

## TABLE OF CONTENT

|     |  |   |    |
|-----|--|---|----|
| 1.  | Introduction   | — | 3  |
| 2.  | Problem Statement  | — | 3  |
| 3.  | Project Snapshot   | — | 6  |
| 4.  | Question 1.1 Outlier Treatment                                   | — | 8  |
| 5.  | Question 1.2 Missing Value Treatment                             | — | 10 |
| 6.  | Question 1.3 Transform Target Variable to 0 and 1                | — | 12 |
| 7.  | Question 1.4 MultiVariate Analysis                               | — | 13 |
| 8.  | Question 1.5 Train Test Split                                    | — | 18 |
| 9.  | Question 1.6 Logistic Regression Models                          | — | 18 |
| 10. | Question 1.7 Performance Metrics of All Models & Interpretations | — | 21 |

### TABLES

|          |   |   |    |
|----------|---|---|----|
| Table 1. | Data description of all variables of raw data | — | 6  |
| Table 2. | Target Variable - First 5                     | — | 12 |
| Table 3. | Target Variable - Last 5                      | — | 12 |
| Table 4. | StatsModel Summary Report of Model #8         | — | 20 |
| Table 5. | All Model Performance Comparison              | — | 21 |
| Table 6. | Classification Report of Model 8              | — | 22 |

### FIGURES

|         |   |   |    |
|---------|---|---|----|
| Fig 1.  | Class Balance of Target Var                                     | — | 6  |
| Fig 2.  | Boxplot before Outlier Treatment - Top 15 predictors - Z-scaled | — | 9  |
| Fig 3.  | Boxplot after IQR Treatment - Top 15 predictors - Z-scaled      | — | 9  |
| Fig 4.  | Boxplot after Z-score Treatment - Top 15 predictors - Z-scaled  | — | 10 |
| Fig 5.  | Missing Values visualized - Raw data (55 vars)                  | — | 11 |
| Fig 6.  | Distribution of top 15 vars - Z-Scaled                          | — | 15 |
| Fig 7.  | Correlation Heatmap - 55 Variables                              | — | 16 |
| Fig 8.  | Correlation Heatmap of top 15 vars                              | — | 17 |
| Fig 9.  | Confusion Matrix of Model 8                                     | — | 22 |
| Fig 10. | Effect of Variables on Default                                  | — | 23 |

## INTRODUCTION

This report consists of Classification modeling of Company Financials using Logistic Regression. It is expected to find whether a given company is in good financial health and will it have a positive Net-worth for the next year

We have used Python for coding - [Click here for the Jupiter Notebook](#)

## 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'

### Data Description:-

| Variable           | Description   |
|--------------------|---|
| Co_Code            | Company Code  |
| Co_Name            | Company Name  |
| Networth_Next_Year | Value of a company as on 2016 - Next Year(difference between the value of total assets and total liabilities) |
| Equity_Paid_Up     | Amount that has been received by the company through the issue of shares to the shareholders                  |
| Networth           | Value of a company as on 2015 - Current Year  |
| Capital_Employed   | Total amount of capital used for the acquisition of profits by a company                                      |
| Total_Debt         | The sum of money borrowed by the company and is due to be paid  |

| Variable                 | Description  |
|--------------------------|--|
| Co_Code                  | Company Code   |
| Co_Name                  | Company Name   |
| Networth_Next_Year       | Value of a company as on 2016 - Next Year(difference between the value of total assets and total liabilities)  |
| Equity_Paid_Up           | Amount that has been received by the company through the issue of shares to the shareholders   |
| Networth                 | Value of a company as on 2015 - Current Year   |
| Capital_Employed         | Total amount of capital used for the acquisition of profits by a company   |
| Total_Debt               | The sum of money borrowed by the company and is due to be paid   |
| Gross_Block              | Total value of all of the assets that a company owns   |
| 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).                                     |
| Curr_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_Liab_and_Prov       | Short-term financial obligations that are due within one year (includes amount that is set aside cover a future liability)   |
| Total_Assets_to_Liab     | Ratio of total assets to liabilities of the company  |
| Gross_Sales              | The grand total of sale transactions within the accounting period  |
| Net_Sales                | Gross sales minus returns, allowances, and discounts   |
| Other_Income             | Income realized from non-business activities (e.g. sale of long term asset)  |
| Value_Of_Output          | Product of physical output of goods and services produced by company and its market price  |
| Cost_of_Prod             | Costs incurred by a business from manufacturing a product or providing a service   |
| 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) |
| PBIDT                    | Profit Before Interest, Depreciation & Taxes   |
| PBDT                     | Profit Before Depreciation and Tax   |
| PBIT                     | Profit before interest and taxes   |
| PBT                      | Profit before tax  |
| PAT                      | Profit After Tax   |
| Adjusted_PAT             | Adjusted profit is the best estimate of the true profit  |
| CP                       | Commercial paper, a short-term debt instrument to meet short-term liabilities.   |
| Rev_earn_in_forex        | Revenue earned in foreign currency   |
| Rev_exp_in_forex         | Expenses due to foreign currency transactions  |
| Capital_exp_in_forex     | Long term investment in forex  |
| Book_Value_Unit_Curr     | Net asset value  |
| Book_Value_Adj_Unit_Curr | Book value adjusted to reflect asset's true fair market value  |
| Market_Capitalisation    | Product of the total number of a company's outstanding shares and the current market price of one share  |

| Variable                          | Description  |
|-----------------------------------|--|
| 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 |
| Cash_Flow_From_Opr                | Use of cash from ongoing regular business activities   |
| Cash_Flow_From_Inv                | Cash used in the purchase of non-current assets—or long-term assets— that will deliver value in the future                                       |
| Cash_Flow_From_Fin                | Net flows of cash that are used to fund the company (transactions involving debt, equity, and dividends)   |
| ROG_Net_Worth_perc                | Rate of Growth - Networth  |
| ROG_Capital_Employed_perc         | Rate of Growth - Capital Employed  |
| ROG_Gross_Block_perc              | Rate of Growth - Gross Block   |
| ROG_Gross_Sales_perc              | Rate of Growth - Gross Sales   |
| ROG_Net_Sales_perc                | Rate of Growth - Net Sales   |
| ROG_Cost_of_Prod_perc             | Rate of Growth - Cost of Production  |
| ROG_Total_Assets_perc             | Rate of Growth - Total Assets  |
| ROG_PBDT_perc                     | Rate of Growth- PBDT   |
| ROG_PBDT_perc                     | Rate of Growth- PBDT   |
| ROG_PBIT_perc                     | Rate of Growth- PBIT   |
| ROG_PBT_perc                      | Rate of Growth- PBT  |
| ROG_PAT_perc                      | Rate of Growth- PAT  |
| ROG_CP_perc                       | Rate of Growth- CP   |
| ROG_Rev_earn_in_forex_perc        | Rate of Growth - Revenue earnings in forex   |
| ROG_Rev_exp_in_forex_perc         | Rate of Growth - Revenue expenses in forex   |
| ROG_Market_Capitalisation_perc    | Rate of Growth - Market Capitalisation   |
| Curr_Ratio_Latest                 | Liquidity ratio, company's ability to pay short-term obligations or those due within one year  |
| Fixed_Assets_Ratio_Latest         | Solvency ratio, the capacity of a company to discharge its obligations towards long-term lenders indicating                                      |
| Inventory_Ratio_Latest            | Activity ratio, specifies the number of times the stock or inventory has been replaced and sold by the company                                   |
| Debtors_Ratio_Latest              | Measures how quickly cash debtors are paying back to the company   |
| Total_Asset_Turnover_Ratio_Latest | The value of a company's revenues relative to the value of its assets  |
| Interest_Cover_Ratio_Latest       | Determines how easily a company can pay interest on its outstanding debt   |
| PBDTM_perc_Latest                 | Profit before Interest Depreciation and Tax Margin   |
| PBITM_perc_Latest                 | Profit Before Interest Tax Margin  |
| PBDTM_perc_Latest                 | Profit Before Depreciation Tax Margin  |
| CPM_perc_Latest                   | Cost per thousand (advertising cost)   |
| APATM_perc_Latest                 | After tax profit margin  |

| Variable                        | Description   |
|---------------------------------|---|
| Debtors_Vel_Days                | Average days required for receiving the payments                          |
| Creditors_Vel_Days              | Average number of days company takes to pay suppliers                     |
| Inventory_Vel_Days              | Average number of days the company needs to turn its inventory into sales |
| Value_of_Output_to_Total_Assets | Ratio of Value of Output (market value) to Total Assets                   |
| Value_of_Output_to_Gross_Block  | Ratio of Value of Output (market value) to Gross Block                    |

