Business Report

SMDM Project Business Report DSBA

***Sanjay Srinivasan***

***PGP-DSBA Online***

***JULY’ 21 Batch***

***Date: 08-05-2022***

# *INDEX*

|  |  |  |
| --- | --- | --- |
| ***S. No*** | ***Contents*** | ***Page* No** |
| **1.** | **Problem - 1** | 4 |
|  | Summary | 4 |
|  | Introduction | 4 |
|  | Data Description | 4 |
|  | Sample dataset | 5 |
|  | Exploratory data analysis | 5 |
|  | 1) Outlier Treatment | 6 |
|  | 2) Missing Value Treatment | 6 |
|  | 3) Transform Target variable into 0 and 1 | 10 |
|  | 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) | 11 |
|  | 5) Train Test Split | 15 |
|  | 6) 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. | 19 |
|  | 7) Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model | 25 |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| ***S.No*** | ***Content*** | ***Page No*** |
| 1.1 | Dataset Sample Before Changing Column Names | 5 |
| 1.2 | Dataset Sample After Changing Column Names | 5 |
| 1.3 | Sample Datatypes of the variable with null values | 6 |
| 1.4 | Shape before Outliers Treatment | 7 |
| 1.5 | Shape After Outliers Treatment | 7 |
| 1.6 | Before Treating Missing value | 8 |
| 1.7 | After Treating Missing value | 8 |
| 1.8 | Default count. | 9 |
| 1.9 | Default count in percentage. | 9 |
| 1.10 | Sample data after Transformation. | 9 |
| 1.11 | Univariate Analysis | 9 |
| 1.12 | Scatterplot for Bivariate Analysis | 10 |
| 1.13 | Sample Multivariate analysis for correlation | 11 |
| 1.14 | Multivariate analysis of plotting correlation in heatmap | 12 |
| 1.15 | Multivariate analysis of plotting correlation in heatmap after dropping insignificant variable. | 13 |
| 1.16 | Taking Highly correlated variables. | 13 |
| 1.17 | X-train shape. | 13 |
| 1.18 | X-test shape. | 13 |
| 1.19 | Y-train shape. | 13 |
| 1.20 | Y-test shap | 13 |
| 1.21 | Selecting the feature with rank 1 | 14 |
| 1.22 | Model-1 summary report | 14 |
| 1.23 | Variance Inflation Factor.of Model-1 | 15 |
| 1.24 | Model-2 Summary Report | 15 |
| 1.25 | Model-3 Summary Report | 16 |
| 1.26 | Model-4 Summary Report | 17 |
| 1.27 | Model-5 Summary Report | 18 |
| 1.28 | Model-6 Summary Report | 19 |
| 1.29 | Model-7 Summary Report | 20 |
| 1.30 | Model-8 Summary Report | 21 |
| 1.31 | Model-9 Summary Report | 22 |
| 1.32 | Optimum threshold | 22 |
| 1.33 | Confusion matrix for train data | 22 |
| 1.34 | Confusion matrix for test data | 22 |
| 1.35 | Classification report for train data | 23 |
| 1.36 | Classification report for test data | 23 |

