FRA PROJECT (MILESTONE-1)

BUSINESS REPORT

SULOCHANA

Table of contents:

* 1. **Outlier Treatment……………………………………………... 14**
  2. **Missing Value Treatment………………………………………17**
  3. **Transform Target variable into 0 and 1……………………..21**
  4. **Univariate (4 marks) & Bivariate (6marks) analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building)…………………………………………………………..22**
  5. **Train Test Split…………………………………………………**
  6. **Build Logistic Regression Model (using stats model library) on most important variables on Train Dataset**

**And choose the optimum cutoff. Also showcase your model building approach…………………………………………………………..42**

* 1. **Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model………………………………………………………………..46.**

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 Networth of the company in the following year (2016) is provided which can be used to drive the labeled field.

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

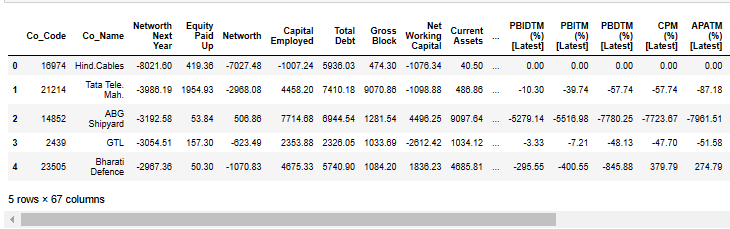
Dependent variable - We need to create a default variable that should take the value of 1 when net worth next year is negative & 0 when net worth next year is positive.

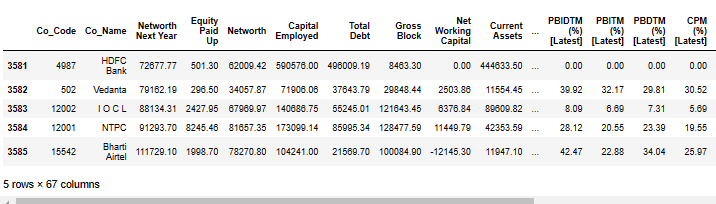
**DATA DICTIONARY:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Field Name** | **Description** | **New Field Name** |  |
| 1 | Co\_Code | Company Code | Co\_Code |  |
| 2 | Co\_Name | Company Name | Co\_Name |  |
| 3 | Networth Next Year | Value of a company as on 2016 - Next Year(difference between the value of total assets and total liabilities) | Networth\_Next\_Year |  |
| 4 | Equity Paid Up | Amount that has been received by the company through the issue of shares to the shareholders | Equity\_Paid\_Up |  |
| 5 | Networth | Value of a company as on 2015 - Current Year | Networth |  |
| 6 | Capital Employed | Total amount of capital used for the acquisition of profits by a company | Capital\_Employed |  |
| 7 | Total Debt | The sum of money borrowed by the company and is due to be paid | Total\_Debt |  |
| 8 | Gross Block | Total value of all of the assets that a company owns | Gross\_Block |  |
| 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). | Net\_Working\_Capital |  |
| 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. | Curr\_Assets |  |
| 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) | Curr\_Liab\_and\_Prov |  |
| 12 | Total Assets/Liabilities | Ratio of total assets to liabailities of the company | Total\_Assets\_to\_Liab |  |
| 13 | Gross Sales | The grand total of sale transactions within the accounting period | Gross\_Sales |  |
| 14 | Net Sales | Gross sales minus returns, allowances, and discounts | Net\_Sales |  |
| 15 | Other Income | Income realized from non-business activities (e.g. sale of long term asset) | Other\_Income |  |
| 16 | Value Of Output | Product of physical output of goods and services produced by company and its market price | Value\_Of\_Output |  |
| 17 | Cost of Production | Costs incurred by a business from manufacturing a product or providing a service | Cost\_of\_Prod |  |
| 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) | Selling\_Cost |  |
| 19 | PBIDT | Profit Before Interest, Depreciation & Taxes | PBIDT |  |
| 20 | PBDT | Profit Before Depreciation and Tax | PBDT |  |
| 21 | PBIT | Profit before interest and taxes | PBIT |  |
| 22 | PBT | Profit before tax | PBT |  |
| 23 | PAT | Profit After Tax | PAT |  |
| 24 | Adjusted PAT | Adjusted profit is the best estimate of the true profit | Adjusted\_PAT |  |
| 26 | CP | Commercial paper , a short-term debt instrument to meet short-term liabilities. | CP |  |
| 27 | Revenue earnings in forex | Revenue earned in foreign currency | Rev\_earn\_in\_forex |  |
| 28 | Revenue expenses in forex | Expenses due to foreign currency transactions | Rev\_exp\_in\_forex |  |
| 29 | Capital expenses in forex | Long term investment in forex | Capital\_exp\_in\_forex |  |
| 30 | Book Value (Unit Curr) | Net asset value | Book\_Value\_Unit\_Curr |  |
| 31 | Book Value (Adj.) (Unit Curr) | Book value adjusted to reflect asset's true fair market value | Book\_Value\_Adj\_Unit\_Curr |  |
| 32 | Market Capitalisation | Product of the total number of a company's outstanding shares and the current market price of one share | Market\_Capitalisation |  |
| 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 | CEPS\_annualised\_Unit\_Curr |  |
| 34 | Cash Flow From Operating Activities | Use of cash from ongoing regular business activities | Cash\_Flow\_From\_Opr |  |
| 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 | Cash\_Flow\_From\_Inv |  |
| 36 | Cash Flow From Financing Activities | Net flows of cash that are used to fund the company (transactions involving debt, equity, and dividends) | Cash\_Flow\_From\_Fin |  |
| 37 | ROG-Net Worth (%) | Rate of Growth - Networth | ROG\_Net\_Worth\_perc |  |
| 38 | ROG-Capital Employed (%) | Rate of Growth - Capital Employed | ROG\_Capital\_Employed\_perc |  |
| 39 | ROG-Gross Block (%) | Rate of Growth - Gross Block | ROG\_Gross\_Block\_perc |  |
| 40 | ROG-Gross Sales (%) | Rate of Growth - Gross Sales | ROG\_Gross\_Sales\_perc |  |
| 41 | ROG-Net Sales (%) | Rate of Growth - Net Sales | ROG\_Net\_Sales\_perc |  |
| 42 | ROG-Cost of Production (%) | Rate of Growth - Cost of Production | ROG\_Cost\_of\_Prod\_perc |  |
| 43 | ROG-Total Assets (%) | Rate of Growth - Total Assets | ROG\_Total\_Assets\_perc |  |
| 44 | ROG-PBIDT (%) | Rate of Growth- PBIDT | ROG\_PBIDT\_perc |  |
| 45 | ROG-PBDT (%) | Rate of Growth- PBDT | ROG\_PBDT\_perc |  |
| 46 | ROG-PBIT (%) | Rate of Growth- PBIT | ROG\_PBIT\_perc |  |
| 47 | ROG-PBT (%) | Rate of Growth- PBT | ROG\_PBT\_perc |  |
| 48 | ROG-PAT (%) | Rate of Growth- PAT | ROG\_PAT\_perc |  |
| 49 | ROG-CP (%) | Rate of Growth- CP | ROG\_CP\_perc |  |
| 50 | ROG-Revenue earnings in forex (%) | Rate of Growth - Revenue earnings in forex | ROG\_Rev\_earn\_in\_forex\_perc |  |
| 51 | ROG-Revenue expenses in forex (%) | Rate of Growth - Revenue expenses in forex | ROG\_Rev\_exp\_in\_forex\_perc |  |
| 52 | ROG-Market Capitalisation (%) | Rate of Growth - Market Capitalisation | ROG\_Market\_Capitalisation\_perc |  |
| 53 | Current Ratio[Latest] | Liquidity ratio, company's ability to pay short-term obligations or those due within one year | Curr\_Ratio\_Latest |  |
| 54 | Fixed Assets Ratio[Latest] | Solvency ratio, the capacity of a company to discharge its obligations towards long-term lenders indicating | Fixed\_Assets\_Ratio\_Latest |  |
| 55 | Inventory Ratio[Latest] | Activity ratio, specifies the number of times the stock or inventory has been replaced and sold by the company | Inventory\_Ratio\_Latest |  |
| 56 | Debtors Ratio[Latest] | Measures how quickly cash debtors are paying back to the company | Debtors\_Ratio\_Latest |  |
| 57 | Total Asset Turnover Ratio[Latest] | The value of a company's revenues relative to the value of its assets | Total\_Asset\_Turnover\_Ratio\_Latest |  |
| 58 | Interest Cover Ratio[Latest] | Determines how easily a company can pay interest on its outstanding debt | Interest\_Cover\_Ratio\_Latest |  |
| 59 | PBIDTM (%)[Latest] | Profit before Interest Depreciation and Tax Margin | PBIDTM\_perc\_Latest |  |
| 60 | PBITM (%)[Latest] | Profit Before Interest Tax Margin | PBITM\_perc\_Latest |  |
| 61 | PBDTM (%)[Latest] | Profit Before Depreciation Tax Margin | PBDTM\_perc\_Latest |  |
| 62 | CPM (%)[Latest] | Cost per thousand (advertising cost) | CPM\_perc\_Latest |  |
| 63 | APATM (%)[Latest] | After tax profit margin | APATM\_perc\_Latest |  |
| 64 | Debtors Velocity (Days) | Average days required for receiving the payments | Debtors\_Vel\_Days |  |
| 65 | Creditors Velocity (Days) | Average number of days company takes to pay suppliers | Creditors\_Vel\_Days |  |
| 66 | Inventory Velocity (Days) | Average number of days the company needs to turn its inventory into sales | Inventory\_Vel\_Days |  |
| 67 | Value of Output/Total Assets | Ratio of Value of Output (market value) to Total Assets | Value\_of\_Output\_to\_Total\_Assets |  |
| 68 | Value of Output/Gross Block | Ratio of Value of Output (market value) to Gross Block | Value\_of\_Output\_to\_Gross\_Block |  |