Table 1 - Data description of all variables of raw data

## PROJECT SNAPSHOT

1. Total Number of Companies (observations) = 3586
2. Total Number of Variables = 67 (1 target and 66 predictors)
3. Target Variable -
  - We create a target variable - 'default'
  - Where, if Net-worth next year is zero or positive → default = 0
  - If Net-worth next year is negative → default = 1
  - Default = 1 → 387 Companies in the data
  -

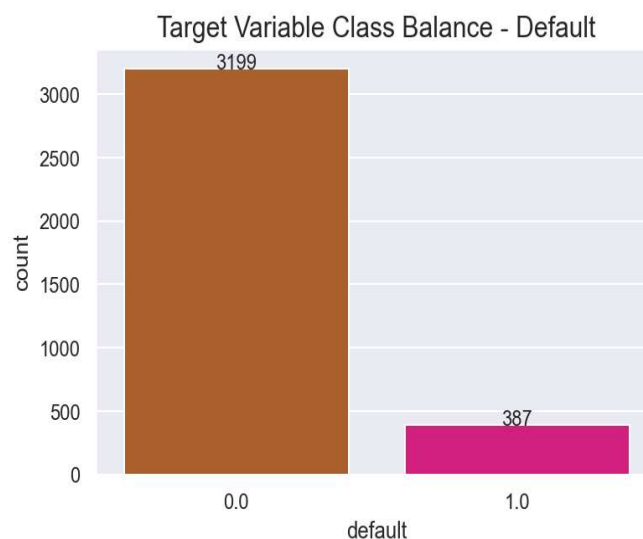


Fig 1 - Class Balance of Target Vary

4. Number of Duplicates = 0
5. Missing Value Treatment -
  - Less than 1% missing values present
  - We impute these missing values by using KNN Imputer (n\_neighbors=10)

## 6. Zero Values -

- Large amount of zero values present (total = 15.1 %)
- We drop columns with more than 30% of zero values (9 columns)
- We found that 164 out of total 387 defaulting companies had more than 5 zero values in their rows—> We conclude, more the missing or zero values, the higher is the probability of default
- For the rest of columns —> we convert zeros to Missing Nan values
- Impute all these missing values using KNN Imputer ( $n\_neighbors=10$ )

## 7. Outlier Treatment -

- IQR and Z-Score methods - used separately to identify and treat outliers
- Different Logistic Regression models fitted and tested using both
- Z-Score outlier treatment was found to give better results on Test Data

## 8. Scaling - We use Z-score Standard scaling

## 9. Multi-Collinearity -

- Many variables in the data are extracts of each other
- Hence, there is a high correlation between many of them
- This causes Multi-Collinearity and can harm a model's interpretability
- Also, these columns don't add any more value to predictions by regression
- Variance Inflation Factor method is used to check and drop columns causing Multi-Collinearity
- Recursively, one-by-one, columns with  $VIF > 5$  are dropped

## 10. Feature Engineering -

- We start with a large number of 66 predictor variables
- There are various methods employed to extract the best features
- Methods and Steps were taken for all modeling -
  - Drop unique identifiers which add no value to predictions - Company Code and name —> 64 vars left
  - Drop variables with zeros > 30% (9 cols dropped) —> 55 vars left
  - Drop vars one-by-one with  $VIF > 5$  —> 27 vars left after IQR outliers —> 23 vars left after Zscore outliers
- Also, for some models, we test by dropping insignificant variables for prediction (vars with p-values > 0.05) at 95% confidence

- For Model #7 - we use the Recursive Feature Elimination (RFE) technique to select the 15 best features for modeling

11. We choose Model #8 as the best model for deployment -

- This has the best combination of Recall and Precision for default=1

- This model -

Outlier treatment → Z-score with values capped to  $\pm 3$  std dev

RFE → with top 15 features

Oversampling method → SMOTE with 50-50 balance of 0 & 1

Choosing Optimum Threshold = 0.5

- Metrics for default = 1 →

Recall = 95%, Precision = 78%, Accuracy = 96%, f1-score = 86%

## 1.1 Outlier Treatment

- Outlier treatment is necessary for any regression model
- In Regression, outliers pull the regression line towards itself thereby affecting its slope. This distorts the reality and leads to faulty predictions
- We employ 2 types of Outlier detection and treatments in this case study
  - Inter-Quartile Range (IQR) Treatment
  - Z-score treatment
- We show box plots of 15 variables before and after the Outlier treatment. We scale these variables for better comparison
- These 15 vars are finally chosen as the best predictors for Logistic Regression
- IQR Treatment -
  - Q1 = 25th percentile, Q3 = 75th percentile
  - IQR = Q3 - Q1
  - Outlier = any value which lies beyond 1.5 times of IQR from Q1 and Q3 on either side
  - We cap all outliers to this upper or lower level