***List Of Figures***

# Problem - 1

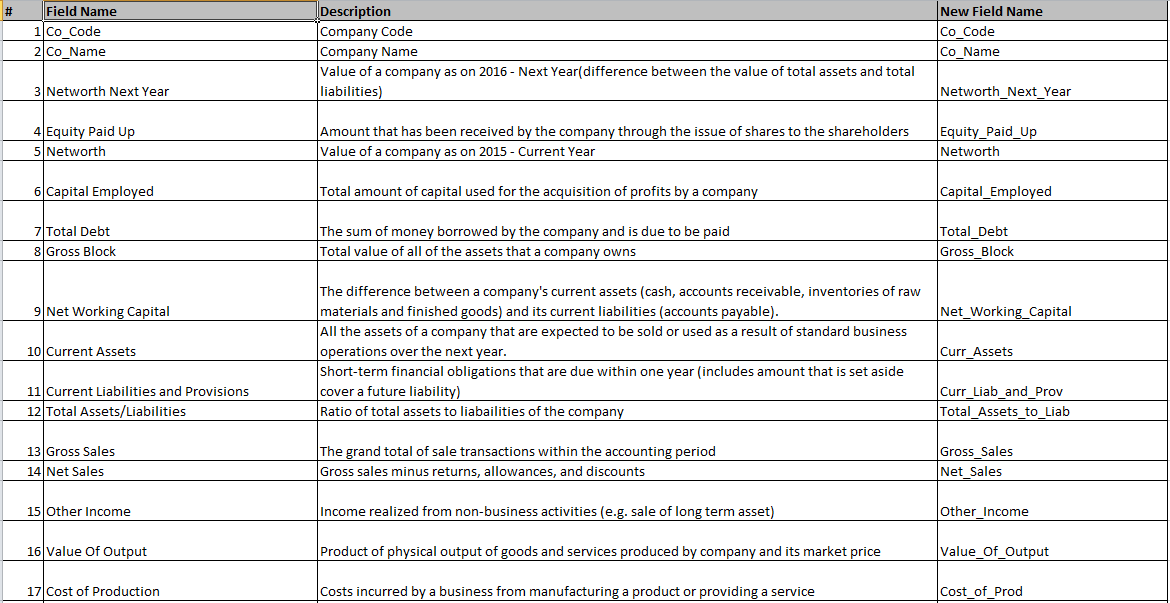
*Summary*

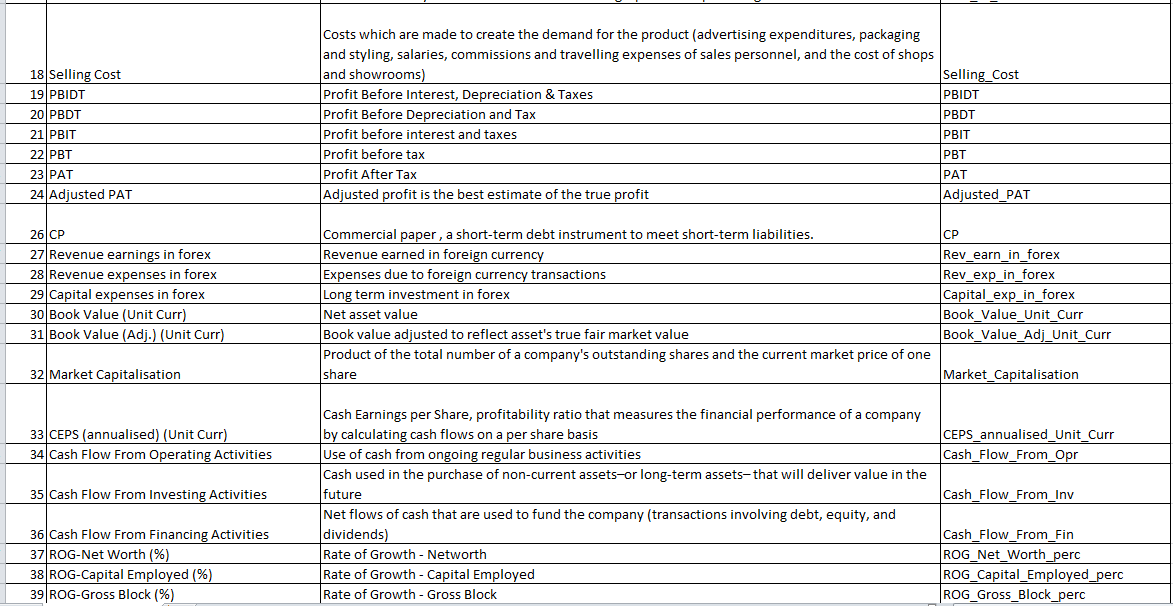
The data is gathered based on the company financial balance sheet, which deals with the company finances. This dataset has financial statements for 3586 company with 67 variables. For investing in the company, to analyse from the investor’s point of view, to predict that the company is capable of handling the financial obligation, can grow quickly and manage the growth scale.

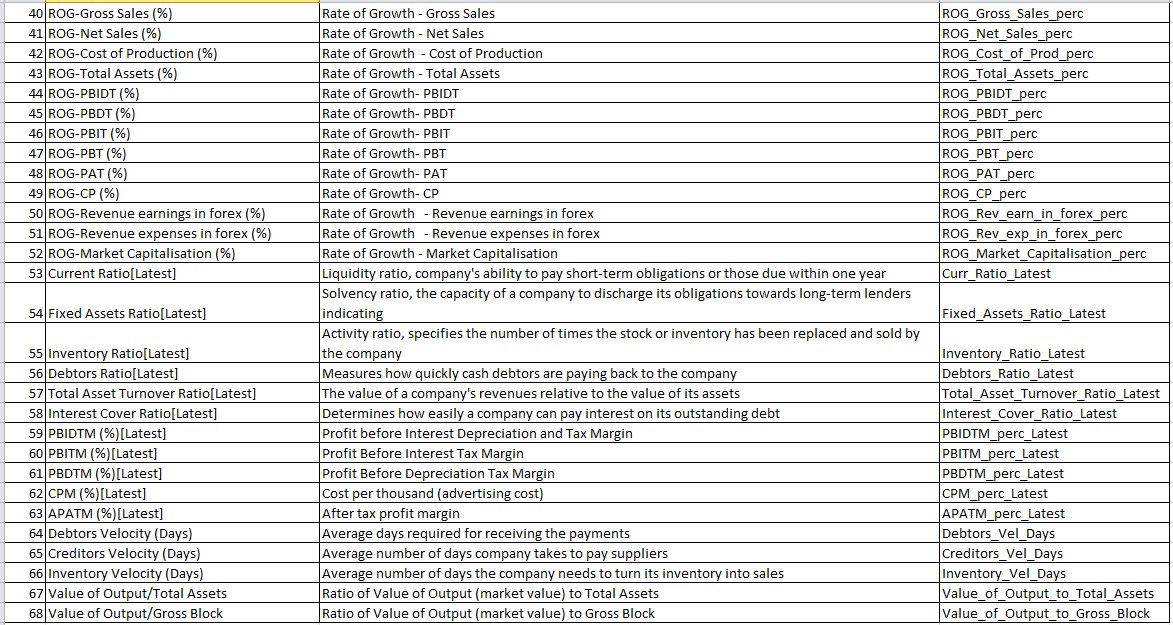
*Introduction*

The purpose of this exercise is to find the company with good credit rating and handling the financial obligation.