**Load the data:**

**Glimpse of Data:**





**Shape:**

The number of rows (observations) is 3586

The number of columns (variables) is 67

## Fixing messy column names (containing spaces) for ease of use:

## After fixing column names:

## 

## Checking data type of all columns:

## 

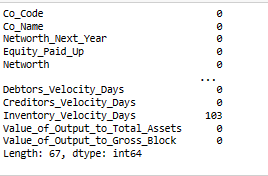
## 

## Descriptive statistics for the continuous variables:

## 

## 

# Checking missing values:



Number of missing values in the dataset is 118.

## Checking duplicates in the data:

Number of duplicates present in the dataset are 0

**Data types:**

float64 63

int64 3

Object 1

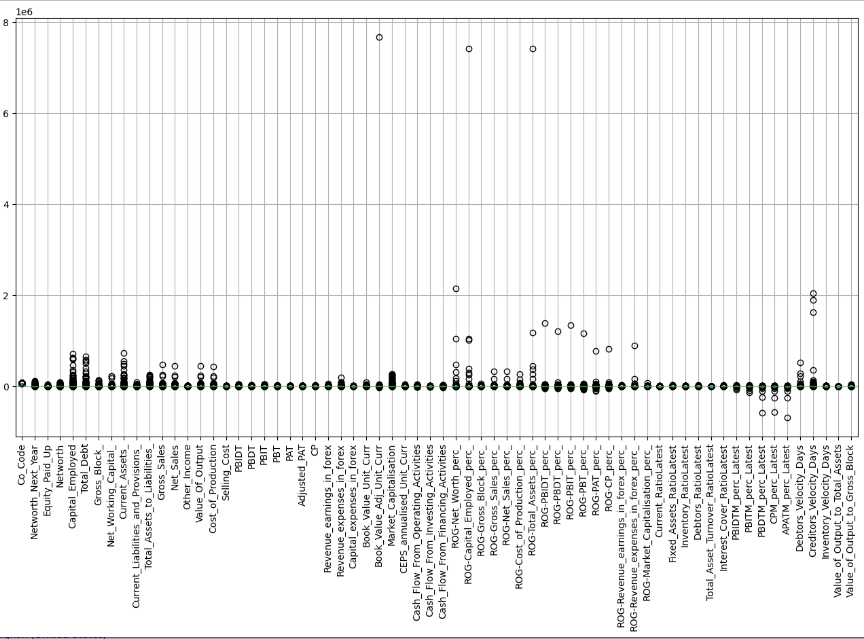
dtype: int64

**Data insights:**

* Data consists of both categorical and numeric variables.
* There are total of 3586 observations and 67 columns in the dataset. Out of 67, 66 columns are of integer type, 1 column is of object type.
* Data contains 118 missing values.
* Data does not contain any duplicated values.
* Column networth\_next\_year can be used to drive the labeled field of the company in the following year. Hence, we will create a default variable.
* This is a target variable and all other are predector variables.
  1. **Outlier Treatment**