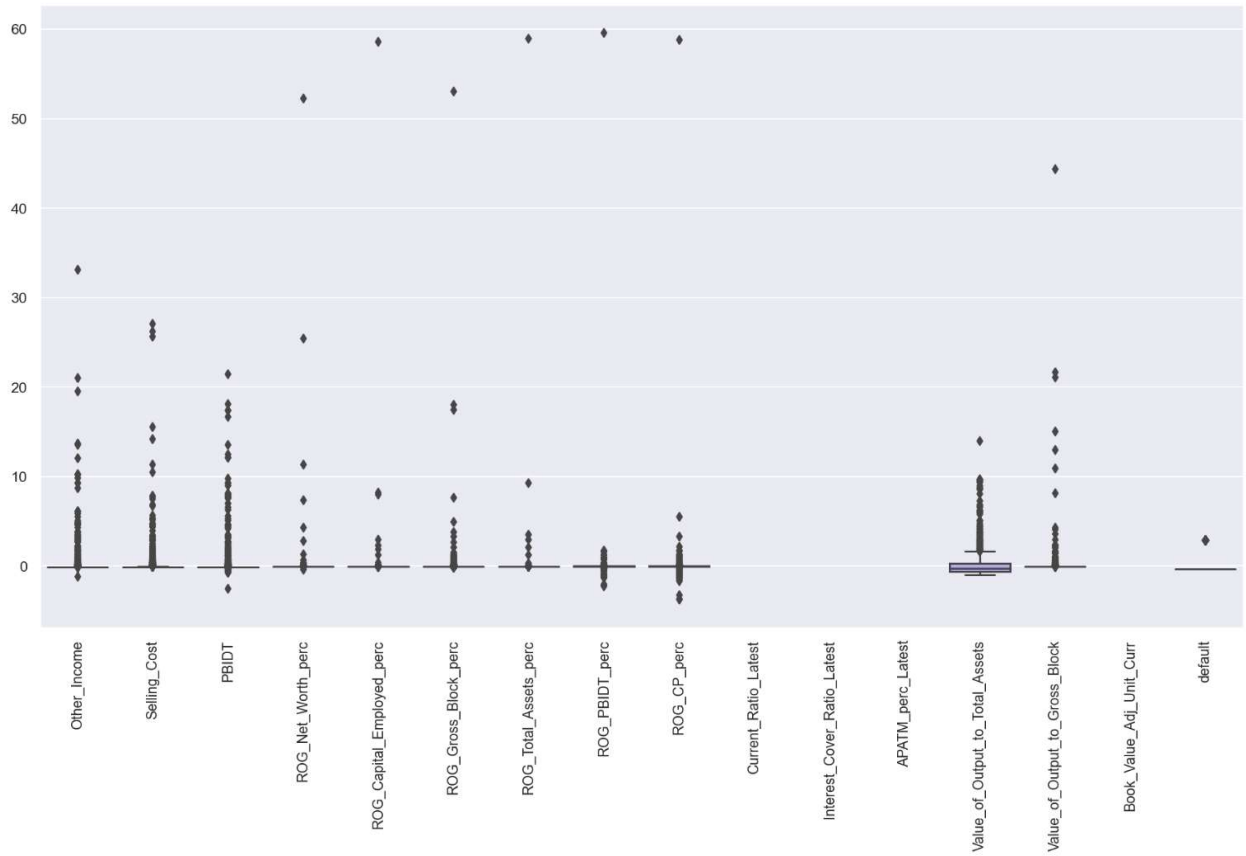


Fig 2 - Boxplot before Outlier Treatment - Top 15 predictors - Z-scaled

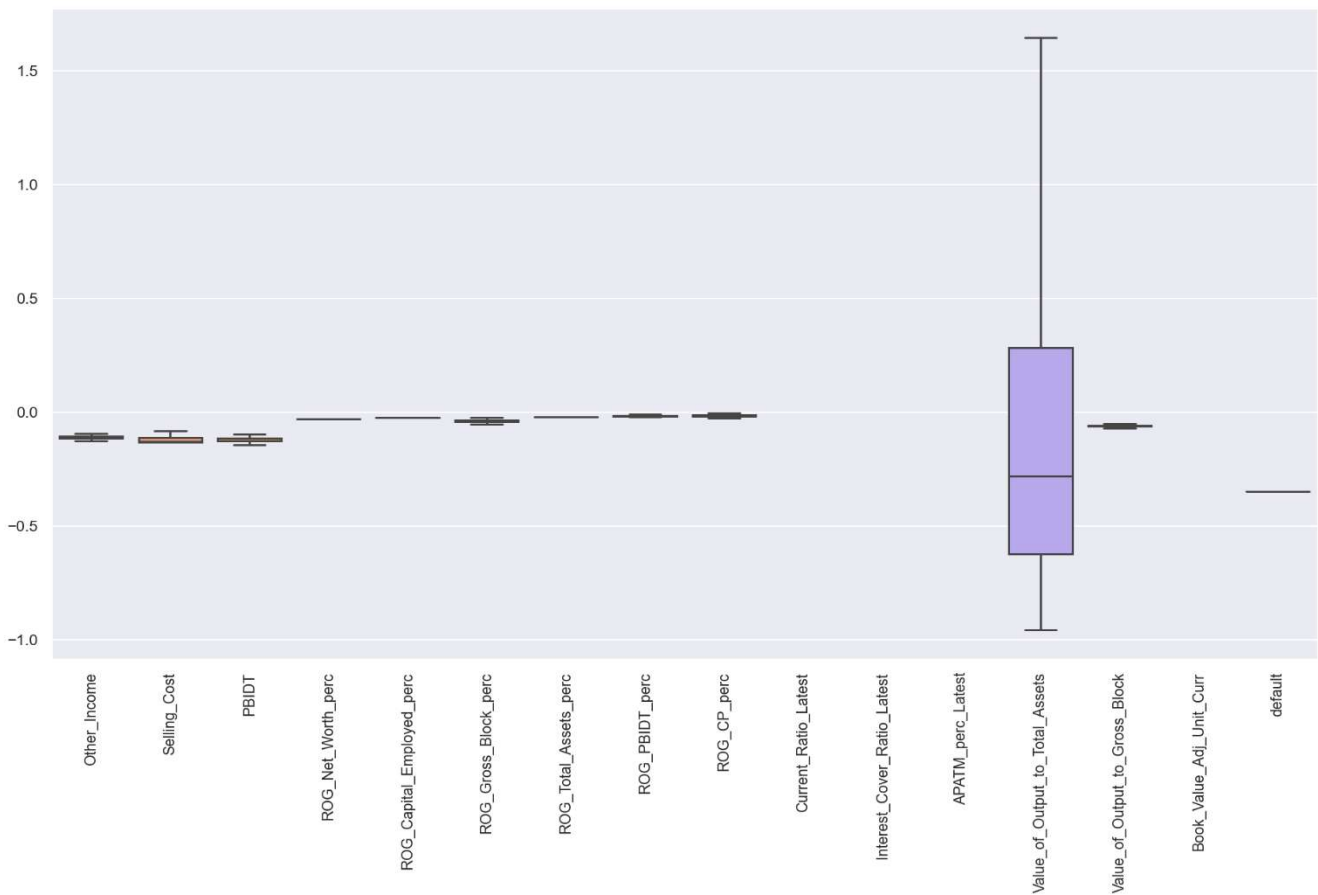


Fig 3 - Boxplot after IQR Treatment - Top 15 predictors - Z-scaled



## Z-score Treatment –

$$\text{z-score} = \frac{\text{value} - \text{mean of feature}}{\text{Std. dev}}$$

- We find a z-score of each value of the feature.
- Outlier = any value with z-score < -3 or > 3
- We cap outliers to  $\pm 3$  on either side.

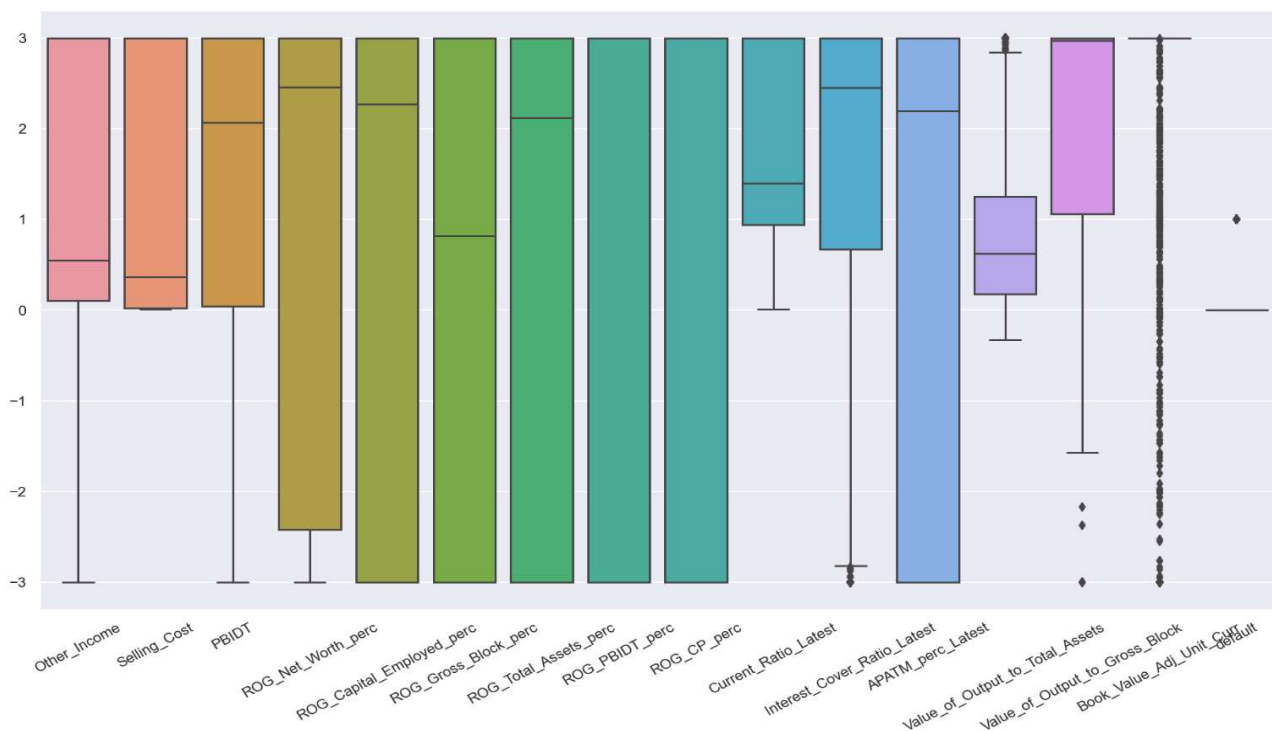


Fig 4 - Boxplot after Z-score Treatment - Top 15 predictors - Z-scaled

## 1.2 Missing Value Treatment

- Missing values in the raw data are very less, about 0.05%
- But there are large number of zero values, which are mostly placeholders for missing values, about 15%
- Also, these zero values add no more value to predictions
- But also mainly, large number of zero values in any feature cause 'Linear Algebra Error' while using StatsModel