*Data Description*







*Sample of the dataset:*

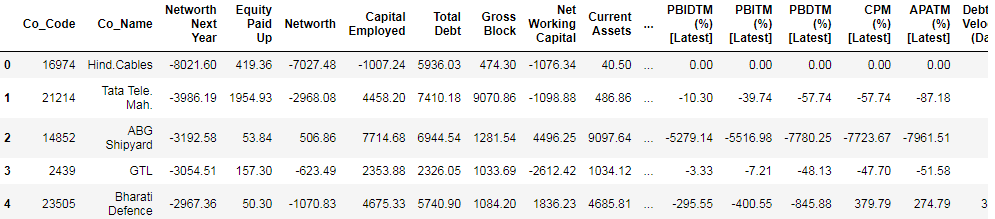


Fig 1.1 Dataset Sample Before Changing Column Names

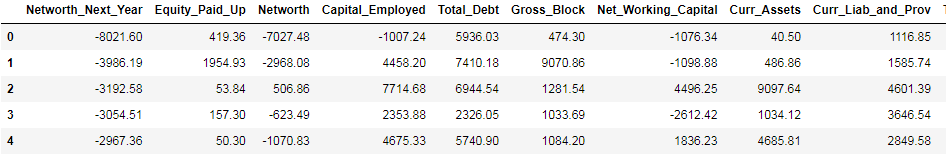


Fig 1.2 Dataset Sample After Changing Column Names

*Exploratory Data Analysis*

*Let us check the types of variables in the data frame.*

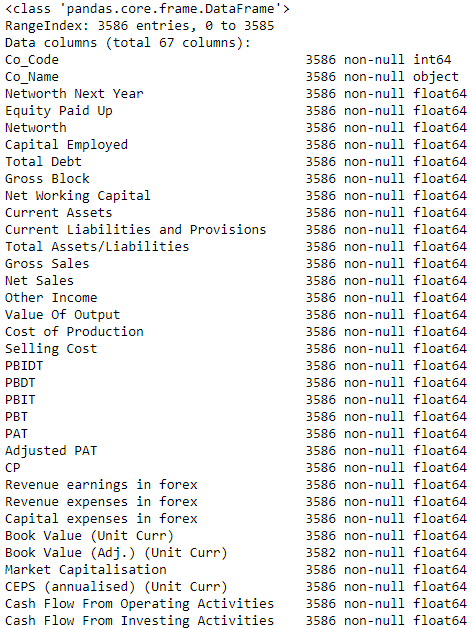


Fig- 1.3. Sample Datatypes of the variable with null values

There are total 3586 rows and 67 columns in the dataset.

**1.1 Outlier Treatment**

The boxplot is plotted for all the variable without treating the outliers.