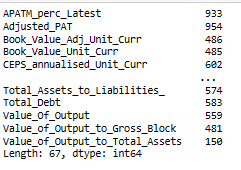
Creating a list of column names that are numeric:

Box plot for continuous variables



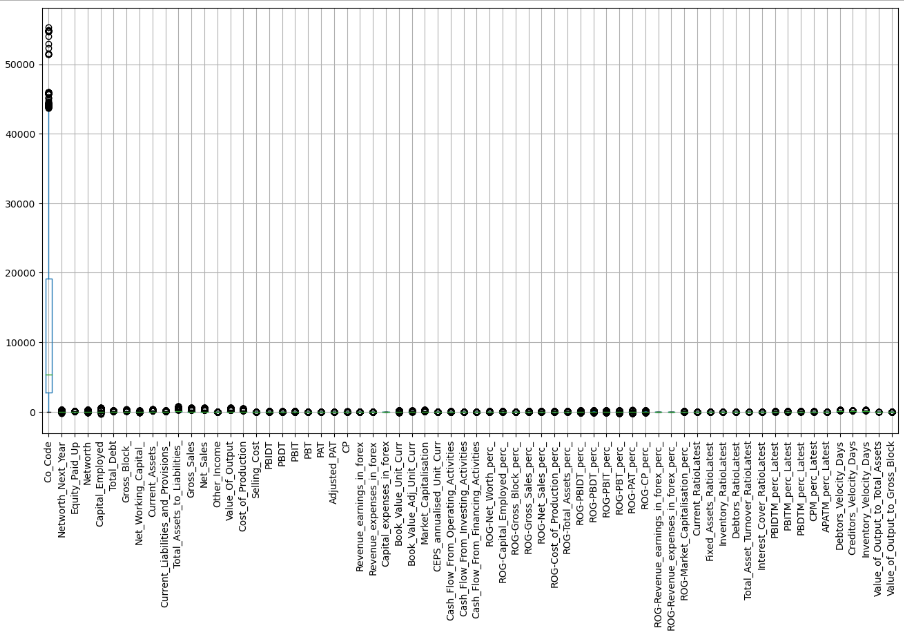
There are outliers present in the data set. Almost all the columns have outliers. We should treat outliers for further analysis.

#### Let's check the number of outliers per column:



### Replacing outliers to Nan values

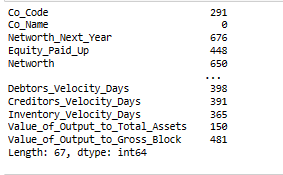
**Box plot after replacing with Nan values:**



Data has very few missing values and roughly 1.5 % of data has outliers.

Here, we are converting outliers to missing values.

* 1. **Missing Value Treatment**

****

Number of missing values after replacing outliers with Nan values is 42440

**Data size:**

240262

42440/240262

Total percentage of missing values 0.17

## Let's visually inspect the missing values in our data

## 

## We can see presence of missing values in some variables can be observed. Blue color in the heat map is indicating occupied cells while red color indicates missing values present in the dataset.

## If missing data in columns is less than 30% of our data and at row level data is atleast at 90% complete, we do not drop the data. Here, we will first check completeness of data and then decide the treatment to be used.

### In order to check the completeness of data at row, we will set axis as 1.

### We should inspect total missing values by each row.

## 

### Let's filter the data which is 90% or more complete at the row level

## 

## We are sorting proportion of missing values by dividing number of missing values by number of applicable rows. We will eliminate anything that is more than 30%.so ROG Revenue expenses in forex\_perc and ROG Revenue earnings in forex perc are the only to values which are more than 30%. Therefore, we can eliminate these values.

### Dropping columns with more than 30% missing values:

ROG-Revenue\_expenses\_in\_forex\_perc\_

ROG-Revenue\_earnings\_in\_forex\_perc\_

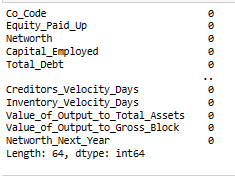
The numbers of rows after dropping columns with more than 30% missing is 3586

The numbers of columns after dropping columns with more than 30% missing is 64

### Imputing the remaining missing values:

The missing values are of numeric in nature. Hence, can be imputed using KNN imputer function from the impute module of the sklearn. This imputer utilizes the K-nearest neighbor’s method to replace the missing values in the dataset by finding the nearest neighbors with the eluclidean distance matrix.

And also it requires us to normalize our data. We need to split the data into response and predictor before applying scale.



We can see above output Missing values have been treated.

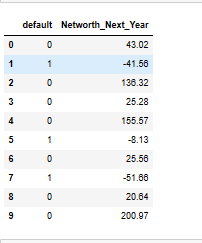
* 1. **Transform Target variable into 0 and 1**

We will now create a default variable that should take the below mentioned values:

1 is when net worth next year is negative and 0 is when net worth next year is positive

### Creating a binary target variable using 'Networth\_Next\_Year':

### Checking top 10 rows:



### Checking proportion of default:

0 3241

1 345

Name: default, dtype: int64

0 0.90

1 0.10

Name: default, dtype: float64

Approximately 10% of the companies from the dataset are likely to default and these are the companies in which investors should probably avoid investing in.