Hence, it is of paramount importance to treat these zero values

- Firstly, we drop all those features with zero values greater than 30%
- Then, we convert all other zero values to Missing Values (Nan values)
- These transformed and original missing values together are imputed using KNN Imputer (n\_neighbors=10)
- A visual of all these missing values is give below - after dropping vars

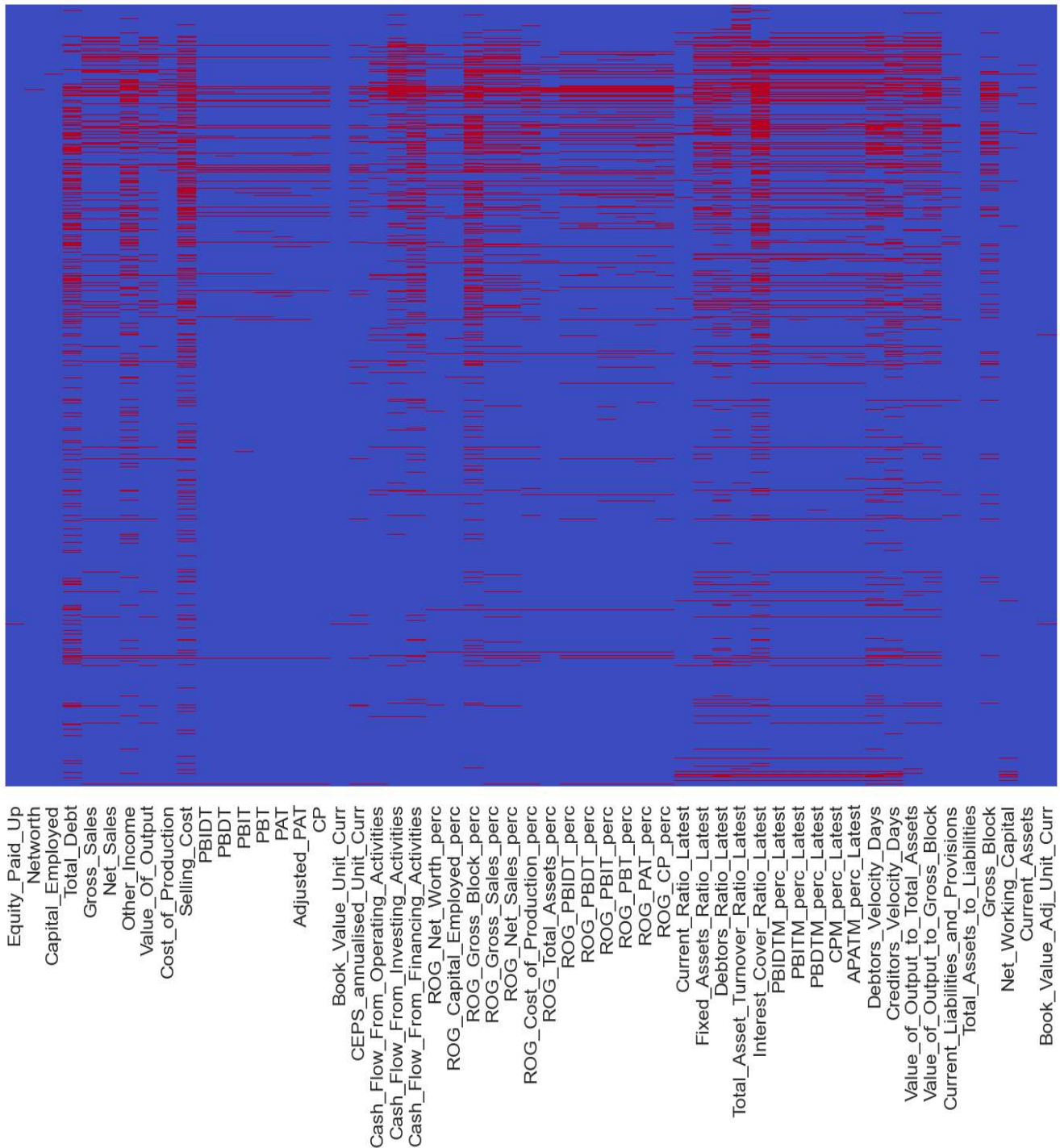


Fig 5 - Missing Values visualized - Raw data (55 vars)

## 1.3 Transform Target variable into 0 and 1

- We check the financial health of companies
- We'll base our prediction on Company's health on whether they will have a positive Net-worth next year or negative
- Hence, We consider 'Networth Next Year' as our Default Variable
- So, we call negative values as Default = 1
- And, zero or positive values as Default = 0
- We convert accordingly - Below is the sample

| default | Networth_Next_Year |
|---------|--------------------|
| 1       | -8021.6            |
| 1       | -3986.19           |
| 1       | -3192.58           |
| 1       | -3054.51           |
| 1       | -2967.36           |

