FRA Milestone 1 Project

PGP – DSBA

May ‘22

Date: 26/03/2022

Swaathi Ramakrishnan

**Table of Contents**

|  |  |
| --- | --- |
| **Contents** | **Page Number** |
|  |  |
| Problem Statement | 4 |
|  |  |
| 1.1 Outlier Treatment | 5 |
|  |  |
| 1.2 Missing Value Treatment | 9 |
|  |  |
| 1.3 Transform Target variable into 0 and 1 | 11 |
|  |  |
| 1.4 Univariate & Bivariate analysis with proper interpretation. | 11 |
|  |  |
| 1.5 Train Test Split | 19 |
|  |  |
| 1.6 Logistic Regression Model | 20 |
|  |  |
| 1.7 Validating the Model on Test Dataset,  showcasing the performance matrices and Interpretation | 27 |
|  |  |

**List of Tables/Figures**

|  |  |
| --- | --- |
| Fig 1. Data Head | 5 |
| Fig 2. Cleansing the column names | 5 |
| Fig 3. Dataset Information | 6 |
| Fig 4. Outliers in the dataset | 7 |
| Fig 5. Sum of outliers in few columns | 7 |
| Fig 6. Boxplot after treating outliers | 8 |
| Fig 7. Heatmap – missing data | 8 |
| Fig 8. Missing values in the dataset per observation | 9 |
| Fig 9. After imputing with KNN Imputer | 9 |
| Fig 10. Correlation Heatmap between independent variables | 10 |
| Fig 11. EDA – Descriptive Statistics | 11 |
| Fig 12. Histogram and Boxplot for Networth\_Next\_Year | 11 |
| Fig 13. Histogram and Boxplot for Equity\_Paid\_Up | 12 |
| Fig 14. Histogram and Boxplot for Networth | 12 |
| Fig 15. Histogram and Boxplot for Capital\_Employed | 12 |
| Fig 16. Histogram and Boxplot for Total\_Debt | 13 |
| Fig 17. Histogram and Boxplot for Gross\_Block\_ | 13 |
| Fig 18. Histogram and Boxplot for Net\_Working\_Capital | 13 |
| Fig 19. Histogram and Boxplot for Current\_Assets\_ | 13 |
| Fig 20. Histogram and Boxplot for Current\_Liabilities\_and\_Provisions\_ | 14 |
| Fig 21. Histogram and Boxplot for Total\_Assets\_to\_Liabilities | 14 |
| Fig 22. Histogram and Boxplot for Gross\_Sales\_ | 14 |
| Fig 23. Histogram and Boxplot for Net\_Sales\_ | 14 |
| Fig 24. Histogram and Boxplot for Other\_Income\_ | 15 |
| Fig 25. Histogram and Boxplot for Value\_Of\_Output | 15 |
| Fig 26. Histogram and Boxplot for Cost\_Of\_Production | 15 |
| Fig 27. Correlation Heatmap | 16 |
| Fig 28. BarPlot – Networth vs Defaulter | 17 |
| Fig 29. BarPlot – Total\_Assets\_to\_Liabilities\_ vs Defaulter | 17 |
| Fig 30. BarPlot – Total\_Debt vs Defaulter | 18 |
| Fig 31. ScatterPlot – Revenue\_Earnings\_in\_Forex vs Revenue\_Expenses\_in\_Forex | 18 |
| Fig 32. Variation\_Inflation\_Factor | 20 |
| Fig 33. Model 1 Summary | 21 |
| Fig 34. Model 2 Summary | 23 |
| Fig 35. Model 3 Summary | 24 |
| Fig 36. Model 4 Summary | 25 |
| Fig 37: Model 1 ROC\_AUC score and ROC curve | 28 |
| Fig 38: Model 2 ROC\_AUC score and ROC curve | 28 |
| Fig 39: Model 3 ROC\_AUC score and ROC curve | 28 |
| Fig 40: Model 4 ROC\_AUC score and ROC curve | 28 |

**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'

Hints :

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.

Test Train Split - Split the data into Train and Test dataset in a ratio of 67:33 and use random\_state =42. Model Building is to be done on Train Dataset and Model Validation is to be done on Test Dataset.

**1.1 Outlier Treatment**

Before starting with outlier treatment, we need to analyse and cleanse the data.

There are 67 columns in the dataset. The column names seem to have some symbols, hence we need to get rid of them by replacing them with ‘\_’.