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

### Dropping unnecessary columns:

Net \_worth

Capital\_Employed

Gross\_Block\_

Gross\_Sales

Net\_Sales

Cost\_of\_Production

PBIDT

PBIT

PBT

PAT

CP

Revenue\_earnings\_in\_forex

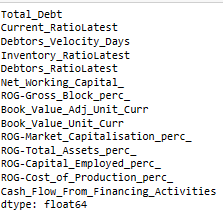
Revenue\_expenses\_in\_forex

Market\_Capitalisation

**Data shape after dropping columns:**

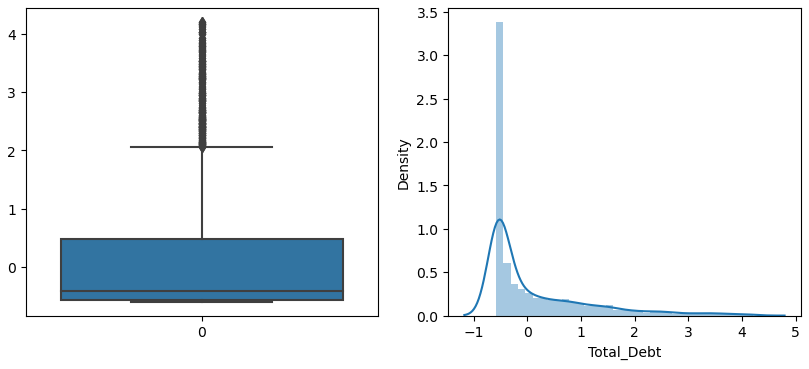
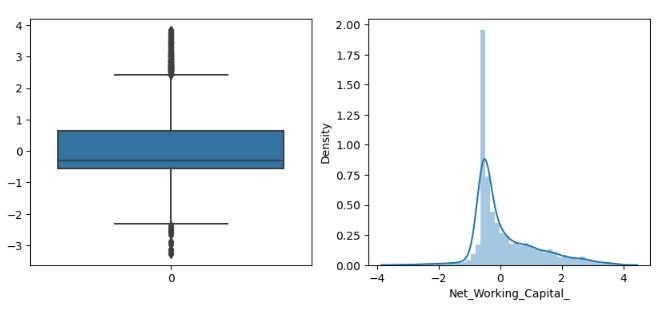
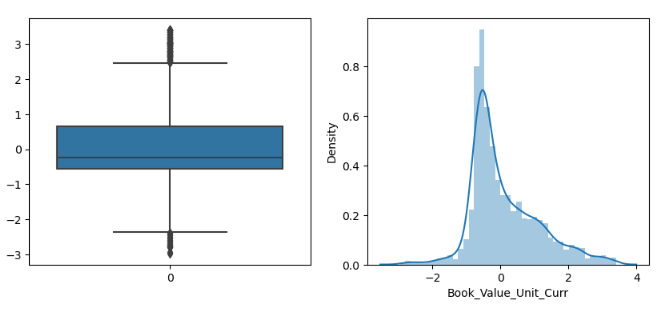
