Business Report

SMDM Project Business Report DSBA

***Sanjay Srinivasan***

***PGP-DSBA Online***

***JULY’ 21 Batch***

***Date: 15-05-2022***

# *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 | 23 |
|  | 8) Build a Random Forest Model on Train Dataset. Also showcase your model building approach | 24 |
|  | 9) Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model | 28 |
|  | 10) Build a LDA Model on Train Dataset. Also showcase your model building approach | 30 |
|  | 11) Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model | 33 |
|  | 12) Compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve) | 35 |
|  | 13) State Recommendations from the above models | 36 |

***List Of Figures***

|  |  |  |
| --- | --- | --- |
| ***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 |
| 1.37 | Initializing Random Forest Classifier | 24 |
| 1.38 | Taking features with Rank 1 | 24 |
| 1.39 | value count of the target column. | 24 |
| 1.40 | Initializing Model-1 using RF model | 24 |
| 1.41 | Model-1 Summary. | 25 |
| 1.42 | Model-1 Variation Inflation Matrix (VIF) | 25 |
| 1.43 | Initializing Model-2 | 26 |
| 1.44 | Model-2 Summary. | 26 |
| 1.45 | Initializing Model-3 | 26 |
| 1.46 | Model-3 Summary | 27 |
| 1.47 | Initializing Model-4 | 27 |
| 1.48 | Model-4 Summary | 28 |

|  |  |  |
| --- | --- | --- |
| 1.49 | Boxplot for Default variable. | 28 |
| 1.50 | Optimum threshold value | 28 |
| 1.51 | Predicted train values | 29 |
| 1.52 | Predicted test values | 29 |
| 1.53 | Boxplot for test values | 29 |
| 1.54 | Optimum threshold values for test data | 29 |
| 1.55 | Confusion matrix train values | 30 |
| 1.56 | Confusion matrix test values | 30 |
| 1.57 | Classification report for train data | 30 |
| 1.58 | Classification report for test data | 30 |
| 1.59 | ROC for train data | 30 |
| 1.60 | AUC score for train data | 30 |
| 1.61 | ROC for test data | 30 |
| 1.62 | AUC score for test data | 30 |
| 1.63 | Initializing LDA model | 30 |
| 1.64 | Taking features with rank 1 for LDA model | 30 |
| 1.65 | Value count for default variable | 31 |
| 1.66 | Model 1 Initializing | 31 |
| 1.67 | Model -1 summary | 31 |
| 1.68 | Model -1 VIF | 32 |
| 1.69 | Model – 2 Intializing | 32 |
| 1.70 | Model – 2 Summary | 32 |
| 1.71 | Intializing Model -3 | 33 |
| 1.72 | Model -3 Summary | 33 |
| 1.73 | Default value for LDA train model | 33 |
| 1.74 | Optimum threshold value for LDA train model | 34 |
| 1.75 | Predicted value for LDA train model | 34 |
| 1.76 | Default value for LDA test model | 34 |
| 1.77 | Predicted value for LDA test model | 34 |
| 1.78 | Confusion matrix train values | 35 |
| 1.79 | Confusion matrix test values | 35 |
| 1.80 | Classification report for train data | 35 |
| 1.81 | Classification report for test data | 35 |
| 1.82 | ROC for train data | 35 |
| 1.83 | AUC score for train data. | 35 |
| 1.84 | ROC for test data | 35 |
| 1.85 | AUC score for test data. | 35 |
| 1.86 | Comparison dataframe for LR,RF and LDA values. | 35 |
| 1.87 | ROC curve for LR model | 36 |
| 1.88 | ROC curve for RF model | 36 |
| 1.89 | ROC curve for LDA model | 36 |