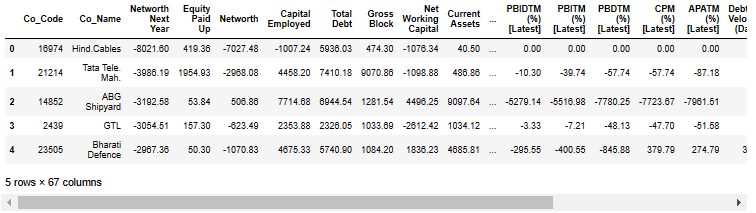
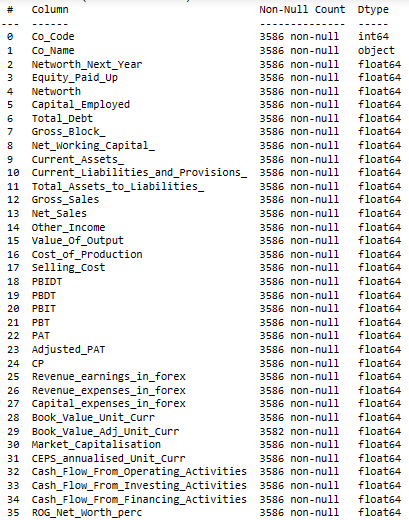


Fig 1. Data head



Fig 2. Cleansing the column names



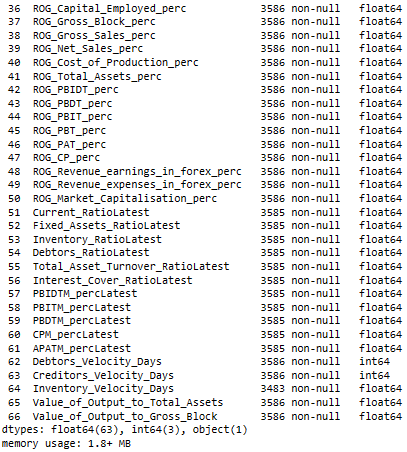
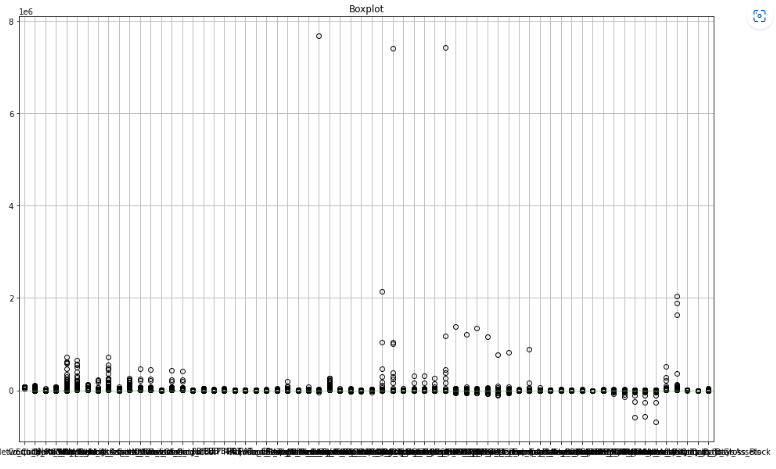


Fig 3. Dataset Information

* The data has 3586 rows and 67 columns.
* Information from data shows us that there are some missing values in the dataset.
* Networth\_Next\_Year variable does not have any missing values.
* Creating the “default” column using this variable with the following condition: when it is > 0, then 0 else 1.
* Proportion of default variable:
  + 0 3198
  + 1 388

After Normalization

* + 0 89 %
  + 1 11 %
* There are many outliers in the dataset.

Fig 4. Outliers in the dataset

For each column we find the Q1 (25th quantile),Q3(75th quantile), IQR(Q3-Q1) values and determine the upper and lower boundary values. Upper boundary is (Q3 + 1.5 \* IQR) and lower boundary is (Q1 – 1.5\*IQR). Any observation above the upper limit or below the lower limit is an outlier.

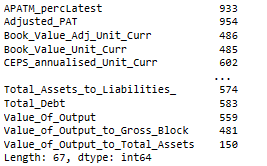


Fig 5. Sum of outliers in few columns

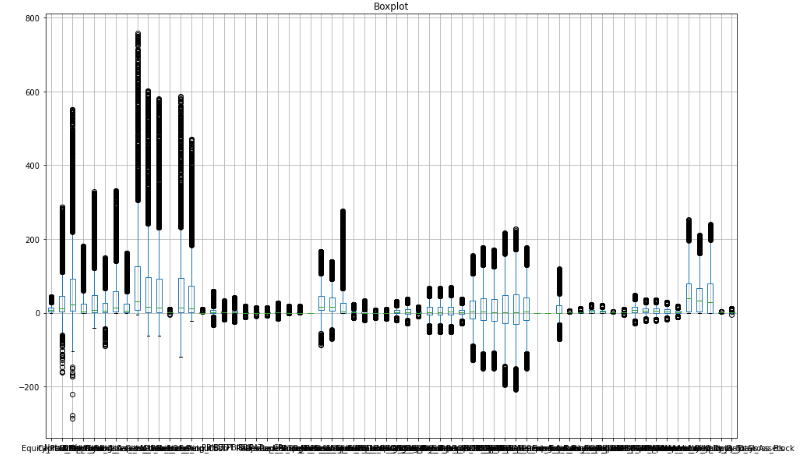


Fig 6. Boxplot after treating outliers

We can see that even after treating outliers, we see some of the columns still having outliers. If we drop them or remove them completely, we might loose the variance in the data. Hence we will substitute them with null and then impute them with KNN imputer and replace the null values for the features we need for the model.