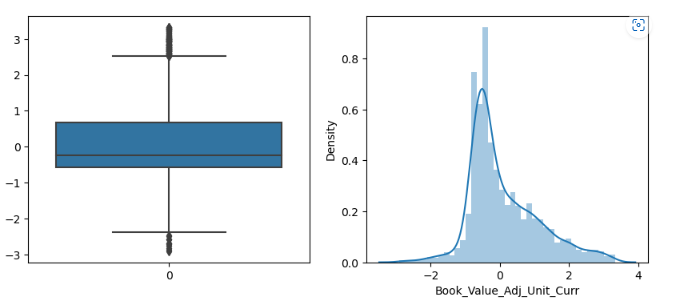
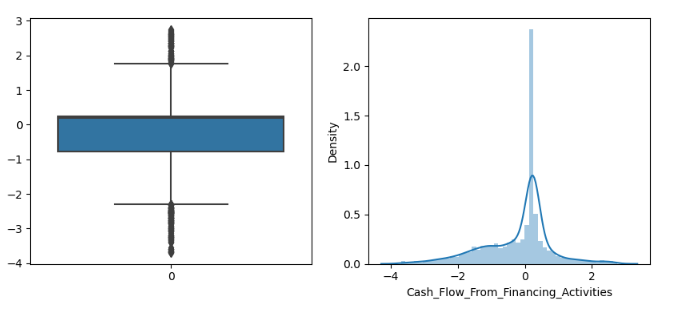
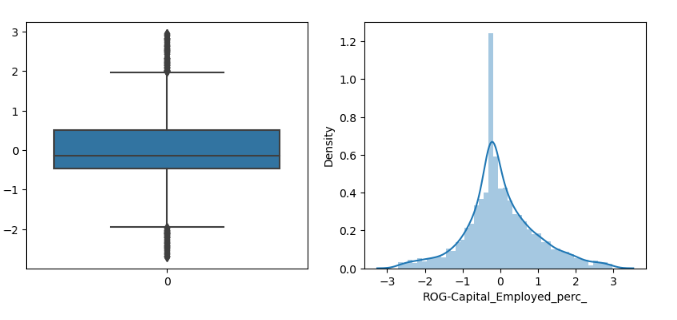
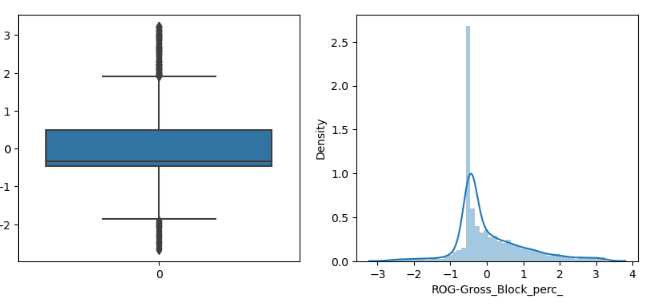
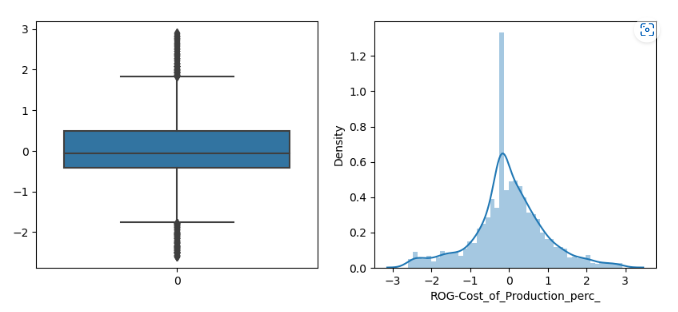
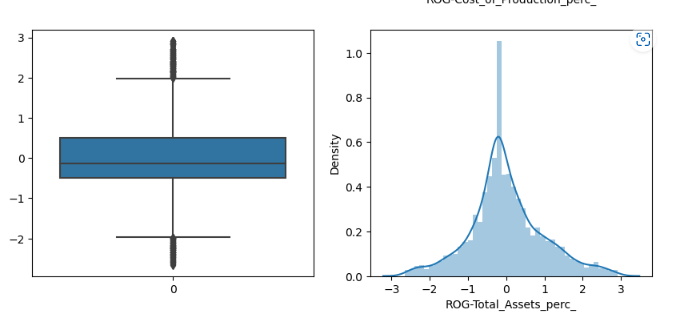
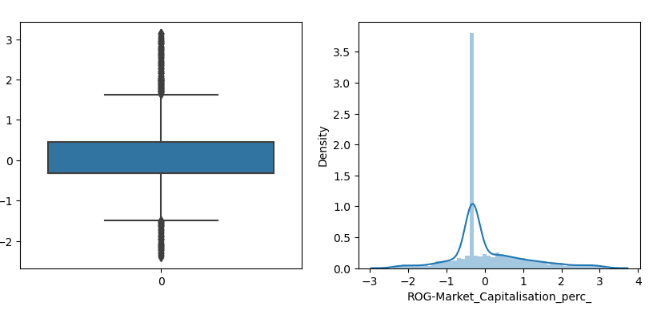
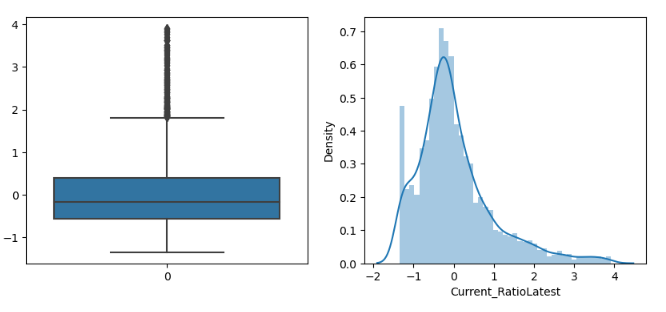
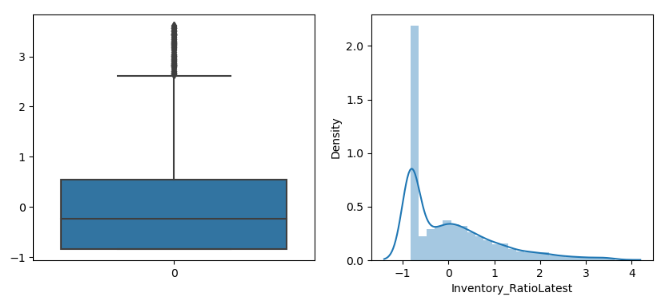
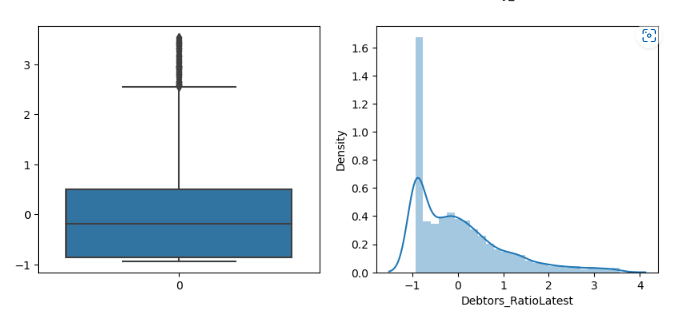
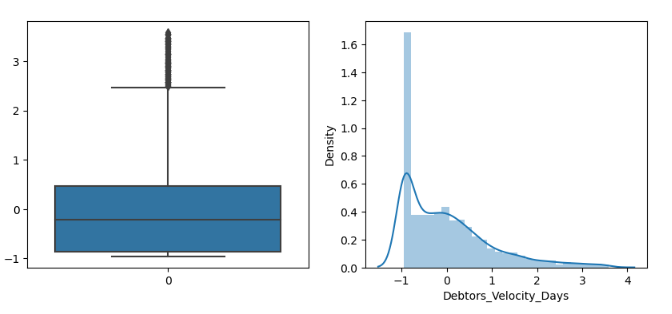
(3586, 50)