# Problem - 1

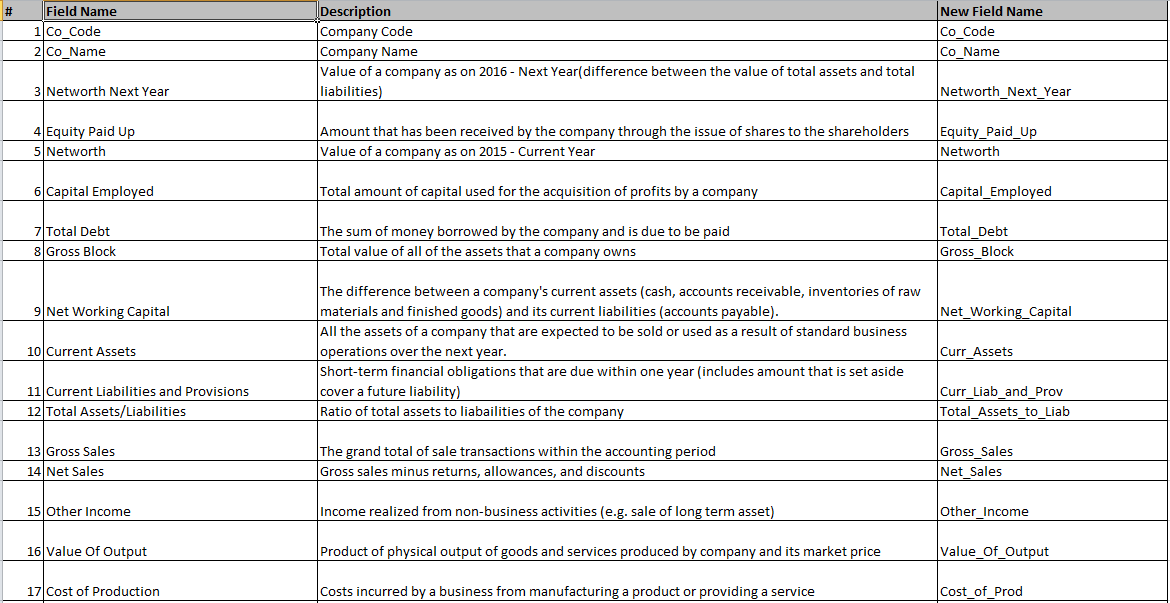
*Summary*

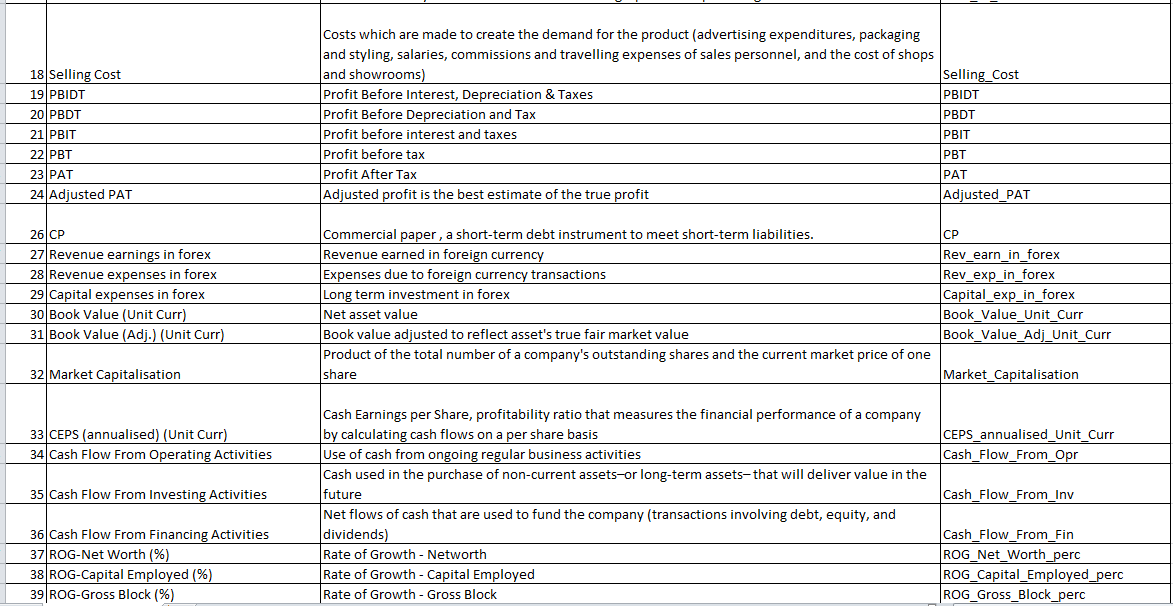
The data is gathered based on the company financial balance sheet, which deals with the company finances. This dataset has financial statements for 3586 company with 67 variables. For investing in the company, to analyse from the investor’s point of view, to predict that the company is capable of handling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