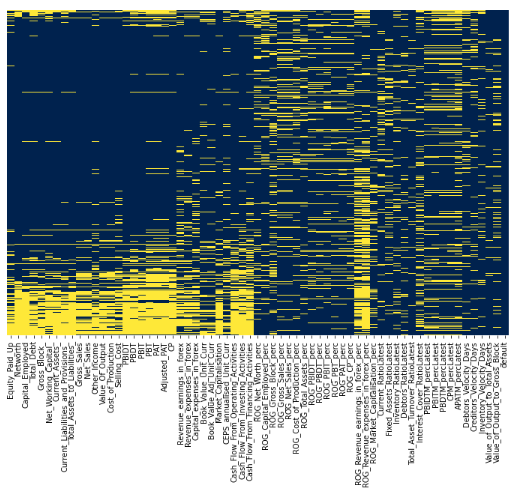


Fig 7. Heatmap – missing data

**1.2 Missing Value Treatment**

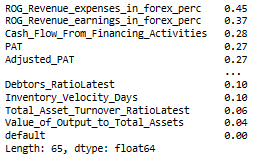


Fig 8. Missing values in the dataset per observation

* We can see that there are occurrences of missing data ranging from 0 to 45 per feature.
* We need to retain only observations that has at max 5 missing data.
* This will lead to a subset of original dataset with 1203 observations against 3586 in input data.
* The number of defaulters in this new subset is 118.
* This means that of the original 388 defaulters, we have 118 in the new subset, which leads to a loss of 70% of the determining data.
* Hence the strategy of limiting observations is not a good one.
* Hence we took a look at how many missing values we have against feature instead of observation.
* We removed columns that had more than 30 % missing values
* We scaled the data using standard scaler and the remaining missing values were imputed with KNN Imputer

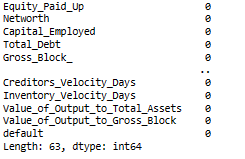


Fig 9. After imputing with KNN Imputer

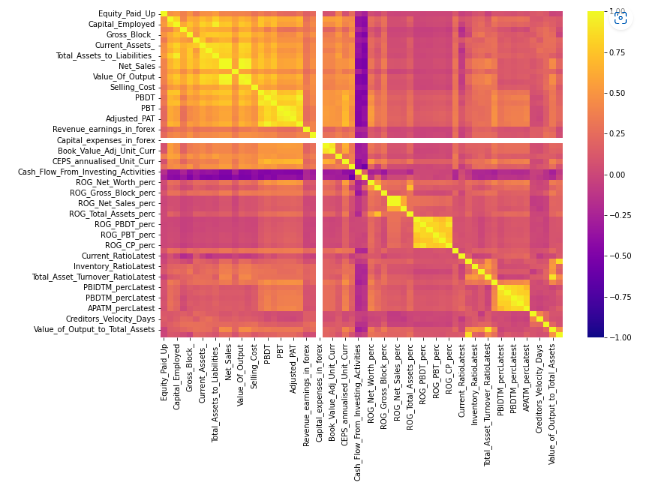


Fig 10. Correlation Heatmap between independent variables

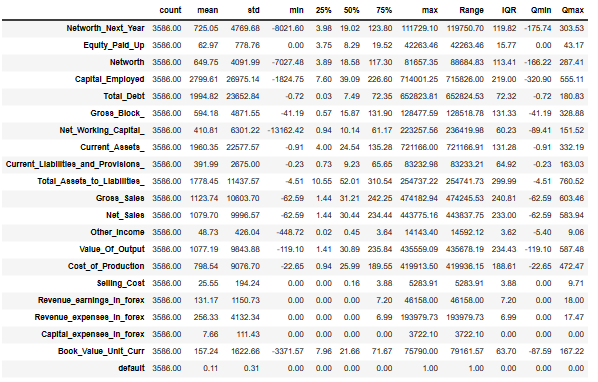
**1.3 Transform Target variable into 0 and 1**

* Networth\_Next\_Year variable does not have any missing values.
* Creating the “default” column using this variable with the following condition: when it is > 0, then 0 else 1.
* Proportion of default variable:
  + 0 3198
  + 1 388

Percentage of default variable

* + 0 89 %
  + 1 11 %

**1.4 Univariate and Bivariate Analysis**

Fig 11. EDA – Descriptive Statistics

Looking at the histogram and boxplot for all the variables. For all of them Distribution is right skewed, and outliers are present.

