very high VIF values are observed in some variables which just reaffirms our findings earlier using the heatmap.
* For our model building, we shall choose the variables that have a VIF of less than 5.

**Logistic Regression Modelling with Statsmodel library:**

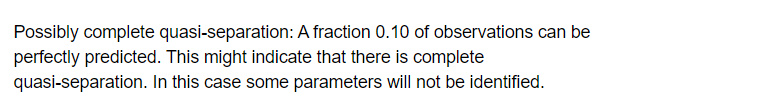
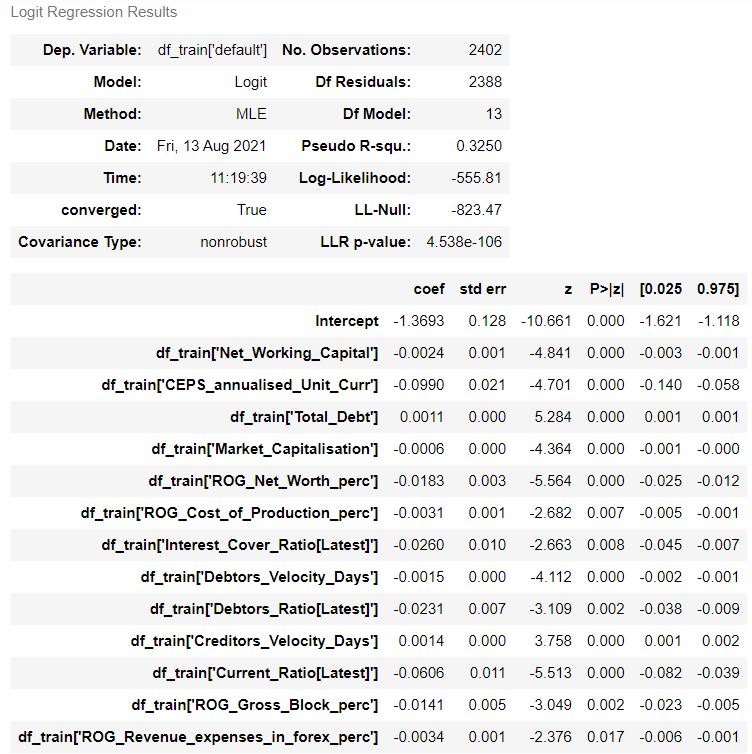
We shall import the logistic regression model and use all the variables with a VIF of less than 5 to execute our model:





We select the variables that have a ‘p value’ of less than 0.05 and rebuild the model.

**Model 2:**



We now have all the variables that have a p value of less than 0.05. This means that all the variables used in model\_2 is useful in predicting the target variable: default.

The pseudo R-squared is 0.32 and the adjusted psuedo R-squared is 0.30. The values are pretty close and this indicate lesser insignificant predicting variables in our model.

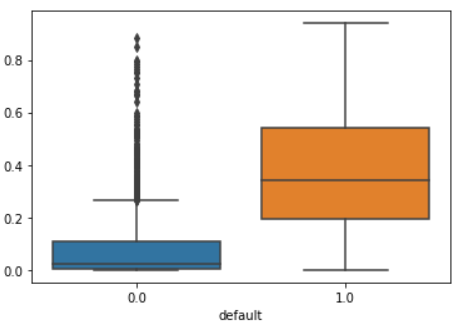
Let us also go ahead and check the Variance Inflation Factor for these predicting variables and thereby get a idea on the multicollinearity:



We see that the VIF values are all low and less than our threshold of 5. So, we will go ahead and use this model. We then fit the model on the train data:

**Validating the model on the train and choosing the optimum cut-off:**

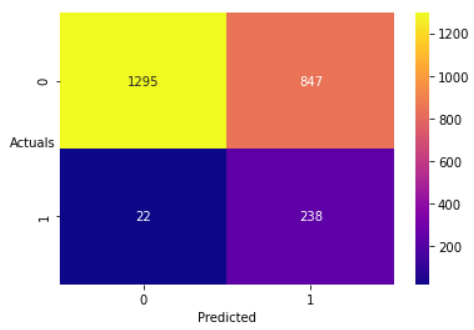
Here is a box-plot to show the distribution of the predicted values with respect to the target variable:

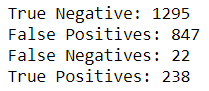


We shall now set a cut-off value for the predicted values on the train set:

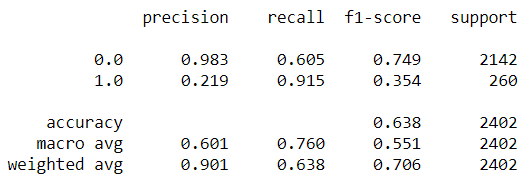
**Cut-off of 0.05:**

**Confusion matrix:**





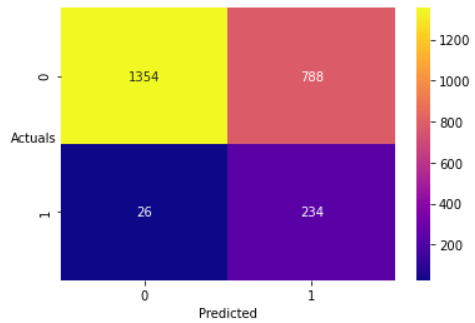
**Classification Report:**

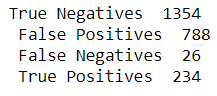


* The accuracy, precision and f1-score are very low.
* The recall is very good. This means that our model for this cut-off is able to accurately identify the actual defaulters(True Positives) from the total list of defaulters (True Positives + False Negatives).
* This is extremely important in our business case. Having a high False Positives will not have any adverse impact on the business as these customers would not end up defaulting anyway. But however, there is a cost involved where those customers would be wrongfully charged a high interest rate because our model would expect them to default but they would not.

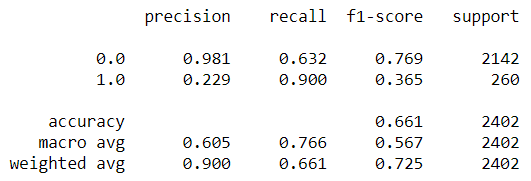
**Cut-off of 0.06:**

**Confusion matrix:**





**Confusion Matrix:**

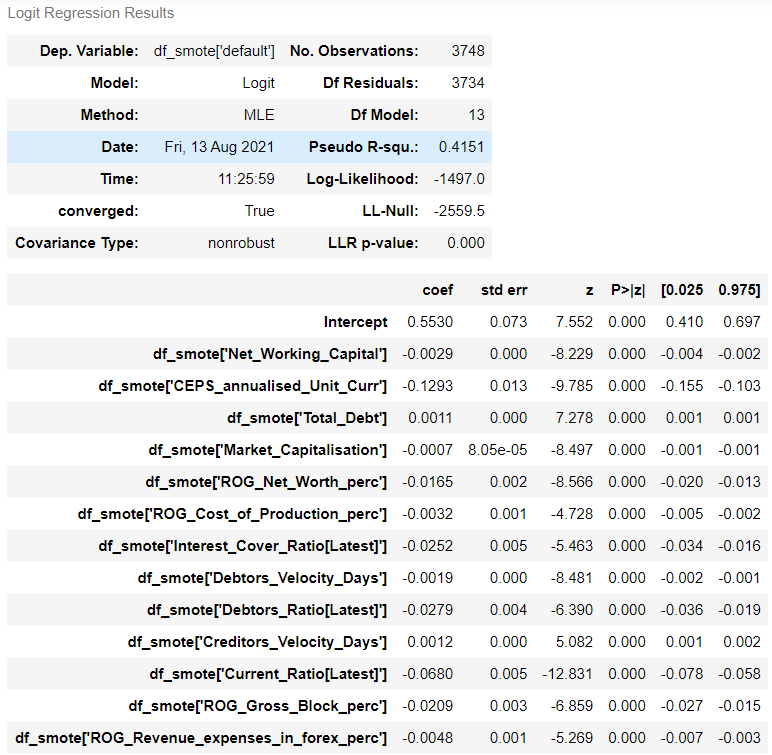


The recall values have actually decreased.

We shall now treat the Target imbalance in our train data using Synthetic Minority Oversampling Technique.

**Model 3:**

We shall use the same variables as in our previous model iteration as those variables proved to be important.



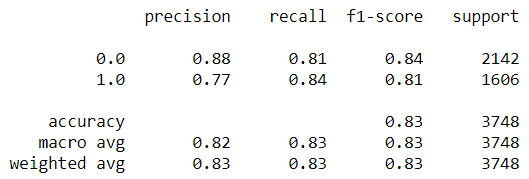
#### **Cut-off of 0.5:**

We shall predict the variable of interest i.e. default using a cut-off of 0.5

**Confusion Matrix:**



**Classification Report:**



* The recall is 84% and the accuracy is 83%. This is a good model as the metric of interest in our business case: Recall is good and the precision is not extremely low as in the previous model iteration.

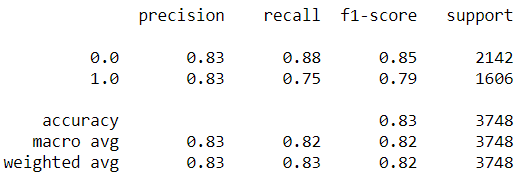
#### **Cut-off of 0.6:**

We shall hope to get better scores using the cut-off as 0.6

**Confusion Matrix:**



**Classification Report:**



* The precision has improved but the recall has decresed.
* The model accuracy has remained the same.
* We shall consider the cut-off of 0.5 as the best model considering the metric of interest: Recall.

**Conclusion:**

* We have not been successful in validating the model on the test data.
* The important variables in predicting the target: default are the ones that are used in the model 3 iteration.