Fig- 1.4 Shape before Outliers Treatment

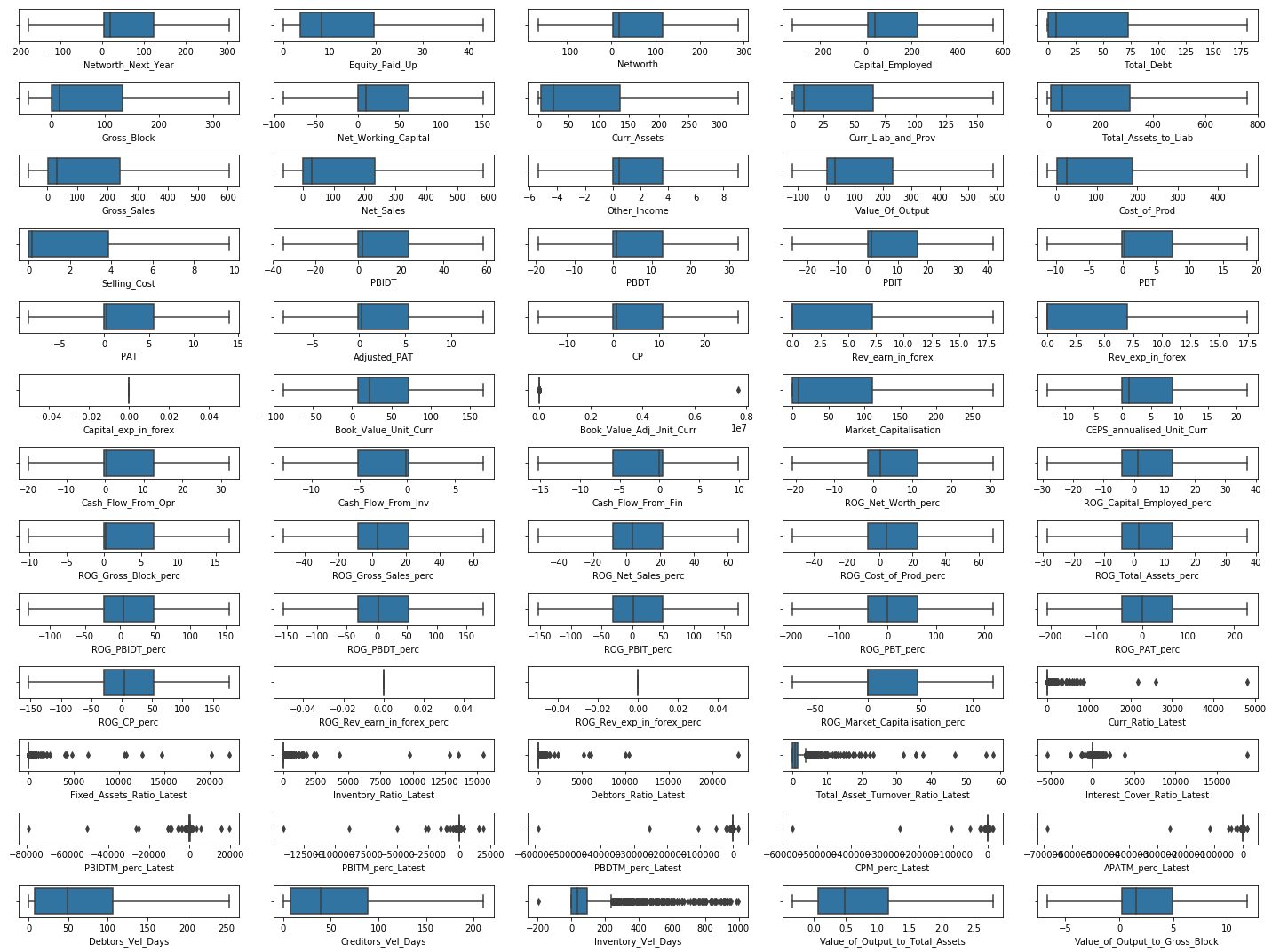


Fig- 1.5 Shape After Outliers Treatment

**1.2 Missing Value Treatment**

Fig- 1.4 Shape After Outliers Treatment

|  |  |
| --- | --- |
| Fig- 1.6 Before Treating Missing value | Fig- 1.7 After Treating Missing value |

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

Target value ‘Networth\_Next\_year’ is transform into 0’s and 1’s.

Networth\_Next\_year < 0 (negative) then target or default variable = 1

Networth\_Next\_year > 0 (positive) then target or default variable = 0

1 - Company might default.

0 – Company might not default.

|  |  |
| --- | --- |
| Fig – 1.8 Default count. | Fig – 1.9 Default count in percentage. |

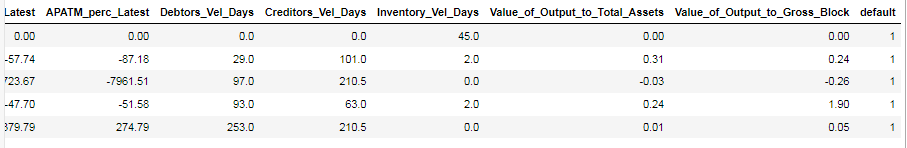


Fig – 1.10 Sample data after Transformation.