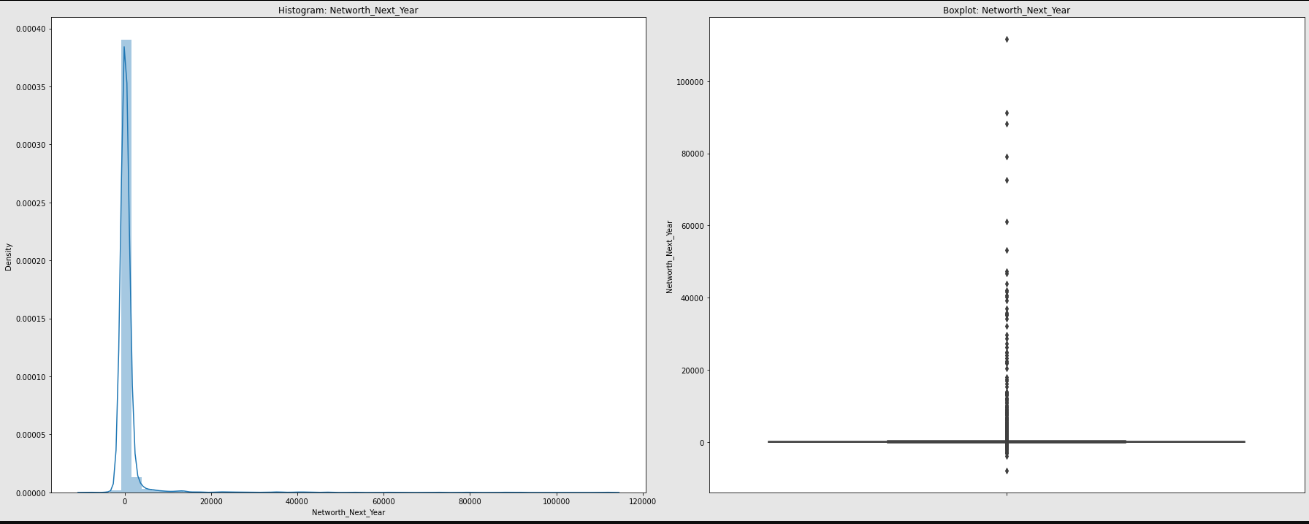


Fig 12. Histogram and Boxplot for Networth\_Next\_Year

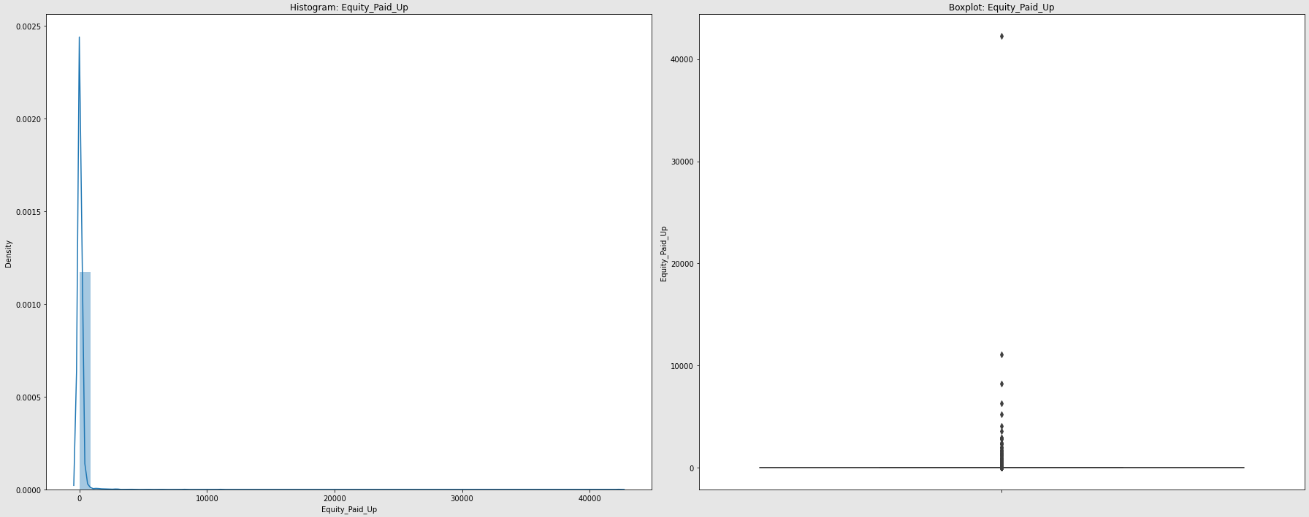