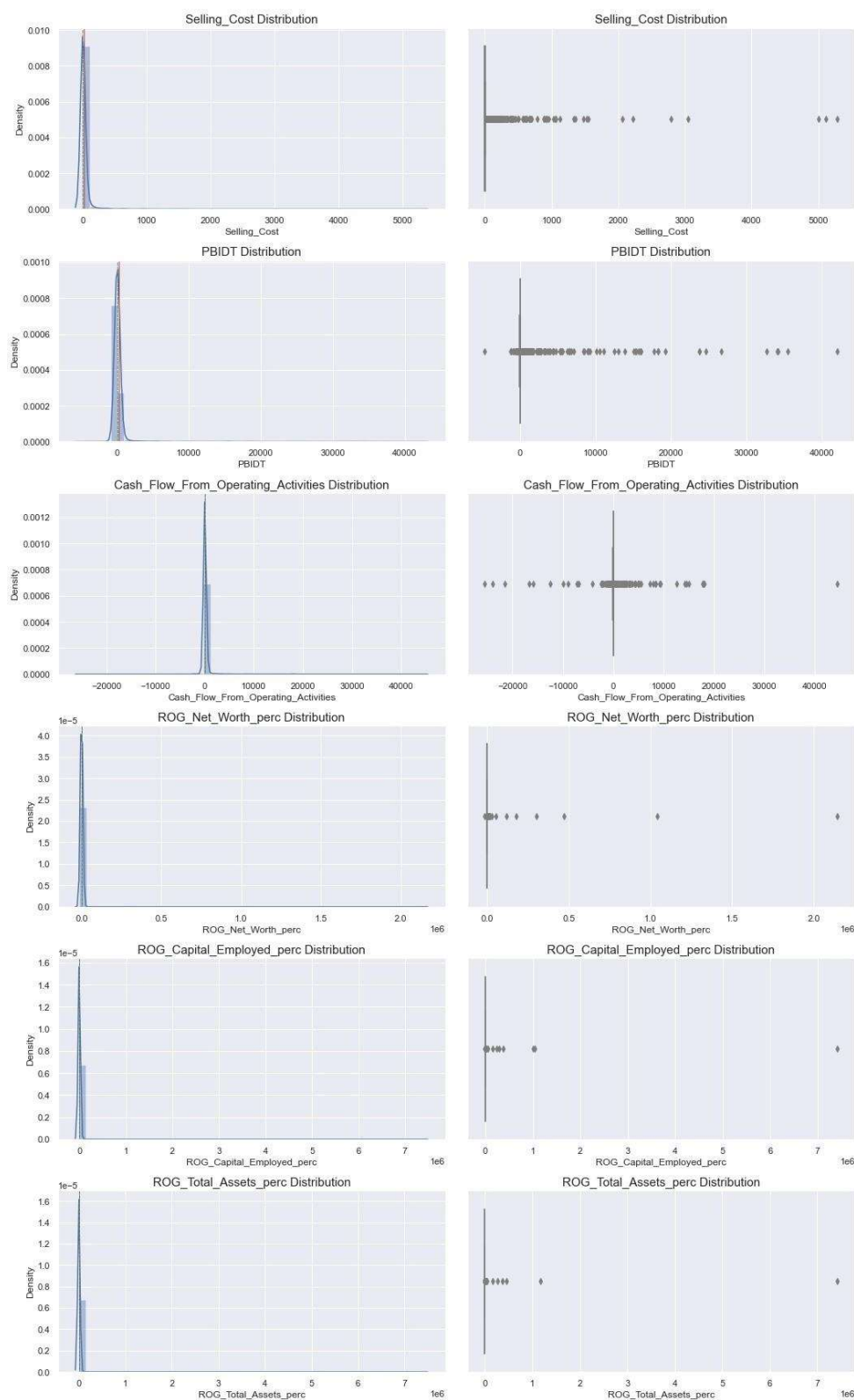
Table 2 - Target Variable - First 5

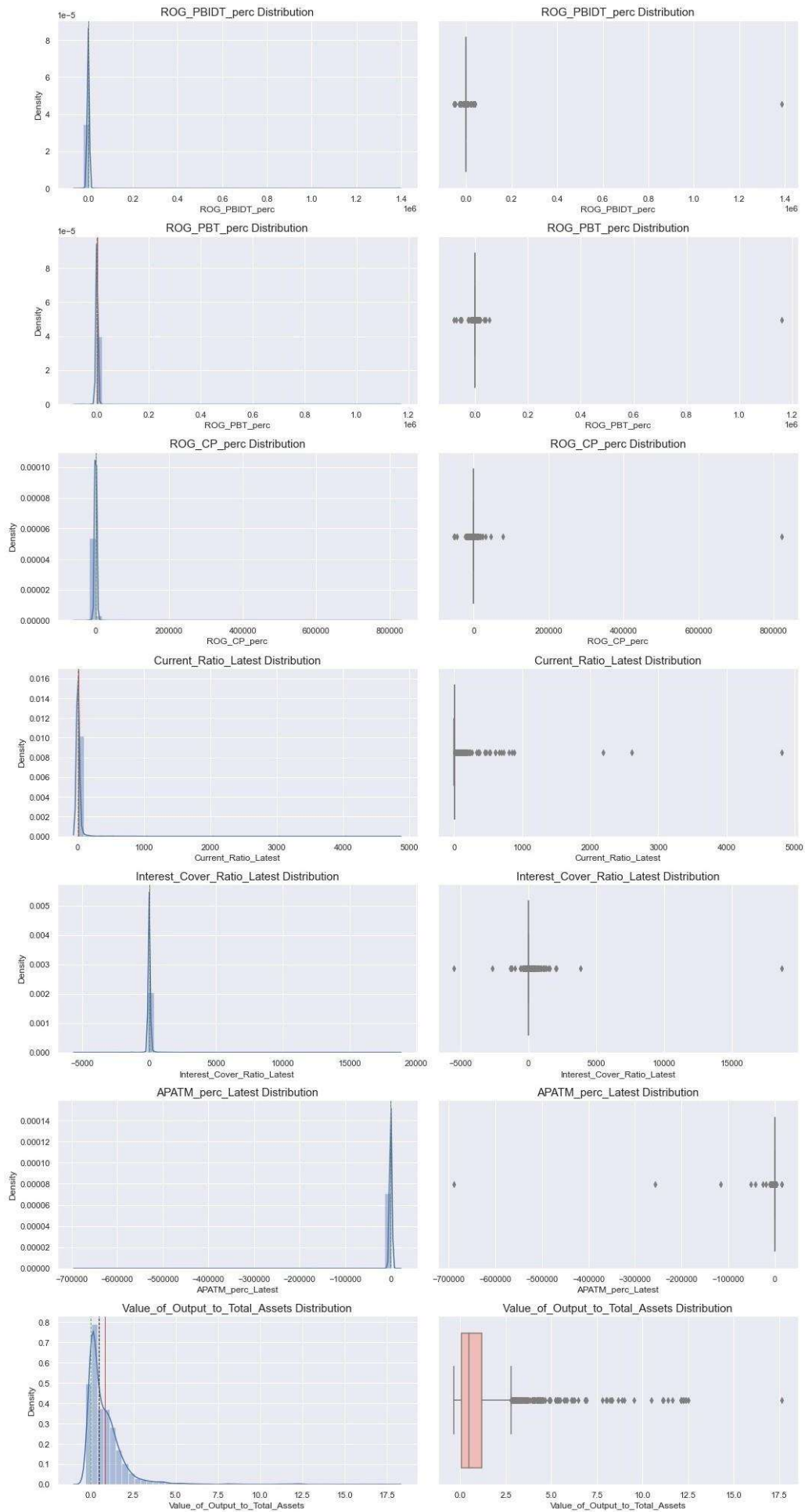
| default | Networth_Next_Year |
|---------|--------------------|
| 0       | 72677.77           |
| 0       | 79162.19           |
| 0       | 88134.31           |
| 0       | 91293.7            |
| 0       | 111729.1           |