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

***Uni-Variate Analysis:***

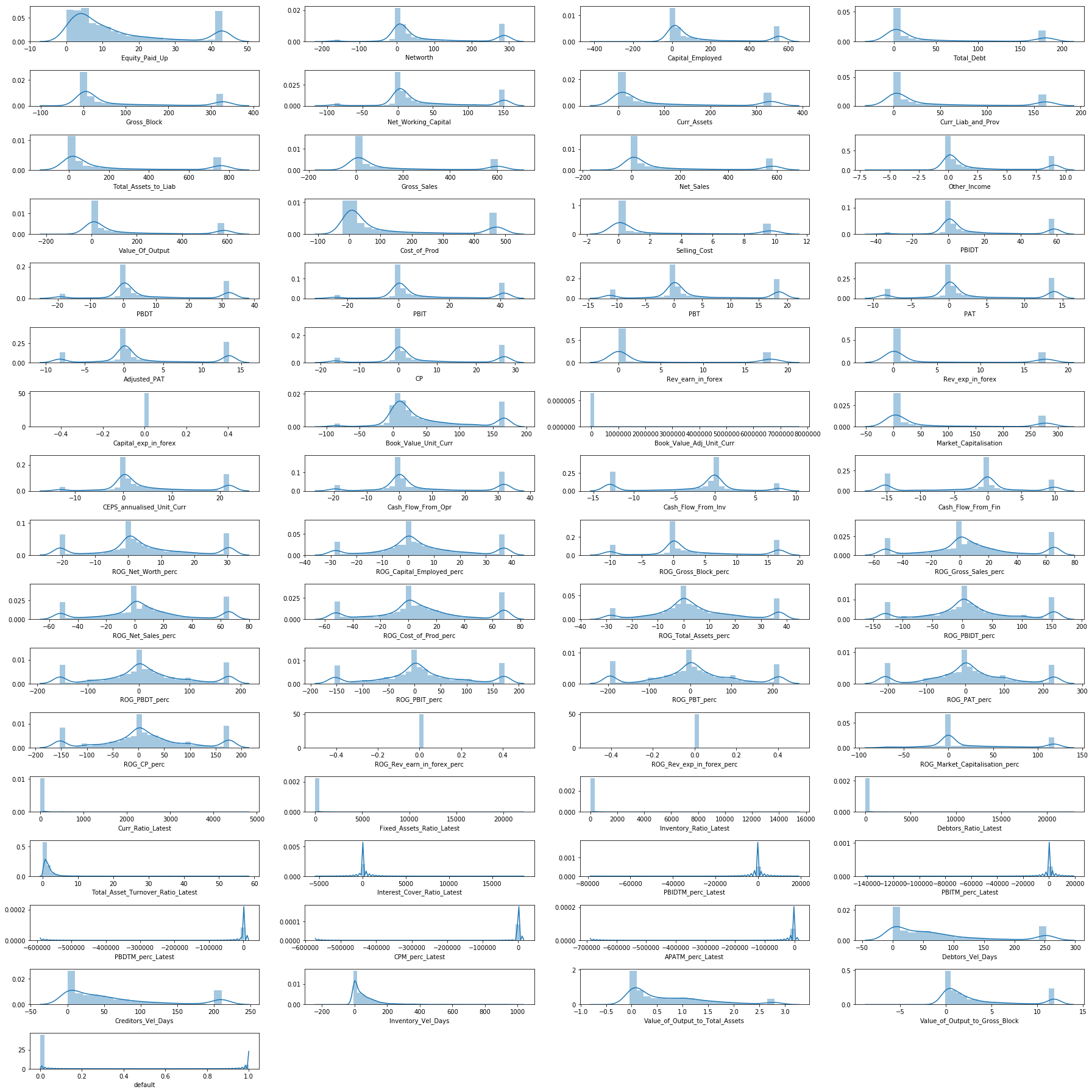


Fig – 1.11 Univariate Analysis

From the above chart (displot and boxplot), there are outliers present in the economic.cond.national and economic.cond.household data. We can infer that there is no trend or pattern that it follows a normal distribution.