Fig 13. Histogram and Boxplot for Equity\_Paid\_Up

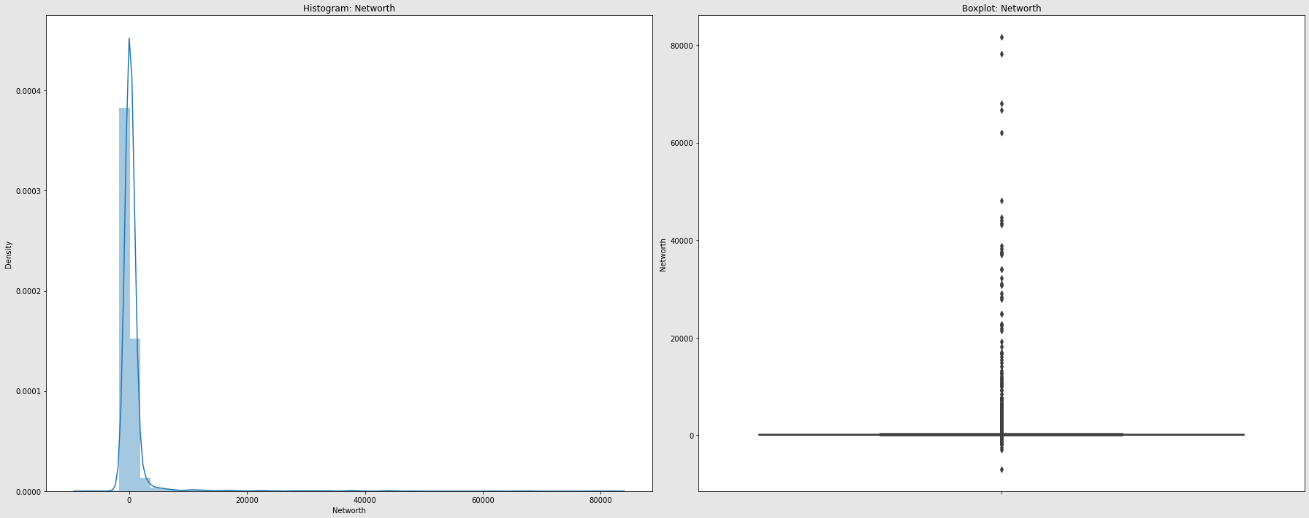


Fig 14. Histogram and Boxplot for Networth