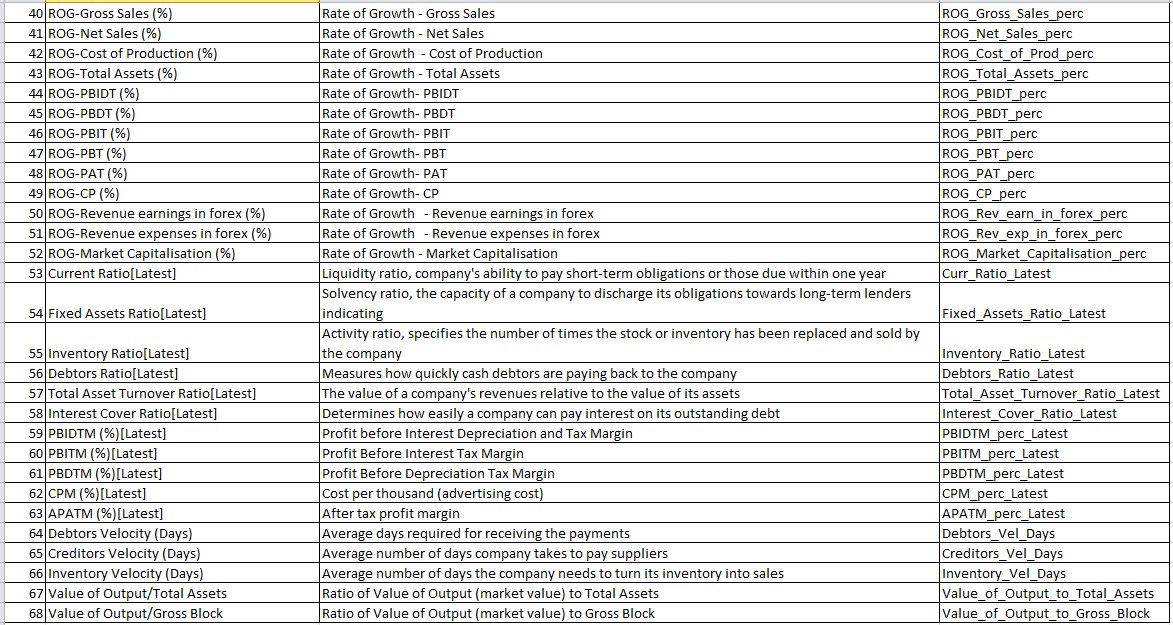
*Introduction*

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:*

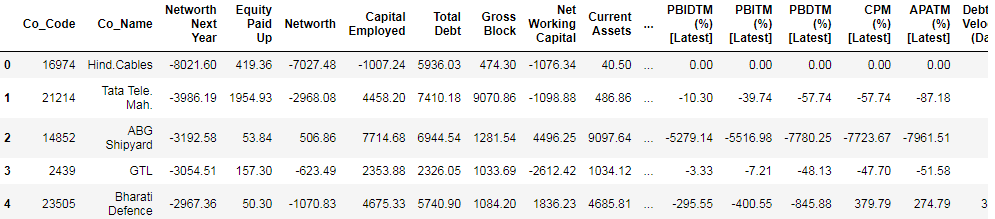


Fig 1.1 Dataset Sample Before Changing Column Names

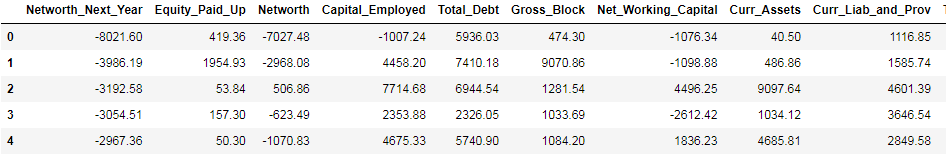


Fig 1.2 Dataset Sample After Changing Column Names

*Exploratory Data Analysis*

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

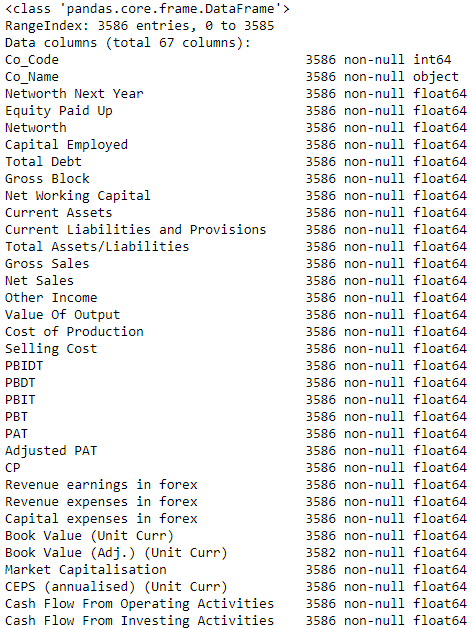


Fig- 1.3. Sample Datatypes of the variable with null values

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

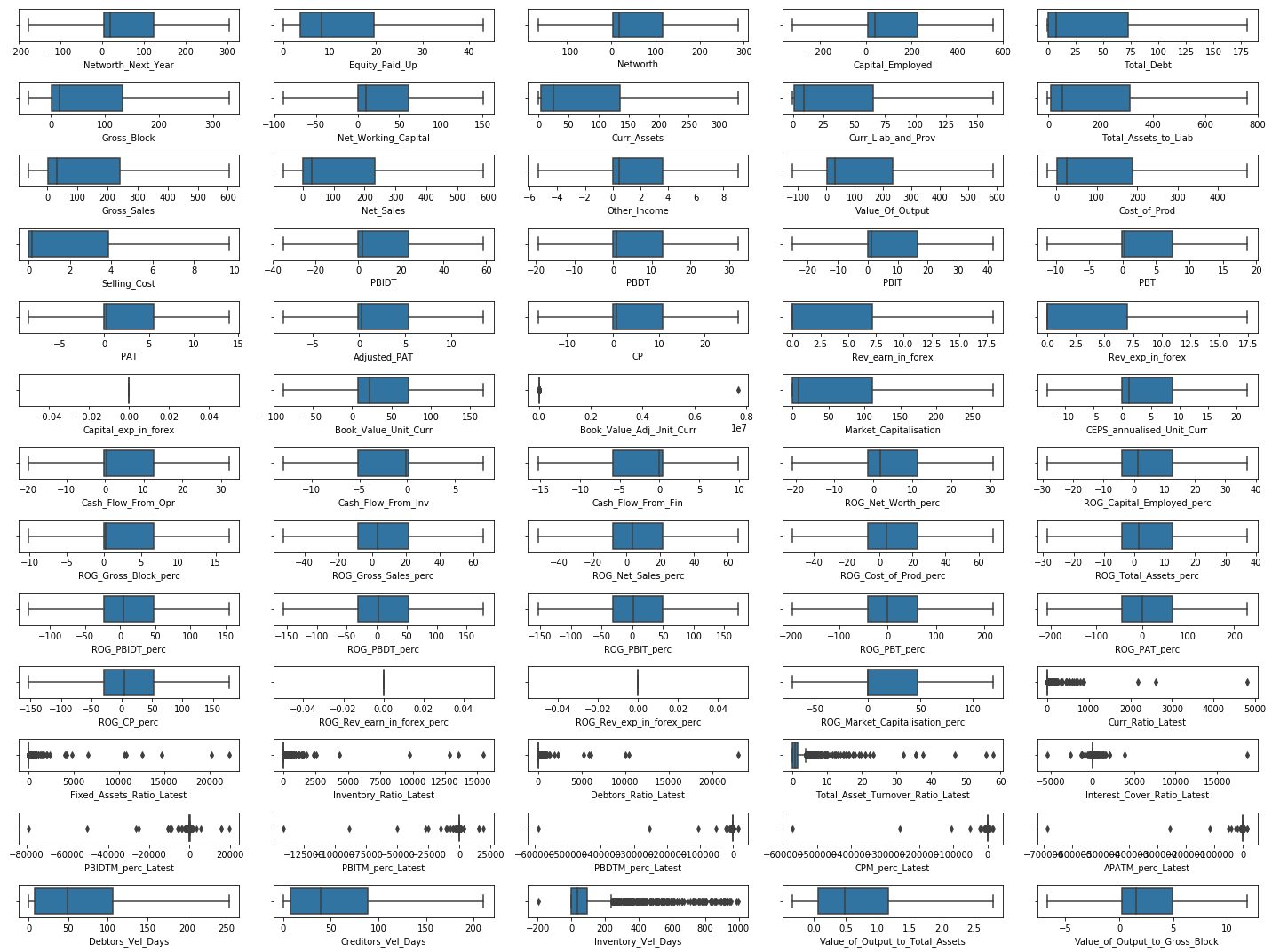


Fig- 1.5 Shape After Outliers Treatment

**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. |

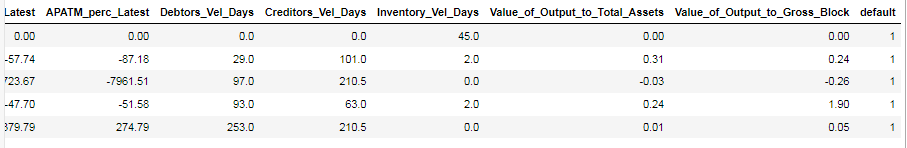


Fig – 1.10 Sample data after Transformation.