***Bi – variate Analysis:***

|  |  |  |
| --- | --- | --- |
|  |  |  |

Fig – 1.12 Scatterplot for Bivariate Analysis

***Multi – variate Analysis:***

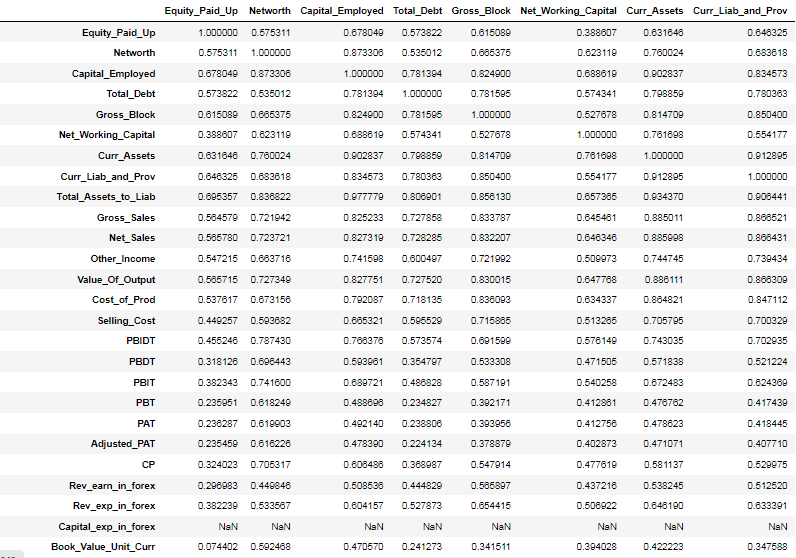


Fig – 1.13 Sample Multivariate analysis for correlation

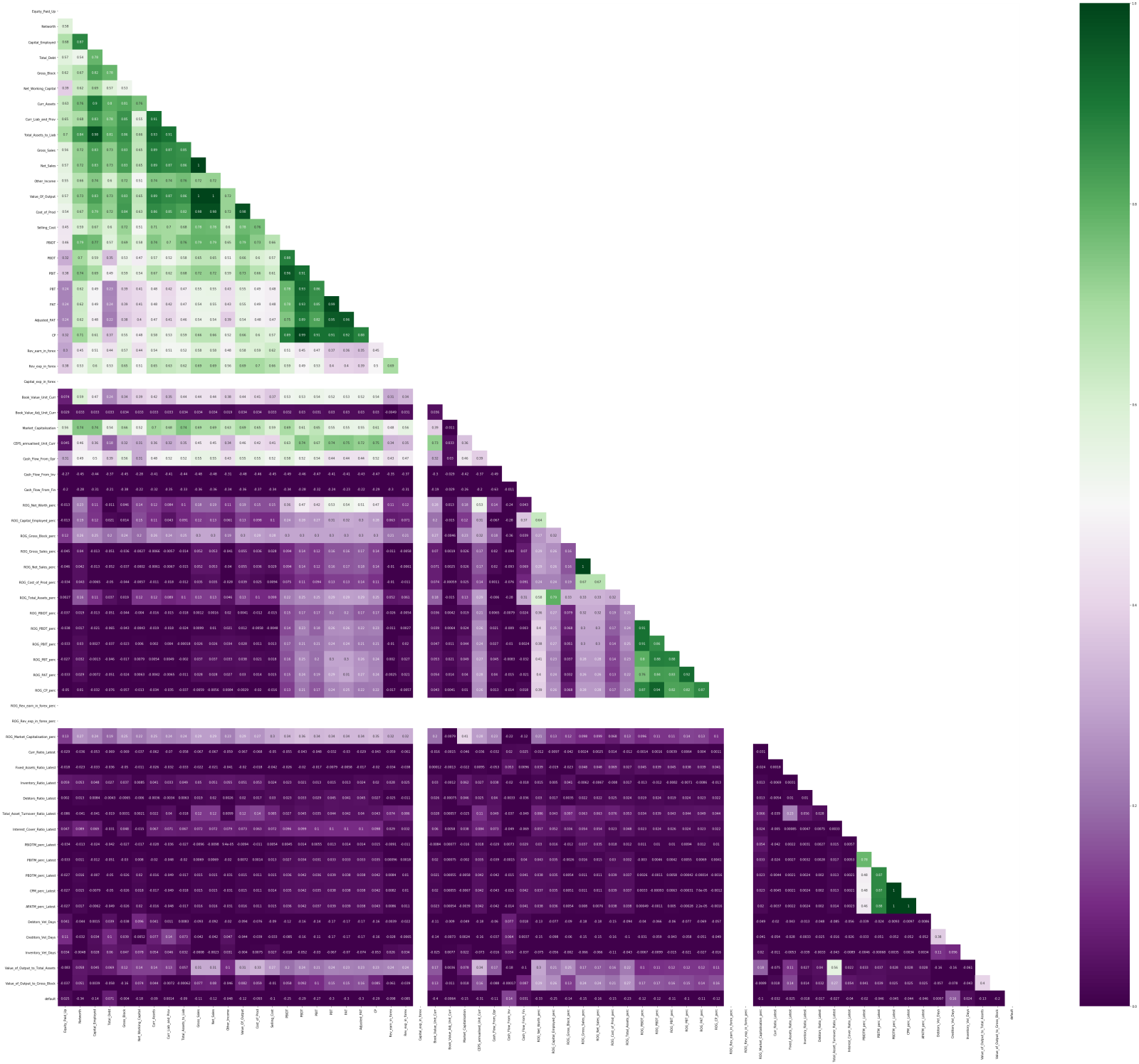


Fig – 1.14 Multivariate analysis of plotting correlation in heatmap

From this Heatmap we can infer that 3 variables do not have any correlation and do not contribute on the output. So, dropping the insignificant variables.

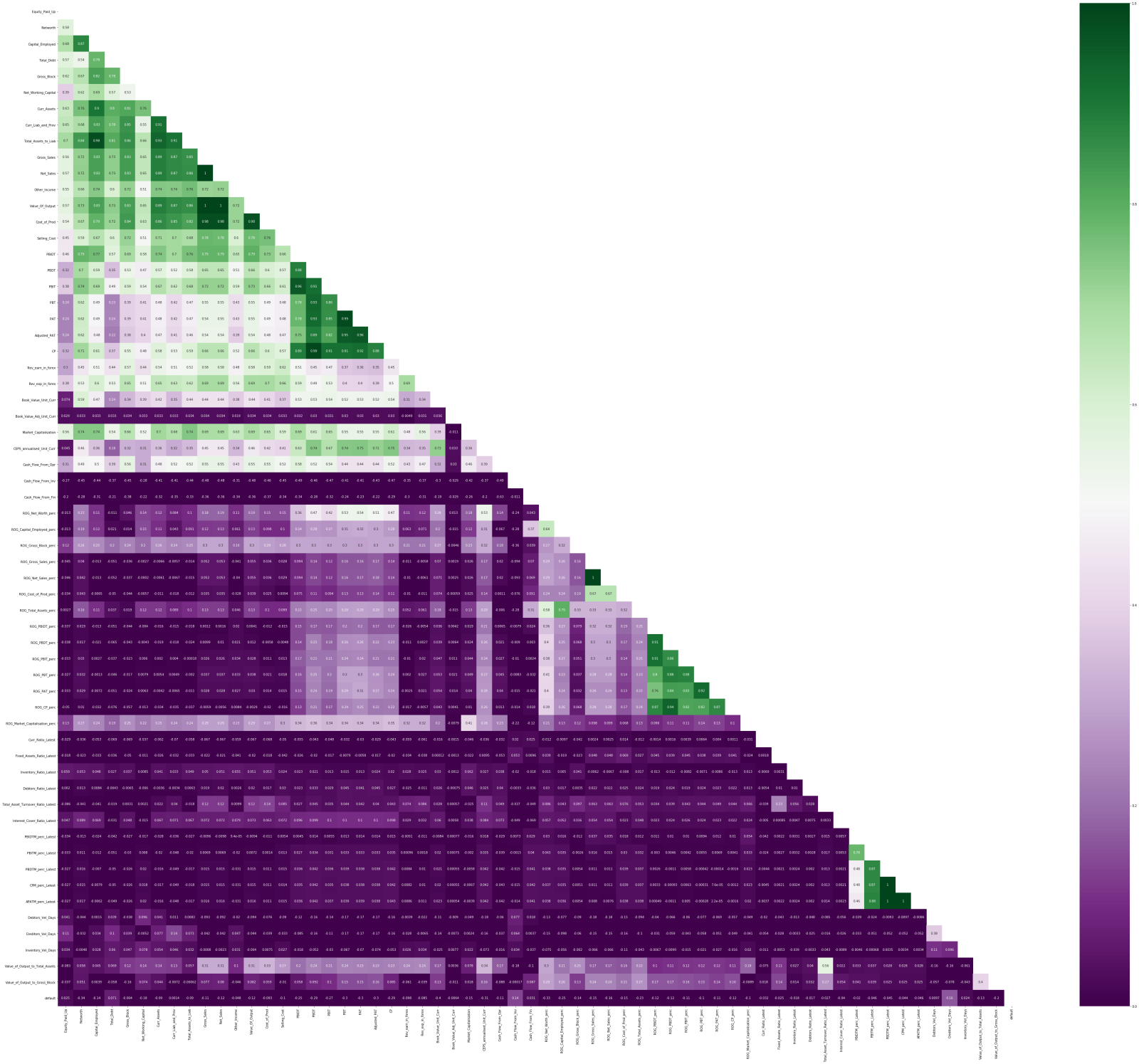


Fig – 1.15 Multivariate analysis of plotting correlation in heatmap after dropping insignificant variable.