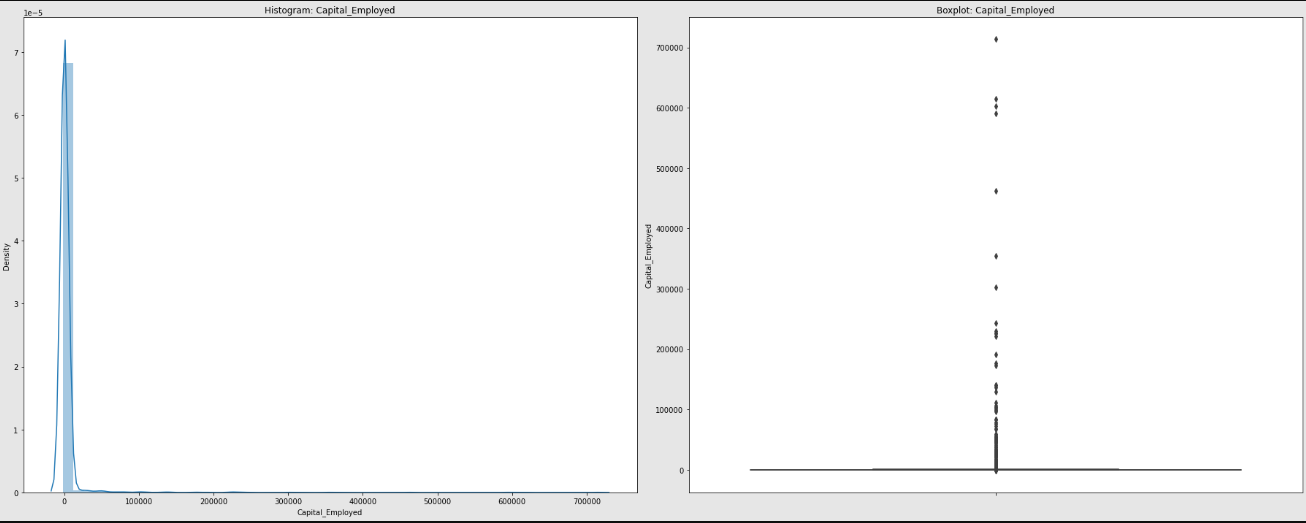
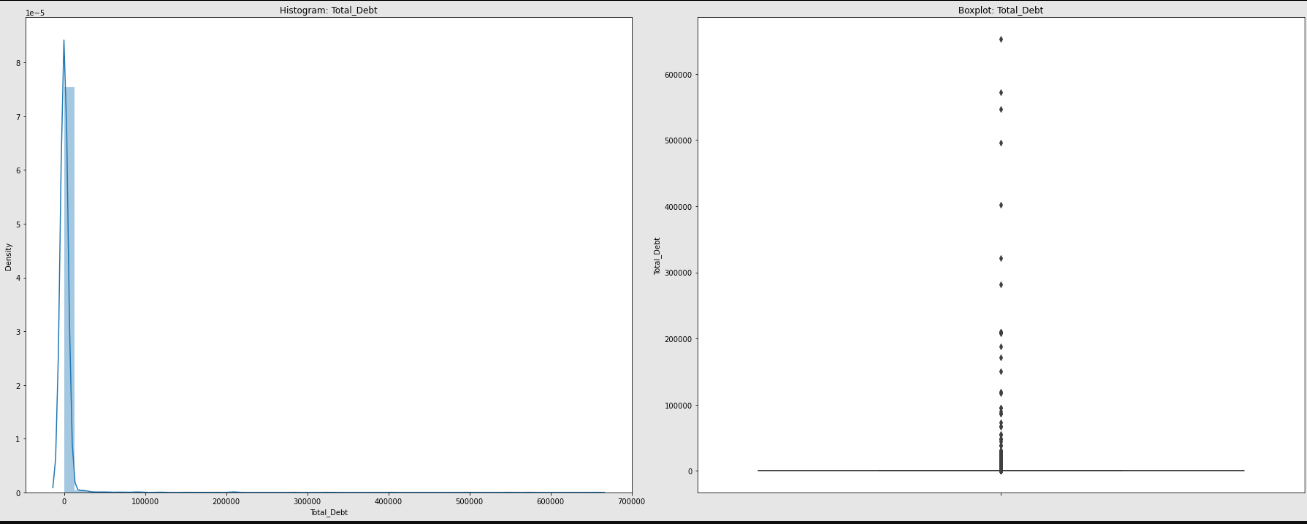
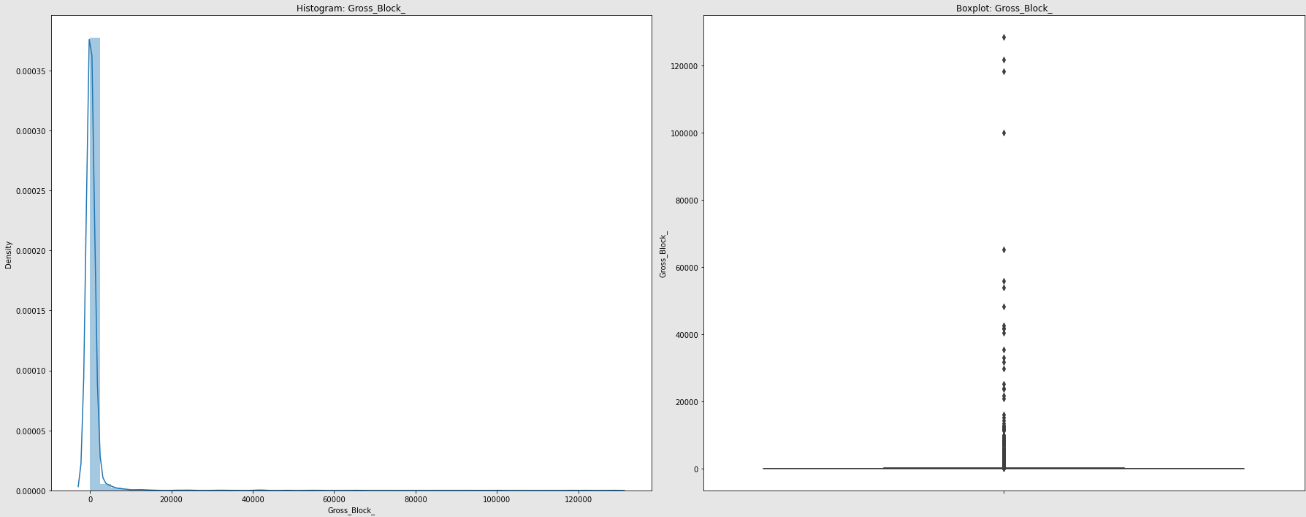