**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:***

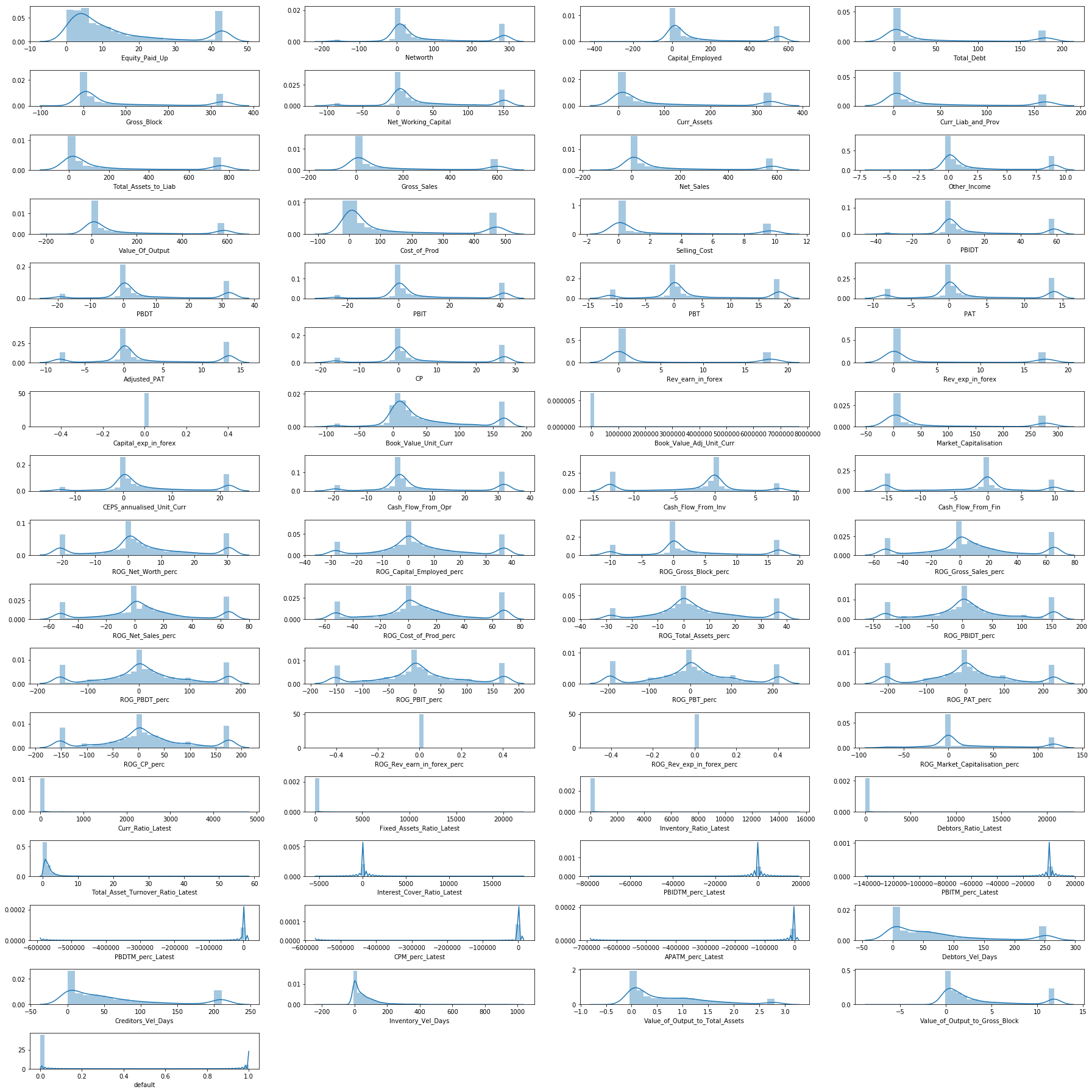


Fig – 1.11 Univariate Analysis

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:***

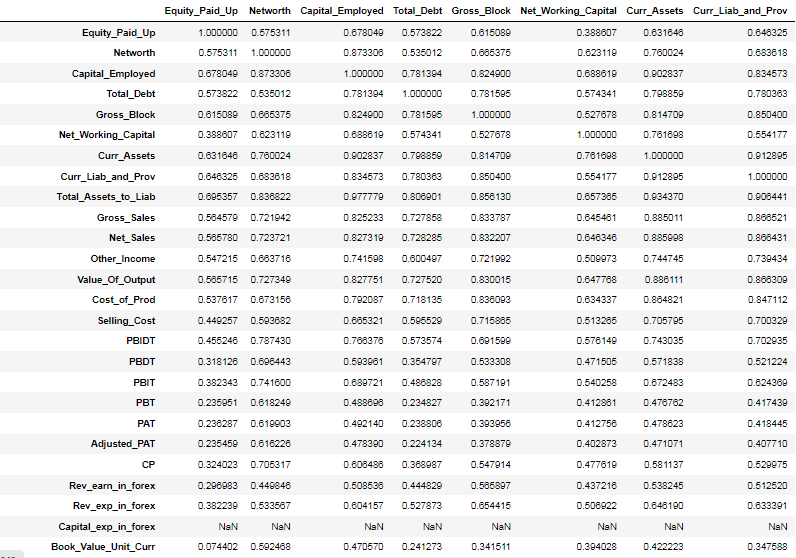


Fig – 1.13 Sample Multivariate analysis for correlation

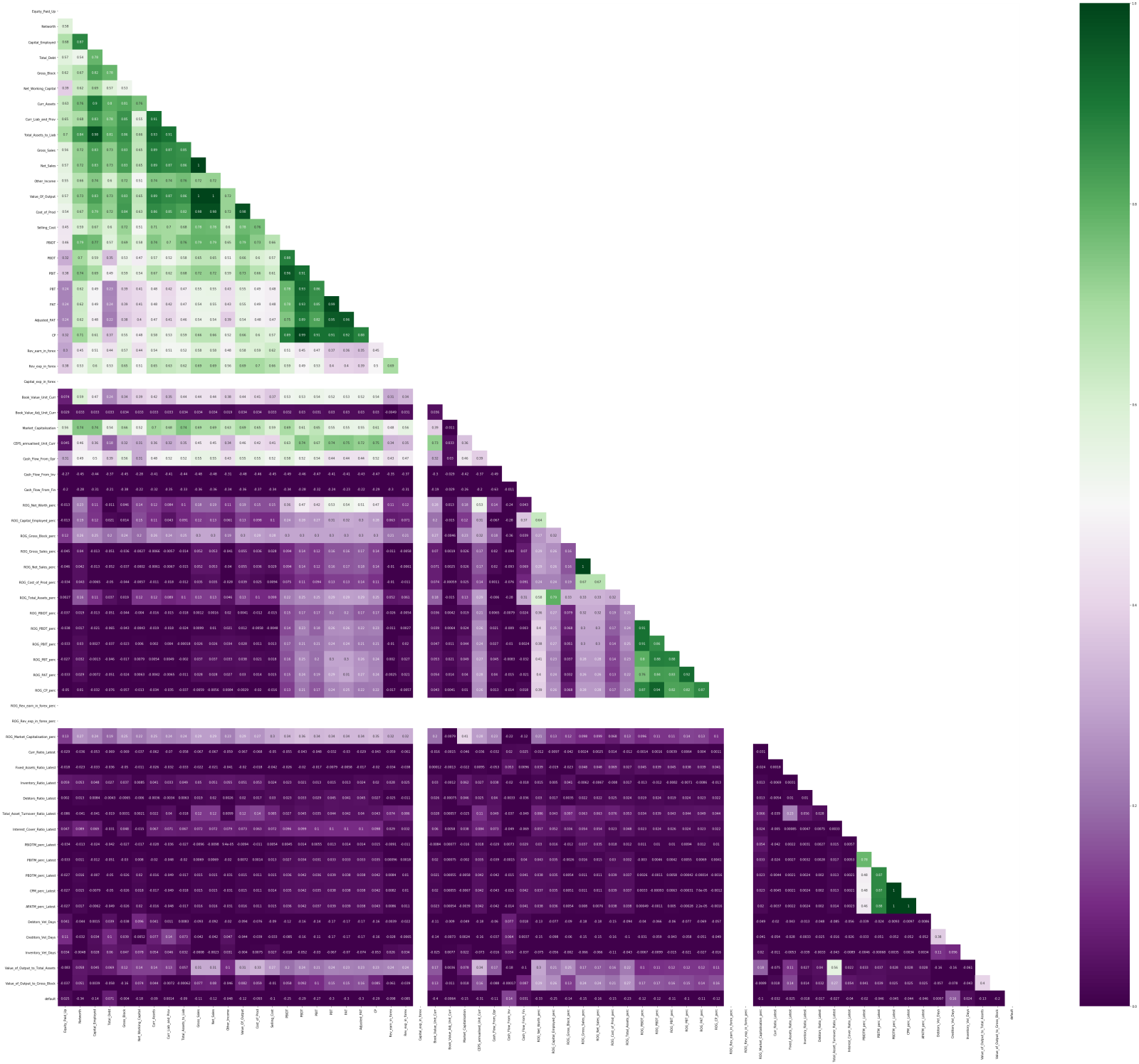


Fig – 1.14 Multivariate analysis of plotting correlation in heatmap

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.