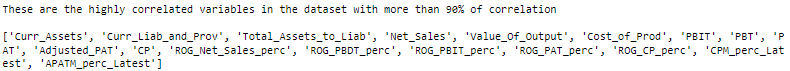


Fig – 1.16 Taking Highly correlated variables.

### 1.5 Train Test Split

The shape of the dataset after splitting the train and test data. The train data and test data are splitted in the ratio of 67:33 with the random state – 42.

|  |  |
| --- | --- |
| Fig – 1.17 X-train shape. | Fig – 1.18 X-test shape. |
| Fig – 1.19 Y-train shape | Fig – 1.20 Y-test shape |

**1.6 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.**

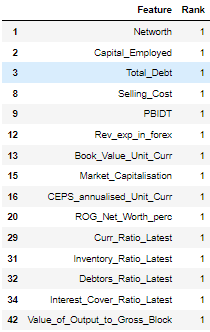


Fig – 1.21 Selecting the feature with rank 1

***Model - 1***

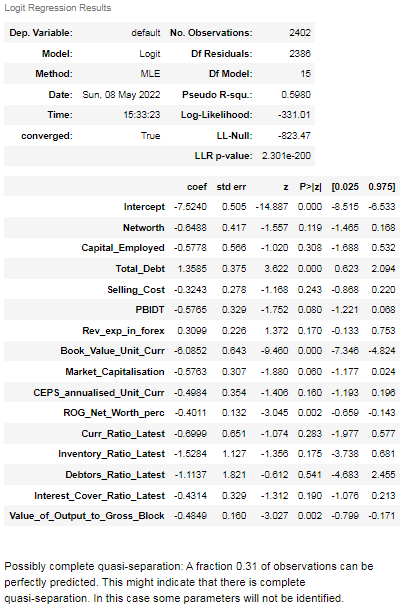


Fig – 1.22 Model-1 summary report

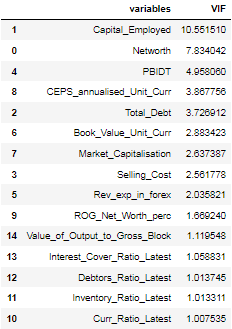


Fig – 1.23 Variance Inflation Factor.of Model-1

The capital Employed has the highest vif and p-value is greater than the alpha value(0.05), capital\_employed variable is dropped.

***Model-2***

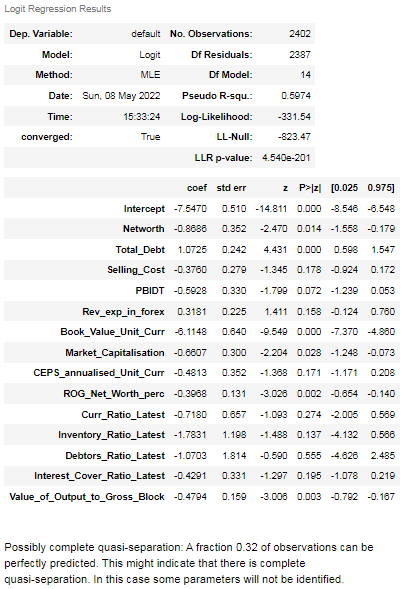


Fig – 1.24 Model-2 Summary Report

###### Debtors\_Ratio\_Latest has the highest p-value and is insignificant, therefore, we need to eliminate it.

##### ***Model 3:***

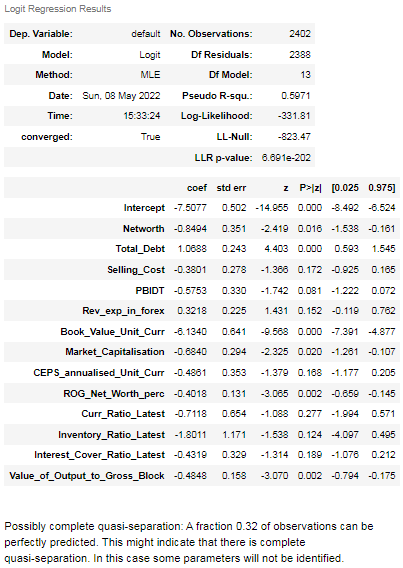


Fig – 1.25 Model-3 Summary Report

###### Curr\_Ratio\_Latest has the highest p-value and is insignificant, therefore, we need to eliminate it.

##### ***Model 4:***

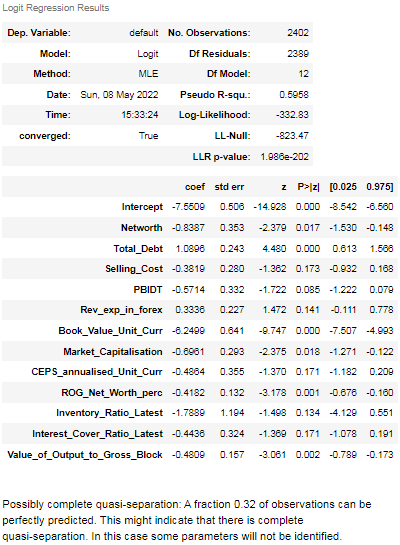


Fig – 1.26 Model-4 Summary Report

Selling\_Cost has the highest p-value and is insignificant, therefore, we need to eliminate it.