Fig 15. Histogram and Boxplot for Capital\_Employed

Fig 16. Histogram and Boxplot for Total\_Debt

Fig 17. Histogram and Boxplot for Gross\_Block\_

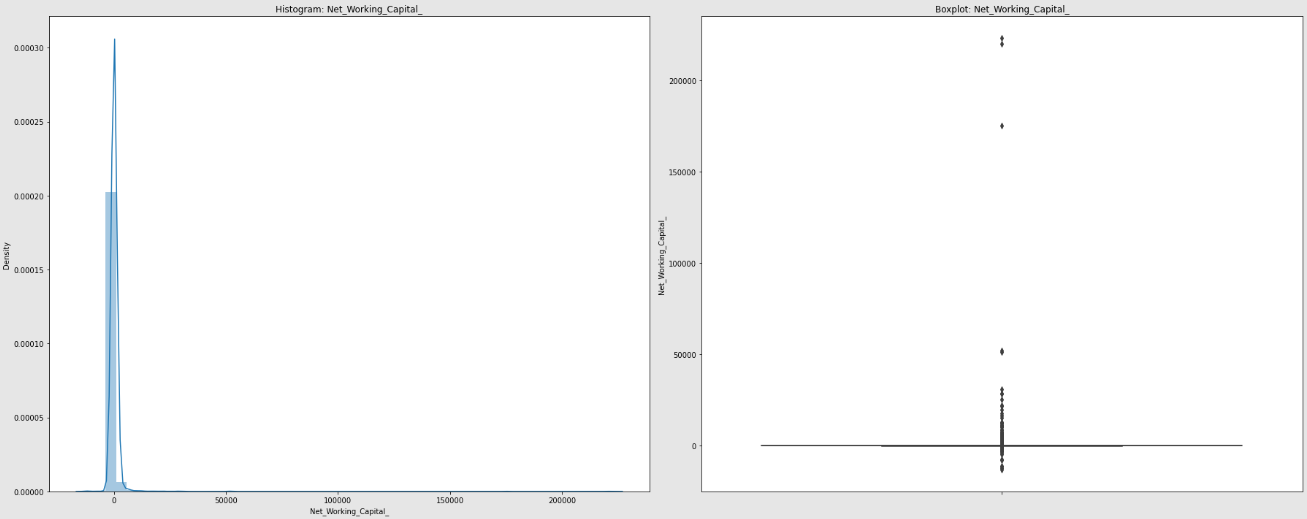