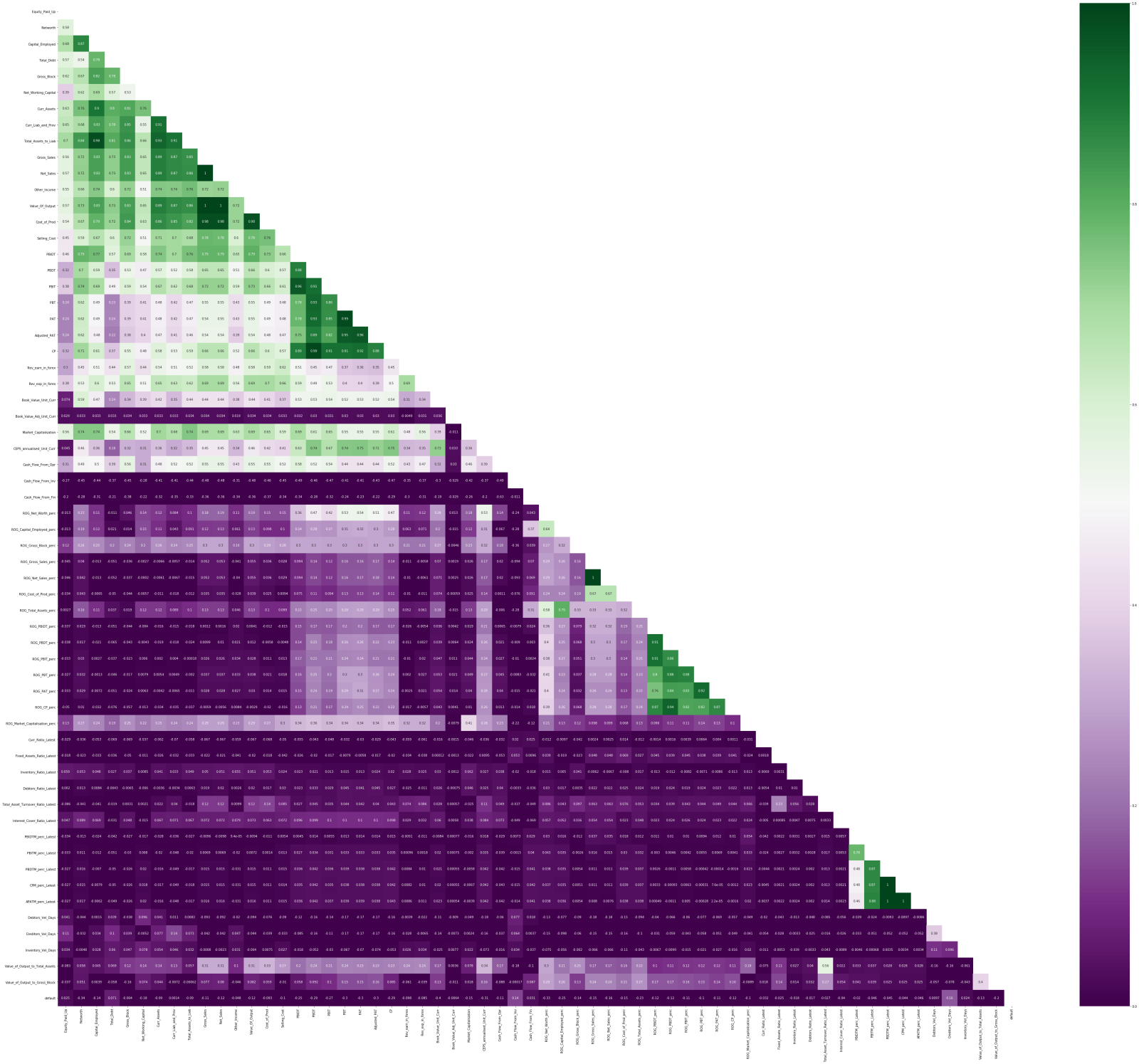


Fig – 1.15 Multivariate analysis of plotting correlation in heatmap after dropping insignificant variable.

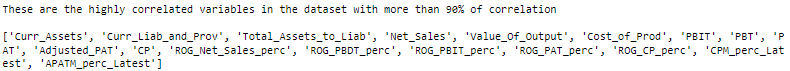


Fig – 1.16 Taking Highly correlated variables.