Model – 5:

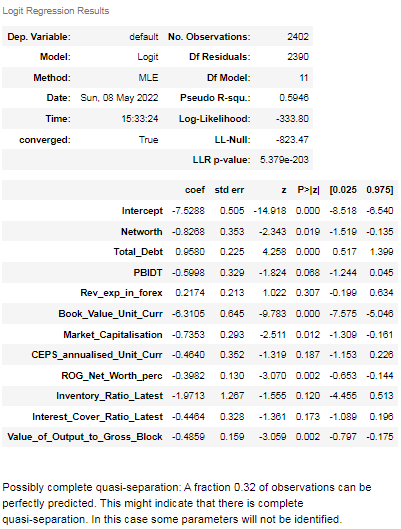


Fig – 1.27 Model-5 Summary Report

Rev\_exp\_in\_forex has the highest p-value and is insignificant, therefore, we need to eliminate it.

###### **Model 6:**

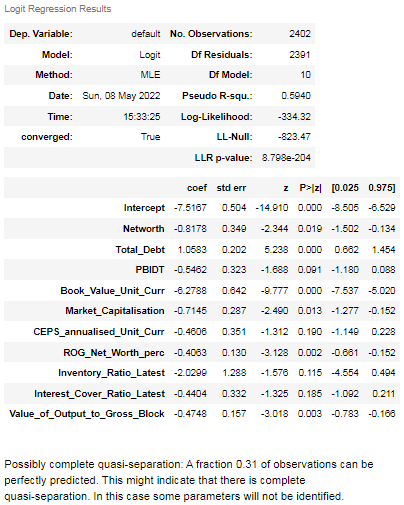


Fig – 1.28 Model-6 Summary Report

CEPS\_annualised\_Unit\_Curr has the highest p-value and is insignificant, therefore, we need to eliminate it.

###### **Model 7:**

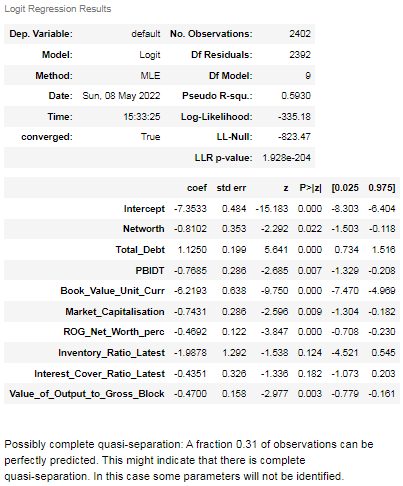


Fig – 1.29 Model-7 Summary Report

Interest\_Cover\_Ratio\_Latest has the highest p-value and is insignificant, therefore, we need to eliminate it.

##### ***Model 8:***

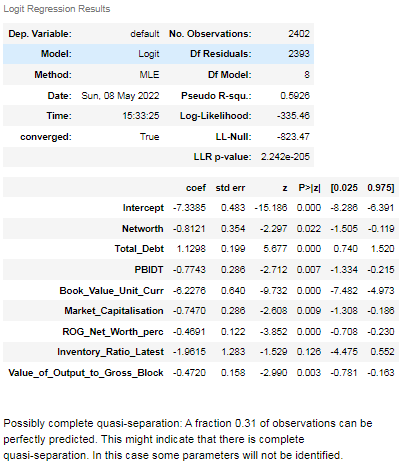


Fig – 1.30 Model-8 Summary Report

Inventory\_Ratio\_Latest has the highest p-value and is insignificant, therefore, we need to eliminate it.

***Model – 9:***

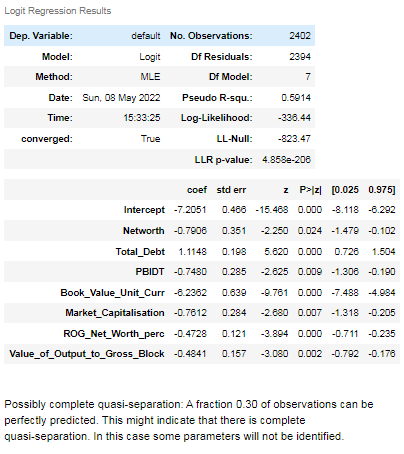


Fig – 1.31 Model-9 Summary Report

### Now, all the variables are significant and p-value is less than the alpha value 0.05. Therefore, we don't need to eliminate the other variables.

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



Fig – 1.32 Optimum threshold

|  |  |
| --- | --- |
| Fig – 1. 33 Confusion matrix for train data | Fig – 1.34 Confusion matrix for test data |

|  |  |
| --- | --- |
| Fig – 1. 35 Classification report for train data | Fig – 1. 36 Classification report for test data |

From the train data and test data we can infer that recall is good for both training and test data classification report .

The test data has 89.8% recall that company might default.

The precision of test data is slightly greater than the train data, test data is slightly over fitting.