### Creating data frame with some important features which were significant for model building after performed RFE (recursive feature elimination) on dataset.



### Univariate Analysis

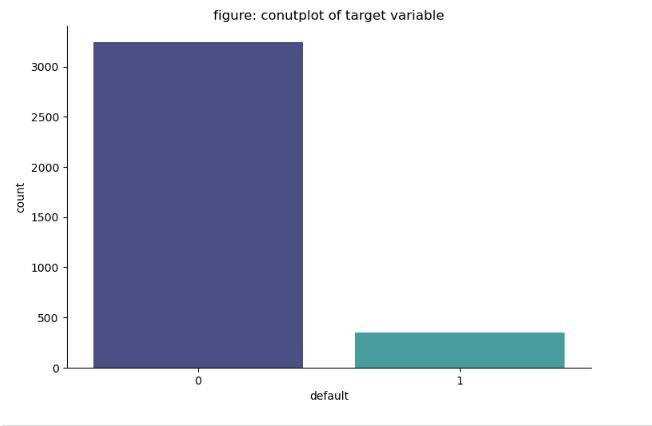
**Box plot & density plot of each important feature as a subplot**

**             **

**We can observe, even though we have treated outliers but some of the variables still indicate the presence of outliers.**

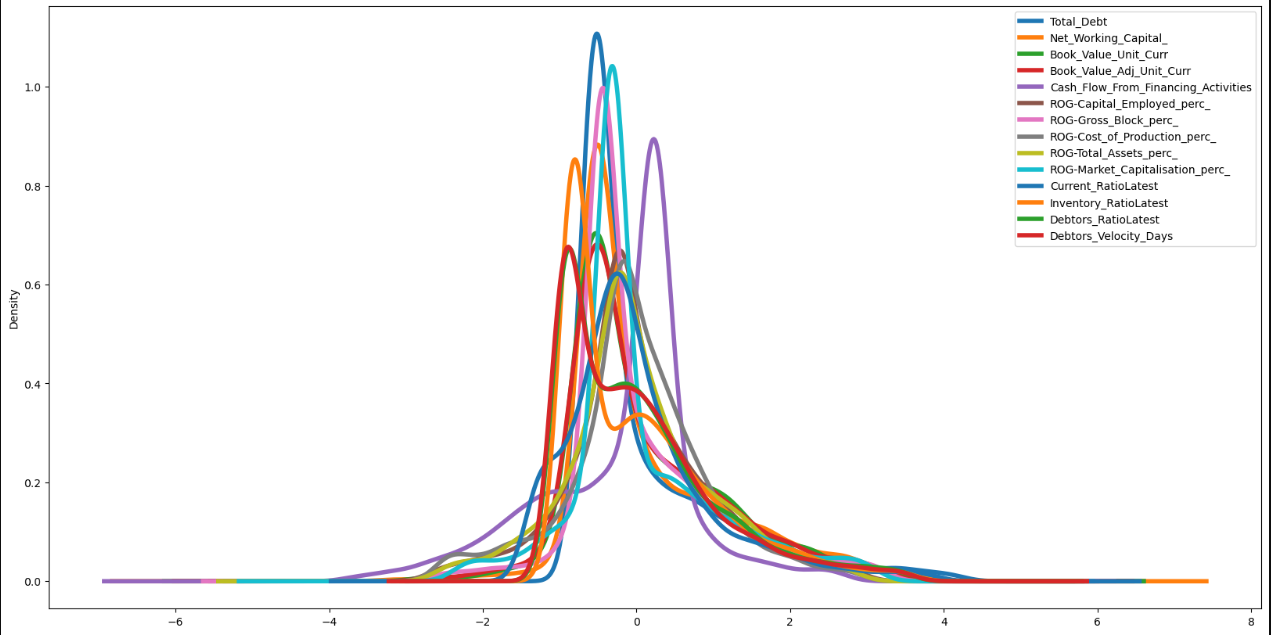
### Bi-Variate Analysis

### countplot of target variable

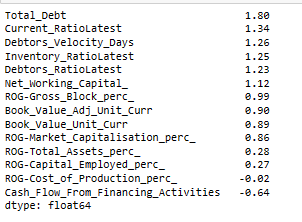
****

The data has higher Non default companies i.e., the companies which are expected to have a positive net worth year.

**Plotting multiple density plot:**

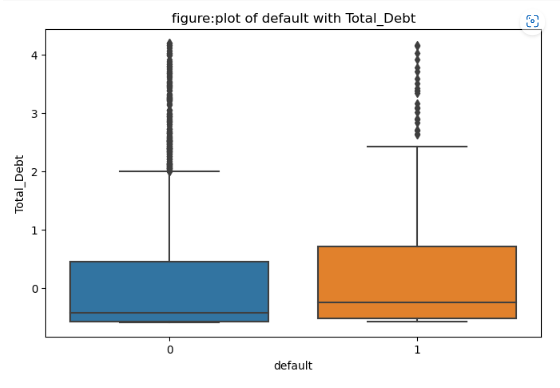
****

**skewness of data:**

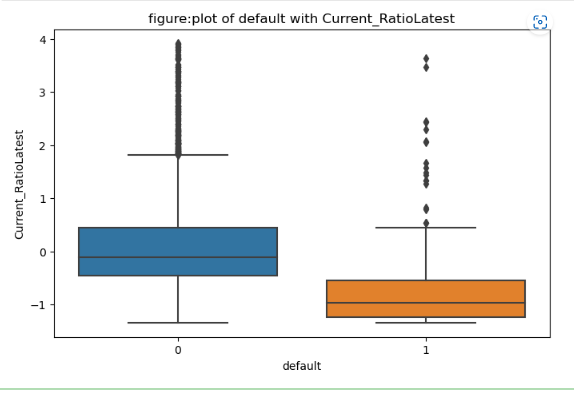
****

**Since skewness is more than 1 indicating is highly skewed for variables total\_debt, current\_ratiolatest, debtors\_velocity days, inventory\_ratiolatest, debtors\_ratiolatest, net\_working\_capital.**

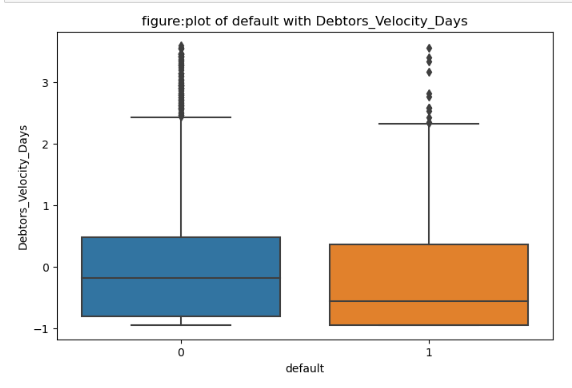
**Plot of default with Total\_Debt:**

****

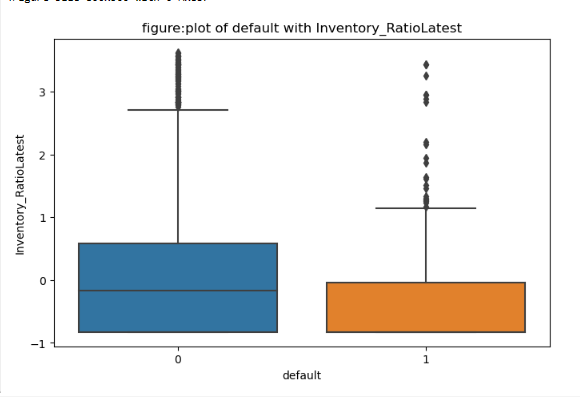
**Plot of default with Current\_RatioLatest:**

****

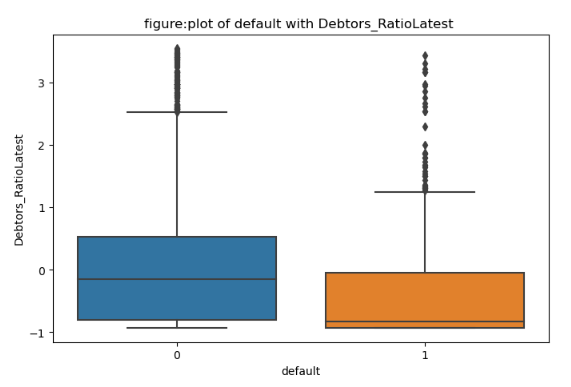
**Plot of default with Debtors\_Velocity\_Days:**