Table 3 - Target Variable - Last 5

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

### UNIVARIATE ANALYSIS





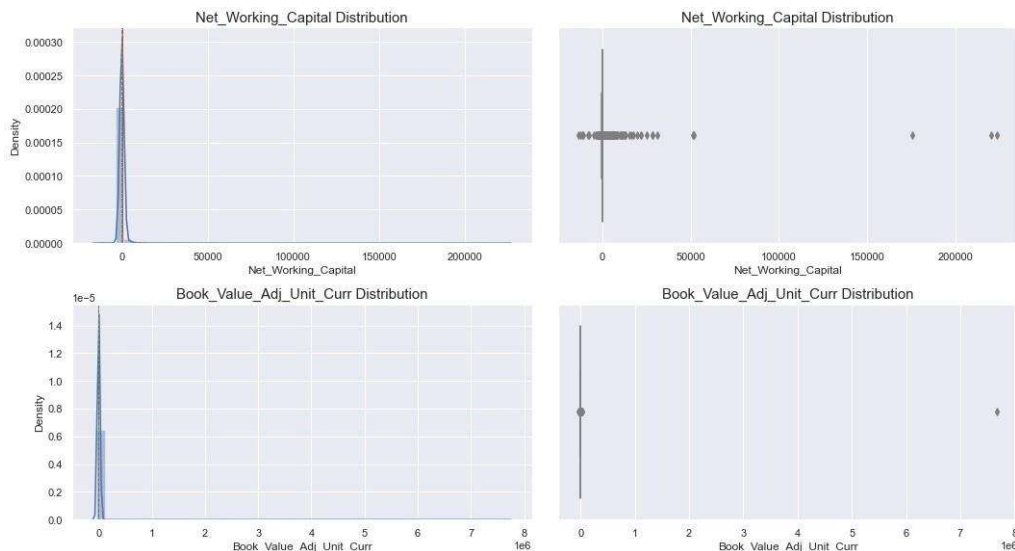


Fig 6 - Distribution of top 15 vars - Z-Scaled

- Distribution of Z-Scaled data of top 15 variables given
- Coloured vertical lines in the distribution indicate central tendencies
  - Mean by Red
  - Median by Black
  - Mode by Green
- 'Selling Cost' - has max companies around its mean. They have Right Skew with outliers on higher side.
- 'PBIDT' - 'Profit before Int Depreciation and Tax' - max companies are around the mean with a prominent right skew. This indicates that there are still many companies with high PBIDT
- 'Cash Flow from Operating Activities' - normal distribution with max companies lying around the mean
- 'ROG Network', 'ROG Capital Employed', 'ROG Total Assets', 'ROG PBIDT', 'ROG PBT (Profit Before Tax)', 'ROG CP', 'Current ratio Latest', 'Interest Cover Ratio Latest', 'Value of Output to Total Assets', 'Net Working Capital', 'Book Value Adjusted' - these variables have max density of companies around its mean with right skew. This indicates outliers on the higher side.
- 'APATM (After Tax Profit Margin)' - has max density around its mean and a prominent left skew. This indicates that there are many companies have their Net Profit on the lower side of the distribution - Possible indication of default
- Largely, it is observed that there are many companies with good margin and financials before tax and all other costs. But, after costs are considered, they slide to the lower half - Shows they need to work on their costs and bottom line



## BIVARIATE ANALYSIS

- Correlation Heat-map is given below of 55 variables (after dropping vars with zeros>30%)

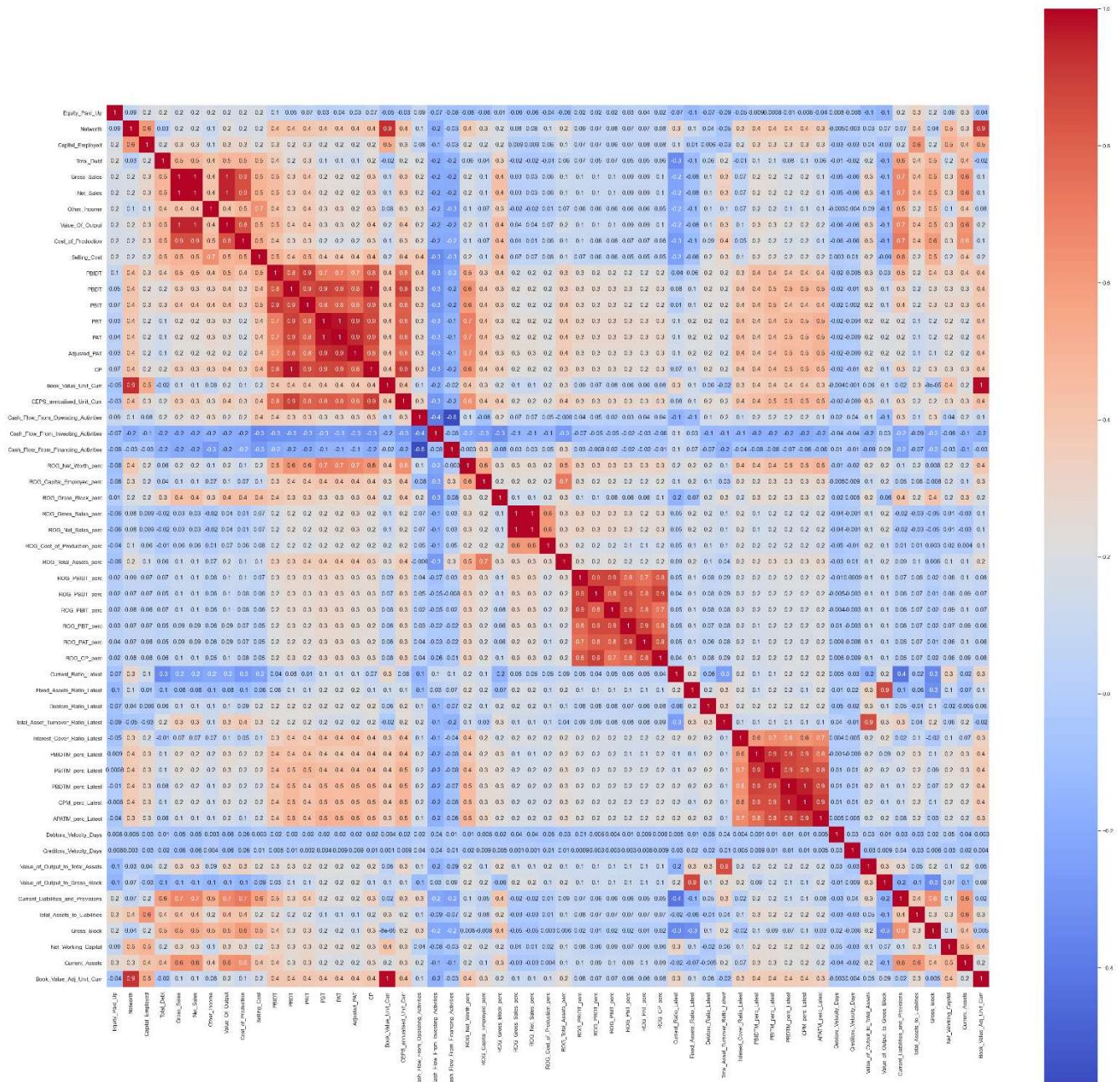


Fig 7 - Correlation Heatmap - 55 Variables

- There are a lot of red patches seen. This indicates high correlation between many variables
- Highly correlated features cause Multi-Collinearity which affect the interpretability of Logistic Regression model. They are best removed.

- We use Variance Inflation factor method and remove all variables with  $VIF > 5$ . This is done recursively, one-by-one
- Correlation heat-map of top 15 predictors and 1 Target, used to get the best model is given below-