### 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.**

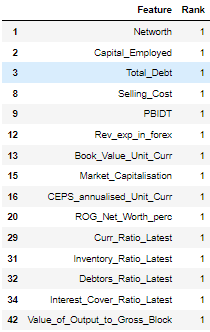


Fig – 1.21 Selecting the feature with rank 1

***Model - 1***

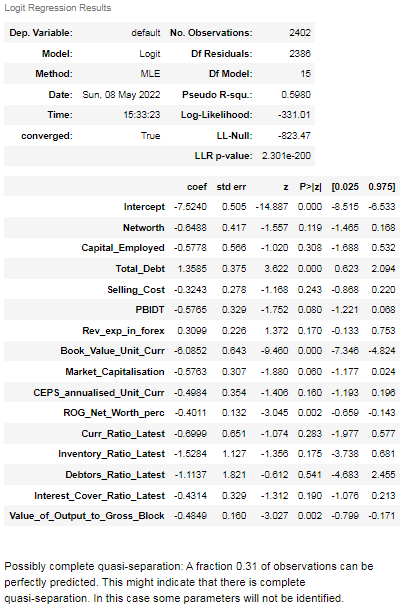


Fig – 1.22 Model-1 summary report

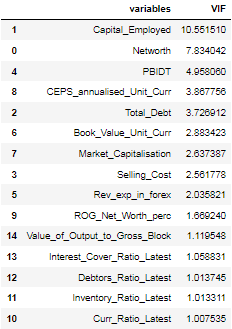


Fig – 1.23 Variance Inflation Factor.of Model-1

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***

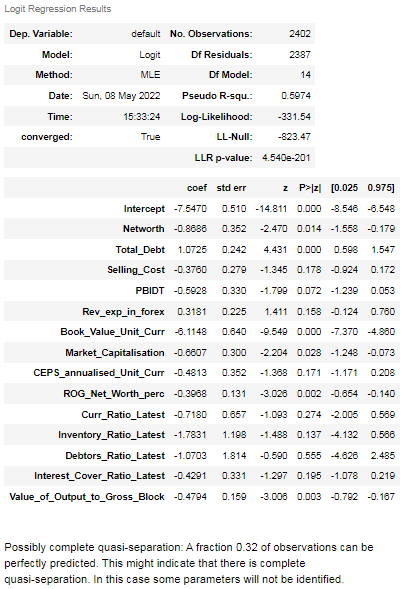


Fig – 1.24 Model-2 Summary Report

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

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

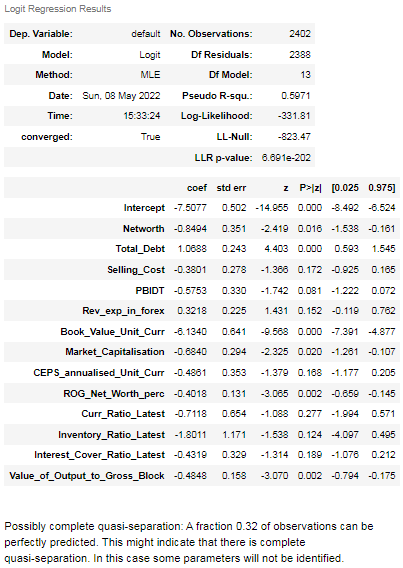


Fig – 1.25 Model-3 Summary Report

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

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

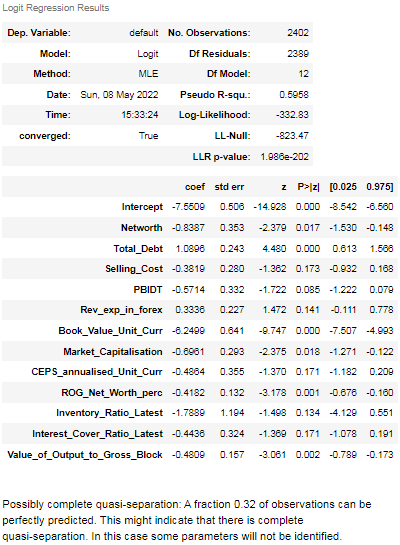


Fig – 1.26 Model-4 Summary Report

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

Model – 5:

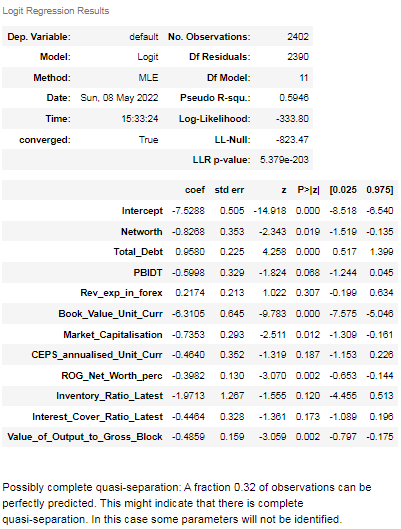


Fig – 1.27 Model-5 Summary Report

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

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

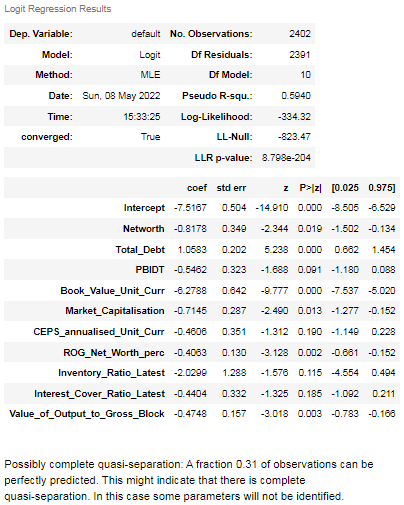


Fig – 1.28 Model-6 Summary Report

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

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

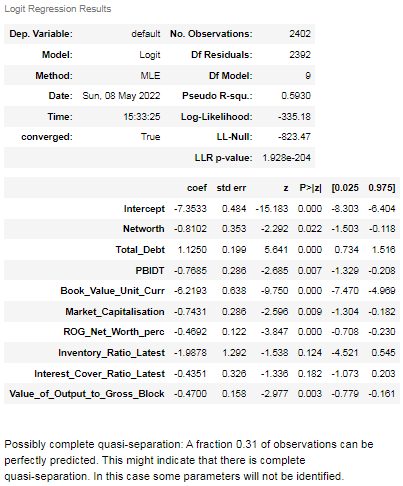


Fig – 1.29 Model-7 Summary Report

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

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

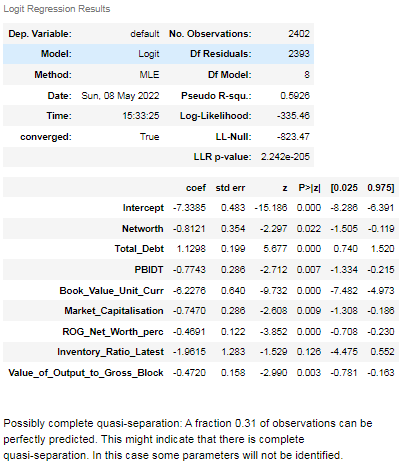


Fig – 1.30 Model-8 Summary Report

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

***Model – 9:***

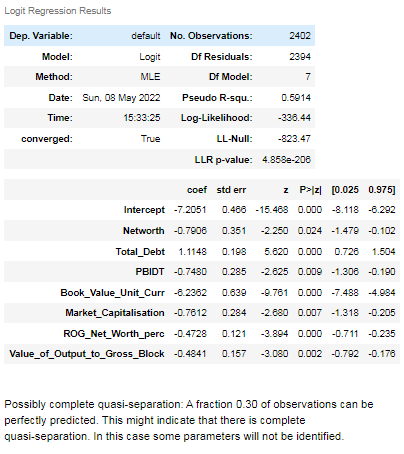


Fig – 1.31 Model-9 Summary Report

### 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.

**1.8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach.**



Fig – 1. 37 Initializing Random Forest Classifier



Fig – 1.38. Taking features with Rank 1



Fig – 1.39 value count of the target column.