****

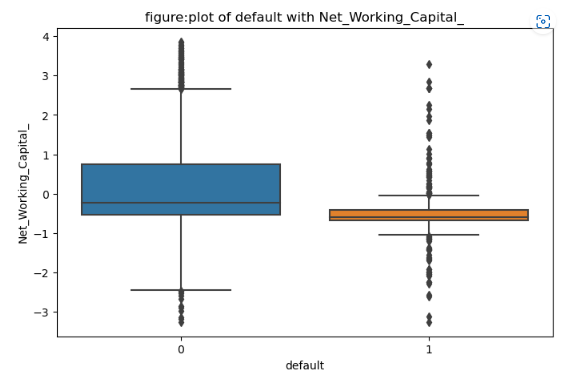
**Plot of default with Inventory\_RatioLatest:**

****

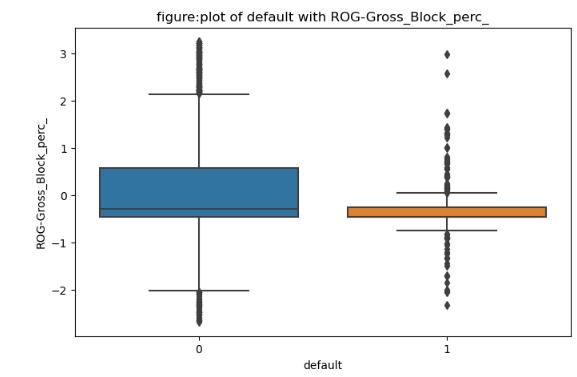
**Plot of default with Debtors\_RatioLatest:**

****

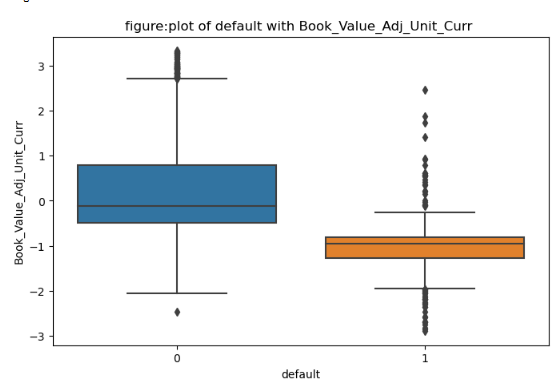
**Plot of default with Net\_Working\_Capital\_:**

****

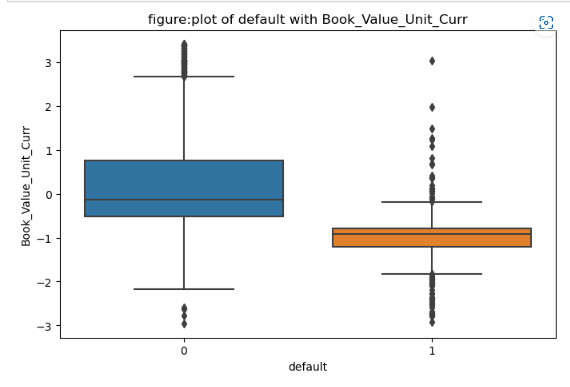
**Plot of default with ROG-Gross\_Block\_perc\_:**

****

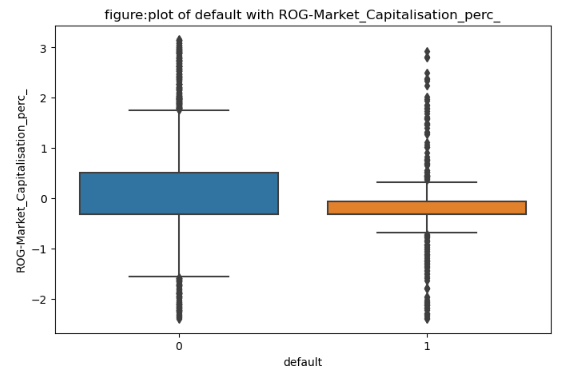
**Plot of default with Book\_Value\_Adj\_Unit\_Curr:**

****

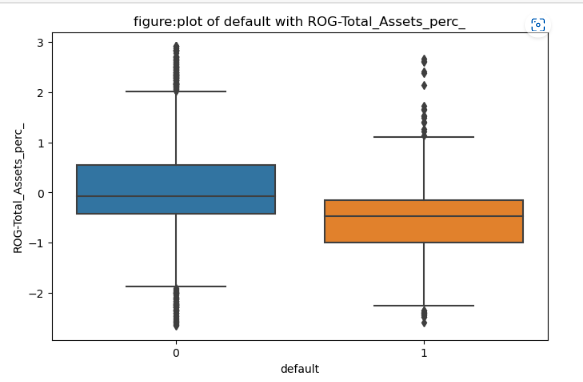
**Plot of default with Book\_Value\_Unit\_Curr:**

****

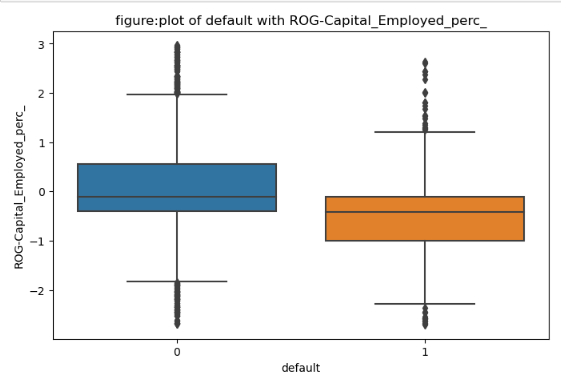
**Plot of default with ROG-Market\_Capitalisation\_perc\_:**

****

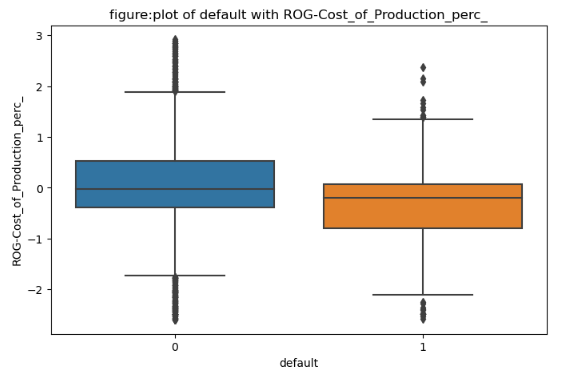
**Plot of default with ROG-Total\_Assets\_perc\_:**