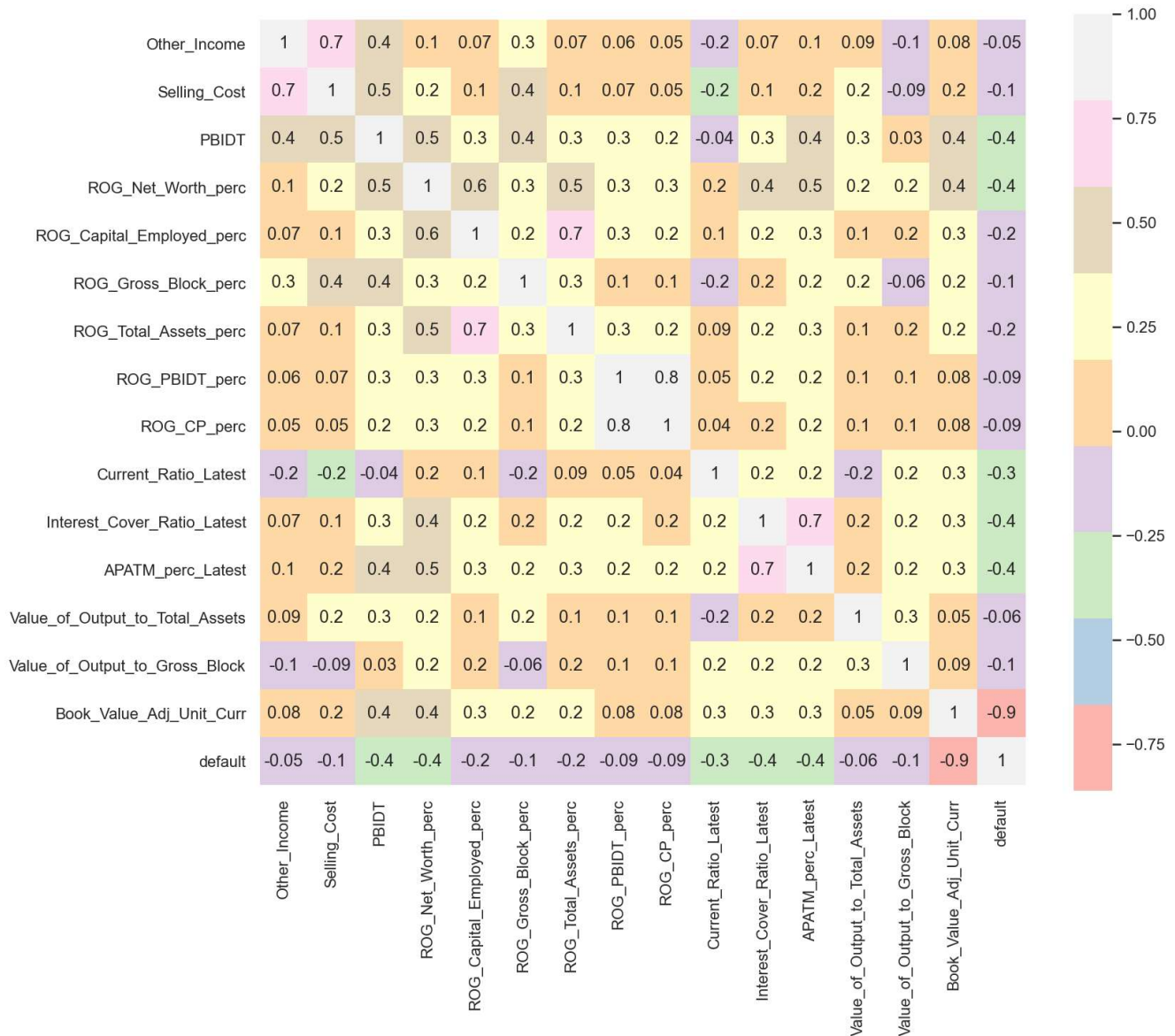


Fig 8 - Correlation Heatmap of top 15 vars

- 'ROG-Capital Employed and ROG-Total Assets' — 'ROG-PBT and ROG-PBIDT' — 'ROG-CP and ROG-PBIDT' — 'ROG-CP and ROG-PBT'
  - the above pairs of features show a high correlation
  - it looks obvious as they seem derived or direct functions of each other
- Target variable 'default' has a high negative correlation with 'Book Value Adj'
  - this indicates as Book Value rises, the Probability of Default falls



## 1.5 Train Test Split

- We use the `train_test_split` function from the scikit-learn library to split the data into train and validation sets
- We split in the ratio of 67-33 - 67% in Training Set and 33% in Testing (Validation) Set
- We seed this split at `random_state=42`
- So after split, Out of Total 3586 —> Train Set has 2402 observations
  - >Test Set has 1184 observations

## 1.1 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

### PREPROCESSING DATA FOR ALL MODELS

- We had large number of 66 predictor variables in the raw data
- 2 unique identifiers - Company Code and Name - were dropped
- There are a large number of zero values in the data. We drop all variables with zeros > 30% (9 vars)
- Highly correlated features exist in the data, which cause Multi-Collinearity. This indicates the existence of redundant features
- We perform Outlier treatments using IQR and Z-score methods
- We drop all features recursively one-by-one with  $VIF > 5$  (VIF - Variance Inflation Factor)
- We are left with - 27 variables after IQR Outlier treatment
  - 23 variables after Z-score treatment

### LOGISTIC REGRESSION MODELS

- We build multiple Logistic Regression models with different approaches and strategies. We test each model on the Test set and fine-tune to improve Recall and Precision of default=1
- We use StatsModel and SciKitLearn libraries to build and test these models

- Model 1 :
  - 27 vars, IQR Outlier treated
- Model 2 :
  - From 27 vars - insignificant vars dropped with p-values > 0.05 (final 10 vars)
  - IQR Outlier treatment
- Model 3 :
  - 23 vars, Z-score Outlier treatment
- Model 4 :
  - From 23 vars - insignificant vars dropped with p-values > 0.05 (final 9 vars)
  - Z-score Outlier treatment
- Model 5 :
  - Above 9 vars, Z Score treatment
  - Regularising the model by Hyper-parameter tuning with GridSearch over 10 folds over following -
 

```
{'penalty':['l2','none', 'l1'],
'solver':['lbfgs', 'liblinear', 'sag', 'saga', 'newton-cg'], 'tol':[0.0001,0.00001]}
```
  - Best Parameters were found as follows -
 

```
{'penalty': 'none', 'solver': 'lbfgs', 'tol': 0.0001}
```
- Model 6 :
  - From 23 vars - insignificant vars dropped with p-values > 0.05 (final 9 vars)
  - Z-score Outlier treatment
  - Check for optimum threshold to get max Recall for default=1
  - This is obtained by maximizing the difference between True Positivity rate and False Positivity rate (tpr - fpr)
  - Optimum Threshold = 0.084
- Model 7 :
  - 23 vars, Z-score Outlier treatment
  - Extracting top 15 features using Recursive Feature Elimination (RFE)
- Model 8 :
  - 23 vars, Z-score Outlier treatment
  - Extracting top 15 features using Recursive Feature Elimination (RFE)
  - This model gave the best metrics on Test Set

- StatsModel report of Model 9 given below -
- We note that 'Book\_Value\_Adj\_Unit\_Curr' has the highest negative coefficient
  - suggesting that this variable has the highest negative impact on Probability of Default
- Also, 'Selling\_Cost' has the highest positive coefficient
  - suggesting that this variable has the highest positive impact on Probability of Default

|                                  | coef    | std err | z       | P> z  | [0.025 | 0.975] |
|----------------------------------|---------|---------|---------|-------|--------|--------|
| Other_Income                     | 0.3180  | 0.101   | 3.145   | 0.002 | 0.120  | 0.516  |
| Selling_Cost                     | 0.6835  | 0.119   | 5.754   | 0.000 | 0.451  | 0.916  |
| PBIDT                            | -0.3132 | 0.056   | -5.553  | 0.000 | -0.424 | -0.203 |
| ROG_Net_Worth_perc               | -0.4148 | 0.049   | -8.388  | 0.000 | -0.512 | -0.318 |
| ROG_Capital_Employed_per c       | 0.4119  | 0.054   | 7.656   | 0.000 | 0.306  | 0.517  |
| ROG_Gross_Block_perc             | 0.0368  | 0.047   | 0.790   | 0.429 | -0.055 | 0.128  |
| ROG_Total_Assets_perc            | -0.2154 | 0.055   | -3.939  | 0.000 | -0.323 | -0.108 |
| ROG_PBIDT_perc                   | 0.2171  | 0.057   | 3.835   | 0.000 | 0.106  | 0.328  |
| ROG_CP_perc                      | -0.1513 | 0.057   | -2.677  | 0.007 | -0.262 | -0.041 |
| Current_Ratio_Latest             | -0.3419 | 0.107   | -3.199  | 0.001 | -0.551 | -0.132 |
| Interest_Cover_Ratio_Lates t     | -0.3591 | 0.059   | -6.077  | 0.000 | -0.475 | -0.243 |
| APATM_perc_Latest                | -0.2973 | 0.055   | -5.371  | 0.000 | -0.406 | -0.189 |
| Value_of_Output_to_Total_A ssets | -0.0817 | 0.177   | -0.461  | 0.645 | -0.429 | 0.266  |
| Value_of_Output_to_Gross_ Block  | 0.2140  | 0.092   | 2.329   | 0.020 | 0.034  | 0.394  |
| Book_Value_Adj_Unit_Curr         | -1.6181 | 0.071   | -22.933 | 0.000 | -1.756 | -1.480 |

Table 4 - StatsModel Summary Report of Model #8

- Model 9
  - 23 vars, Z-score Outlier treatment
  - Extracting top 15 features using Recursive Feature Elimination (RFE)
  - Check for optimum threshold to get max Recall for default=1
  - This is obtained by maximizing the difference between True Positivity rate and False Positivity rate (tpr - fpr).
  - Optimum Threshold = 0.4246
- Model 10 :
  - Z-Score Outlier treatment, Top 15 features through RFE
  - Class balancing 50-50 using SMOTE
  - Optimum Threshold = 0.5
  - Dropping insignificant vars with p-values > 0.05