***Model Building of Random forest model:***

***Model - 1***



Fig – 1.40 Initializing Model-1 using RF model

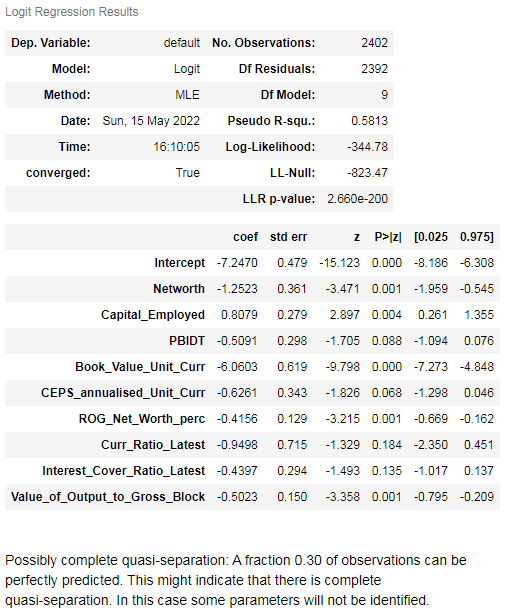


Fig – 1.41 Model-1 Summary.



Fig – 1.42 Model-1 Variation Inflation Matrix (VIF)

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

***Model - 2***



Fig – 1.43 Initializing Model-2

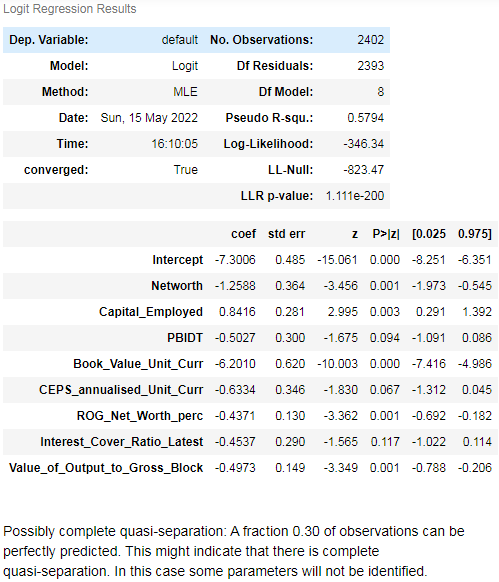


Fig – 1.44 Model-2 Summary

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

***Model - 3***



Fig – 1.45 Initializing Model-3

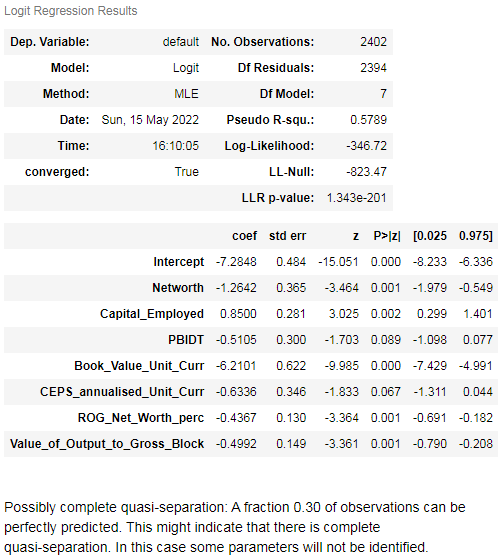


Fig – 1.46 Model-3 Summary

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

***Model - 4***



Fig – 1.47 Initializing Model-4

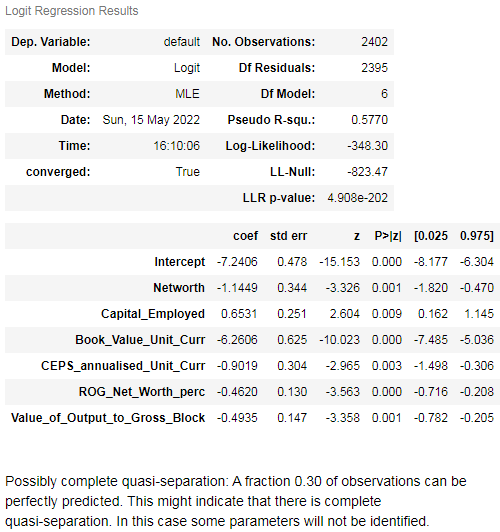


Fig – 1.48 Model-4 Summary

All features are having p-value less than the alpha value. So model building for the Random Forest is over.

**1.9 Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model.**

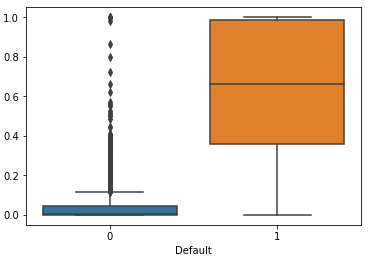


Fig – 1.49 Boxplot for Default variable.



Fig – 1.50 Optimum threshold value

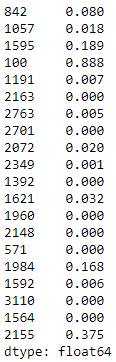


Fig – 1.51 Predicted train values

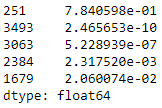


Fig – 1.52 Predicted test values

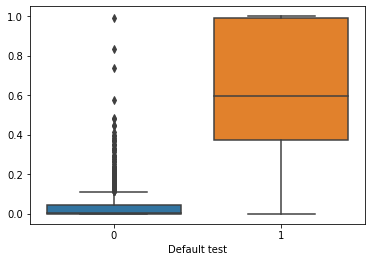


Fig – 1.53 Boxplot for test values



Fig – 1.54 Optimum threshold values for test data

|  |  |  |
| --- | --- | --- |
| Table | Train data | Test data |
| Confusion matrix | Fig – 1.55 Confusion matrix train values | Fig – 1.56 Confusion matrix test values |
| Classification report | Fig – 1.57 Classification report for train data | Fig – 1.58 Classification report for test data |
| AUC  ROC Curve | Fig – 1.59 ROC for train data    Fig 1.60 AUC score for train data. | Fig – 1.61 ROC for train data    Fig 1.62 AUC score for test data. |

**1.10 Build a LDA Model on Train Dataset. Also showcase your model building approach.**



Fig 1.63 Initializing LDA model

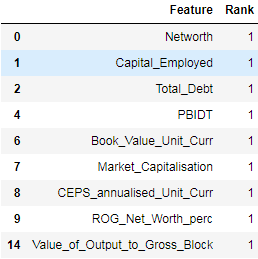
.