****

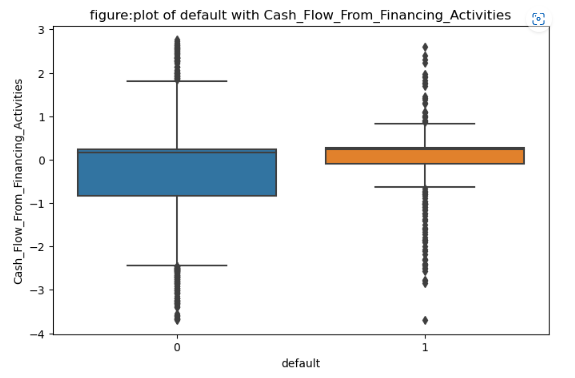
**Plot of default with ROG-Capital\_Employed\_perc\_:**

****

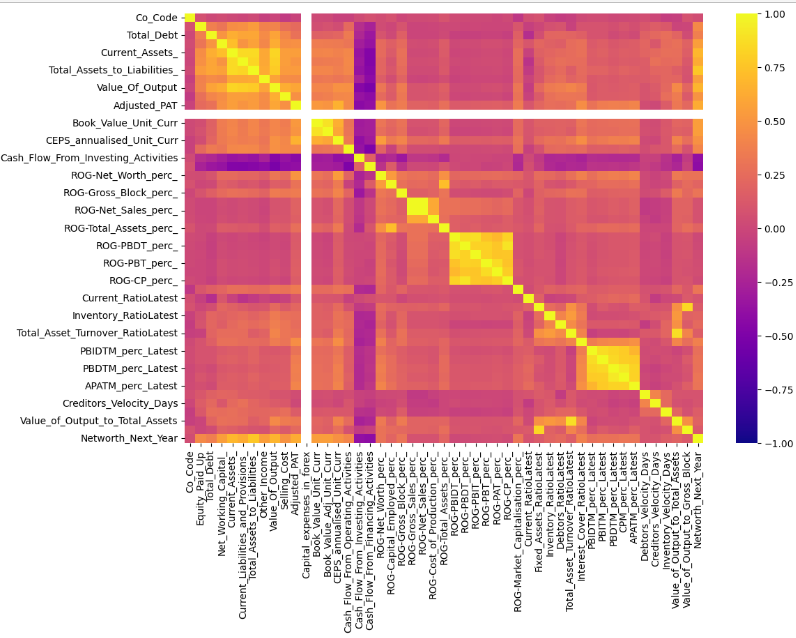
**plot of default with ROG-Cost\_of\_Production\_perc\_:**

****

**plot of default with Cash\_Flow\_From\_Financing\_Activities:**

****

### Inspect possible correlations between independent variables:



* **Using the correlation function and seaborn cluster map, we plotted correlations and obtained a better understanding of the data.**
* **Net worth and net worth next year were highly correlated.**
* **The analysis shows that there is a collinearity issue with this data set**
  1. **Train Test Split**
* Split the data into train and test dataset in a ratio of 67:33 with the fixed random\_state as 42 to ensure uniformity across multiple systems and stratify on default to make sure both the train and test data have similar proportion of defaulters and non- defaulters.
* Before we do the train test split, we will first separate independent and dependent variables using train test split from sklearn .model selection.

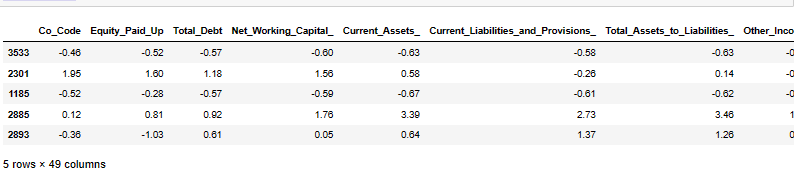
Number of rows and columns of the training set for the independent variables: (2402, 49)

Number of rows and columns of the training set for the dependent variable: (2402, 1)

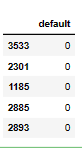
Number of rows and columns of the test set for the independent variables: (1184, 49)

Number of rows and columns of the test set for the dependent variable: (1184, 1)

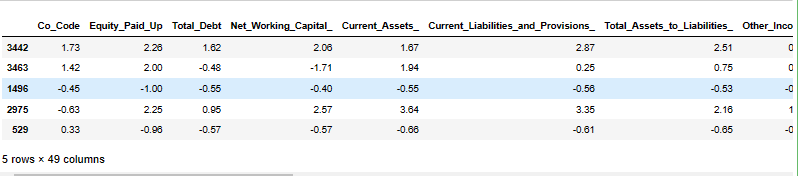
**X\_train.head**

****

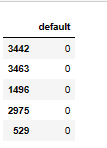
**Y\_train.head**

****

**X\_test.head**

****

**Y\_test.head:**

****

* 1. **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**

Here, we will use stats model logistic regression.

### For modeling we will use Logistic Regression with recursive feature elimination

Splitting arrays into random train and test subsets. Model will be fitted on train set and predictions will be made on the test set.

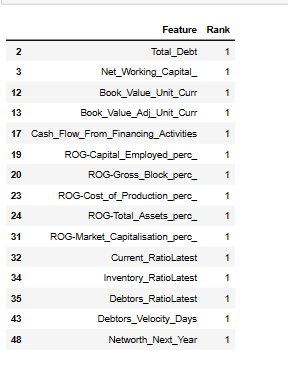
For model building, we try to approach recursive feature elimination and we want to select top 15 features that would contribute to the model well .

For modeling we will use logistic regression with recursive feature elimination.

Below are the highest contributing independent variables to the model building.

**selector.ranking**

****

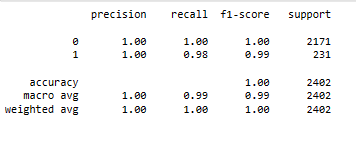
****

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

**Y\_train**



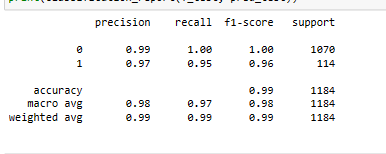
**Confusion matrix for train data**



**Y\_test:**

****

**Confusion matrix for test data**

****

We train the model and then validate the model in both the training and testing sets.

For both sets, we are plotting the confusion matrix and classification report.

The training data shows high precision and accuracy, but the recall seems lower compared to precision and accuracy.

**END**