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

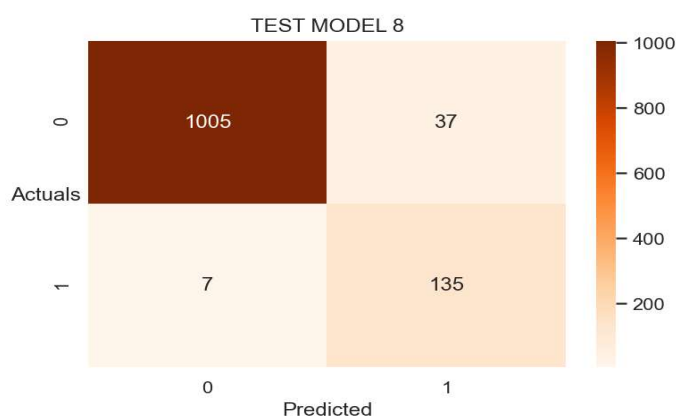
- Performance Metrics of all models on Test Dataset is given below

| MODEL NAME | RECALL FOR 1<br>(in %) | PRECISION<br>FOR 1<br>(in %) | ACCURACY<br>(in %) | F-1 FOR 1<br>(in %) |
|------------|------------------------|------------------------------|--------------------|---------------------|
| Model 1    | 99                     | 23                           | 61                 | 38                  |
| Model 2    | 99                     | 23                           | 60                 | 37                  |
| Model 3    | 88                     | 90                           | 97                 | 89                  |
| Model 4    | 88                     | 89                           | 97                 | 88                  |
| Model 5    | 88                     | 89                           | 97                 | 88                  |
| Model 6    | 95                     | 71                           | 95                 | 81                  |
| Model 7    | 87                     | 89                           | 97                 | 88                  |
| Model 8    | 95                     | 78                           | 96                 | 86                  |
| Model 9    | 95                     | 75                           | 96                 | 84                  |
| Model 10   | 92                     | 78                           | 96                 | 85                  |

Table 5 - All Model Performance Comparison

|                     | precision | recall | f1-score | support |
|---------------------|-----------|--------|----------|---------|
| <b>0</b>            | 0.99      | 0.96   | 0.98     | 1042.00 |
| <b>1</b>            | 0.78      | 0.95   | 0.86     | 142.00  |
|                     |           |        |          |         |
| <b>accuracy</b>     | 0.96      | 0.96   | 0.96     | 0.96    |
| <b>macro avg</b>    | 0.89      | 0.96   | 0.92     | 1184.00 |
| <b>weighted avg</b> | 0.97      | 0.96   | 0.96     | 1184.00 |

Table 6 - Classification Report of Model 8



- We have Models 1 and 2 with Best Recall of 99% but very poor Precision, hence we ignore these models
- We choose Model 8 for deployment, because of its best combination scores of Recall and Precision

### **INTERPRETATION:-**

- Recall of 95% means - 95% of Actual Defaults were Predicted Correctly
- Precision of 78% means - 78% of Predicted Defaults were Actual
- For this modeling, we needed to predict as many of Actual Defaults as possible and minimize Type 2 errors foremost
- Hence Recall and then Precision were considered in choosing the best model
- In Table 4 above, coefficients of all variables indicate the weightage of that variable in predicting Default

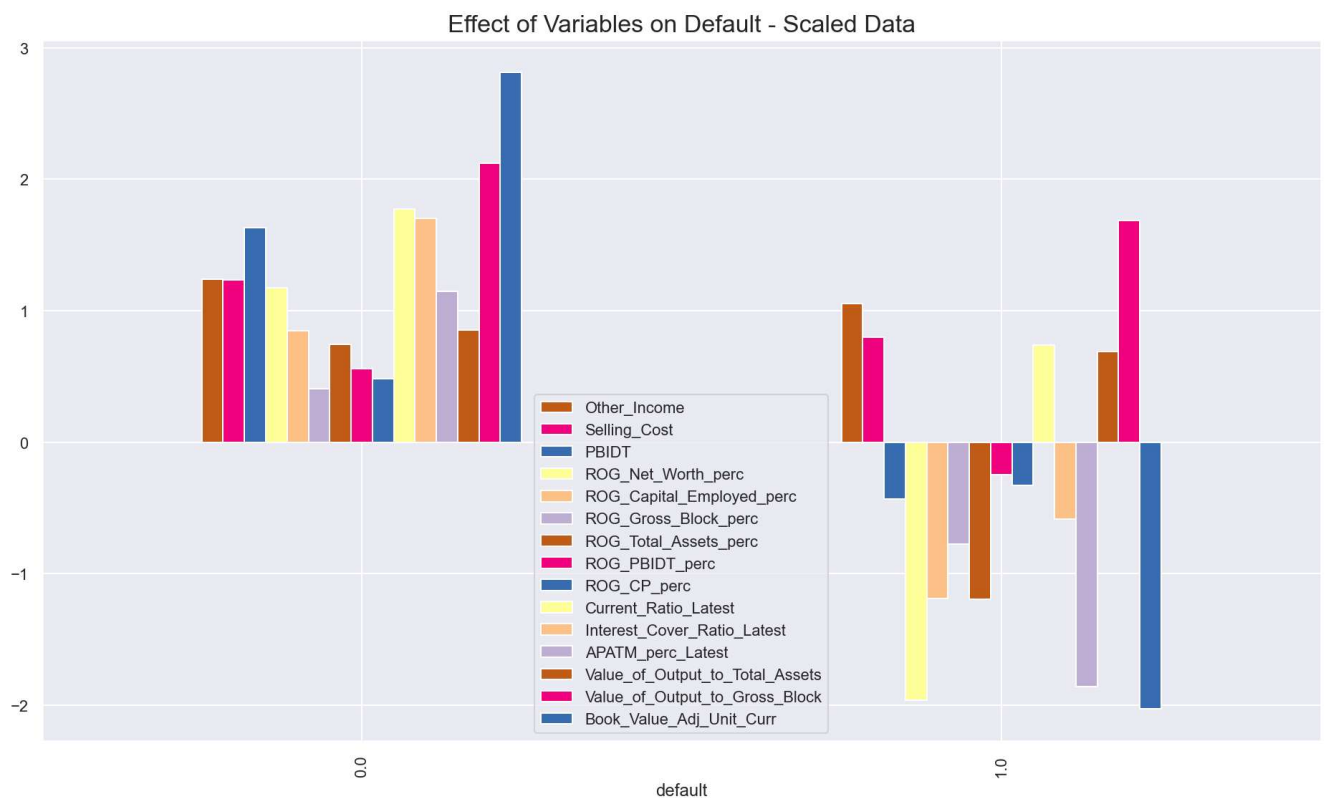


Fig 10 - Effect of Variables on Default

- Positive coefficient means, if all else is equal, then higher value of this variable will lead to higher likelihood of default
- Negative coefficient means, if all else is equal, then higher value of this variable will lead to lower likelihood of Default
- In figure 9 above, we see the effect of various features on Default
- We note that highest negative value is of - 'Book\_Value\_Adj\_Unit\_Curr'
  - suggesting if all else is same, then as Book Value of a Company increases, then the Probability of Default by that Company decreases

- Also, the highest positive coefficient is of - 'Selling\_Cost'
  - suggesting if all else is the same, then as the Selling Cost of a Company increases, then the Probability of Default by that Company increases
- From Multi-variate Analysis, we observed that many companies had good profit margins before considering taxes, interests, and other costs
- But once all costs are considered along-with taxes and depreciation, the majority of these companies slide to the bottom half in Profitability
- These companies should focus on optimizing their bottom line.

---Thanks---