Fig 1.64 Taking features with rank 1 for LDA model



Fig 1.65 Value count for default variable

***Model – 1***



Fig 1.66 Model 1 Initializing

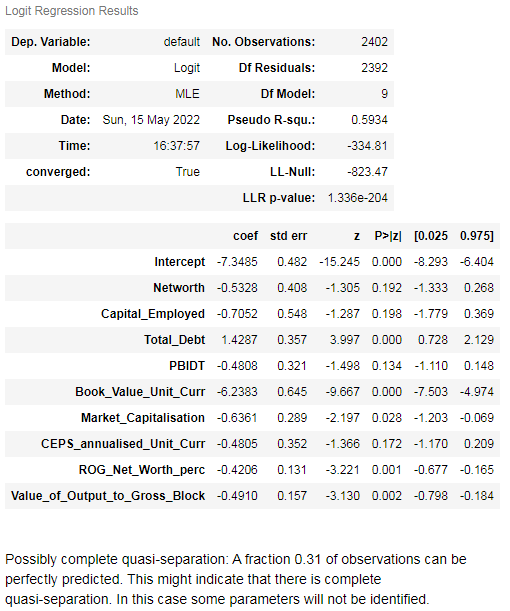


Fig 1.67 Model -1 summary

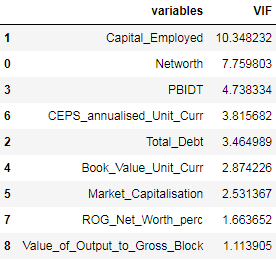


Fig 1.68 Model -1 VIF

Capital\_Employed has the highest VIF value and is insignificant, therefore, we need to eliminate it.

***Model – 2***



Fig 1.69 Model – 2 Intializing

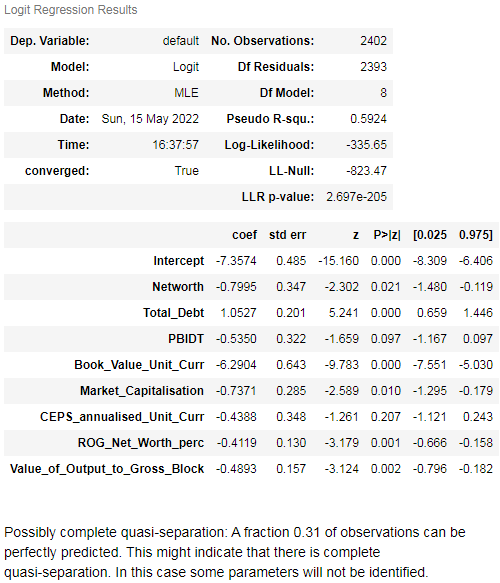


Fig 1.70 Model – 2 Summary

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

***Model – 3***



Fig 1.71 Intializing Model -3

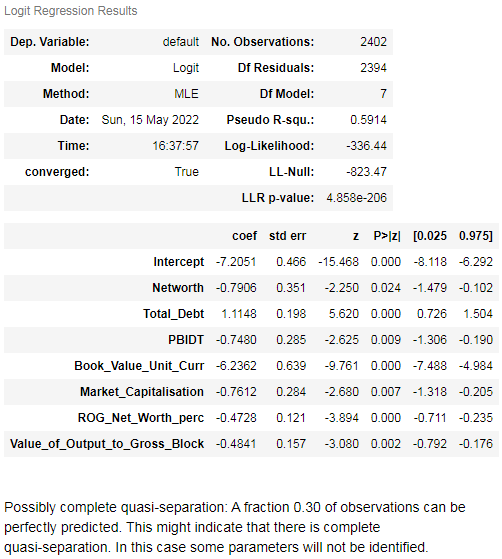


Fig 1.72 Model -3 Summary

All the features are having p-value less than the alpha value. So model building for the LDA model is completed.

**1.11 Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model.**

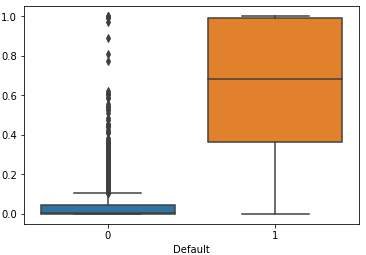


Fig 1.73 Default value for LDA train model



Fig 1.74 Optimum threshold value for LDA train model

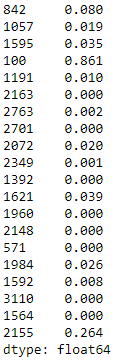


Fig 1.75 Predicted value for LDA train model

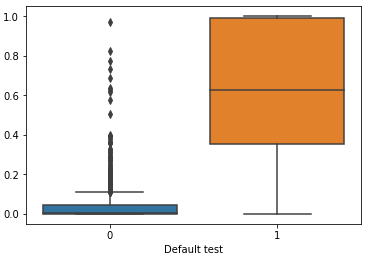


Fig 1.76 Default value for LDA test model

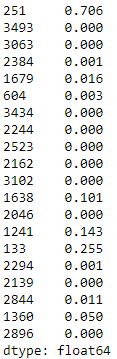


Fig 1.77 Predicted value for LDA test model

|  |  |  |
| --- | --- | --- |
| Table | Train data | Test data |
| Confusion matrix | Fig – 1.78 Confusion matrix train values | Fig – 1.79 Confusion matrix test values |
| Classification report | Fig – 1.80 Classification report for train data | Fig – 1.81 Classification report for test data |
| AUC  ROC Curve | Fig – 1.82 ROC for train data    Fig 1.83 AUC score for train data. | Fig – 1.84 ROC for train data    Fig 1.85 AUC score for test data. |

**1.12 Compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve).**

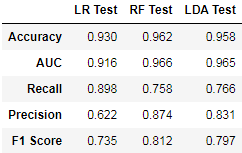


Fig 1.86 Comparison dataframe for LR,RF and LDA values.

|  |  |  |
| --- | --- | --- |
| ***LR ROC Curve*** | ***RF ROC Curve*** | ***LDA ROC Curve*** |
| Fig 1.87 ROC curve for LR model | Fig 1.88 ROC curve for RF model | Fig 1.89 ROC curve for LDA model |

From the above dataframe, Recall is higher in logistic Regression, precision is better in Random Forest classifier.

***Overall Random Forest is better algorithm.***

**1.13 State Recommendations from the above models**

From the above model, Random Forest model is the best model with higher precision and recall.

Company with the following details will lead the investor to invest in the company are

1. Increase in debtor’s shows company turnover has increased.

2. Increase in debtor’s / decrease in creditors will lead to get fresh loan with lower interest rate with good credit rating for the company.

3. Reduction in creditors shows that the company follows the strict/disciplined payment terms.

4. Change in debt equity ratios shows that the company is growing.

5. Growth in current asset and decrease in current liability.

6. Increase in Net worth will help the company to provide good dividends for the share/stake holders.