Fig 18. Histogram and Boxplot for Net\_Working\_Capital

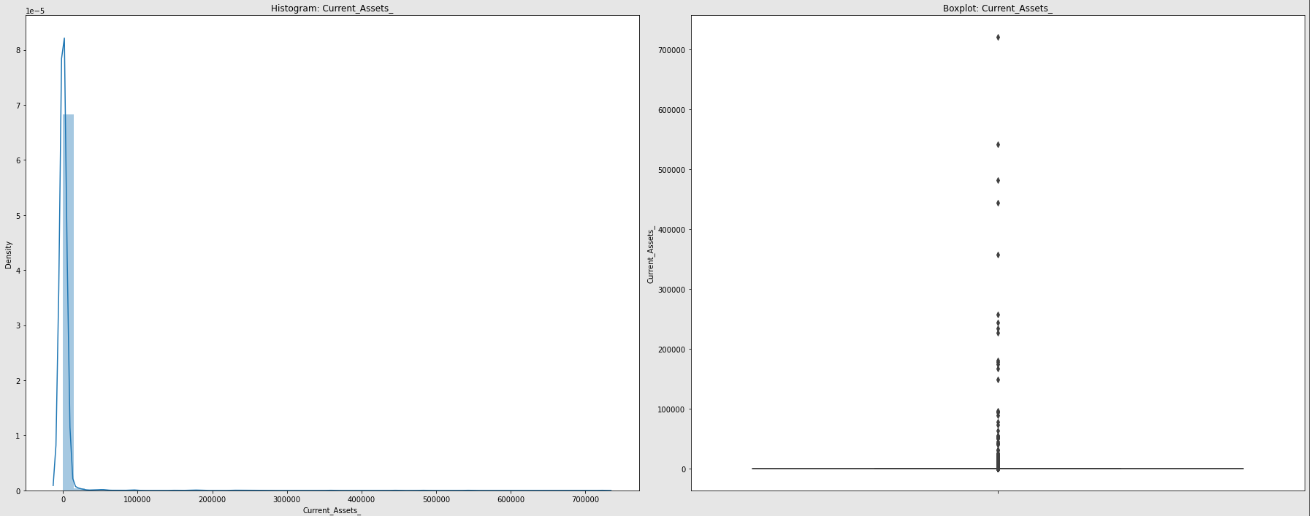


Fig 19. Histogram and Boxplot for Current\_Assets\_

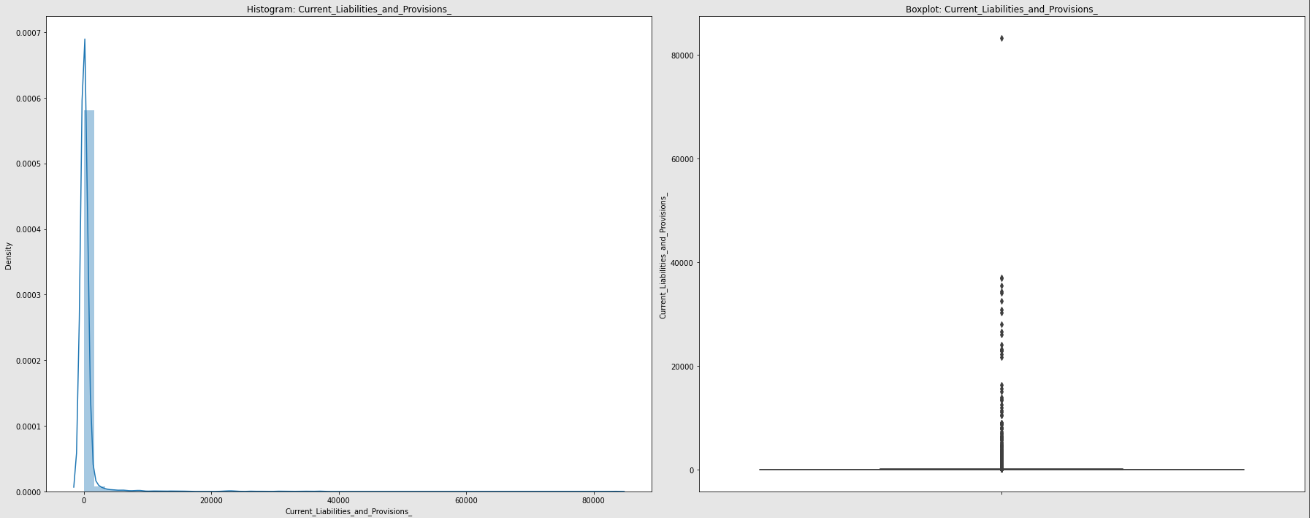
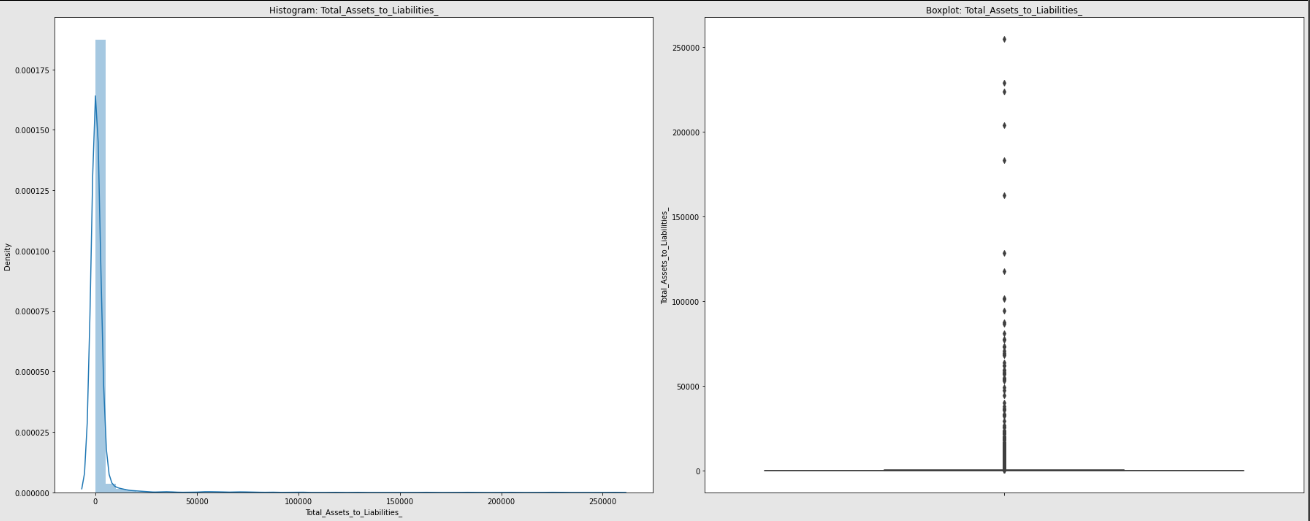
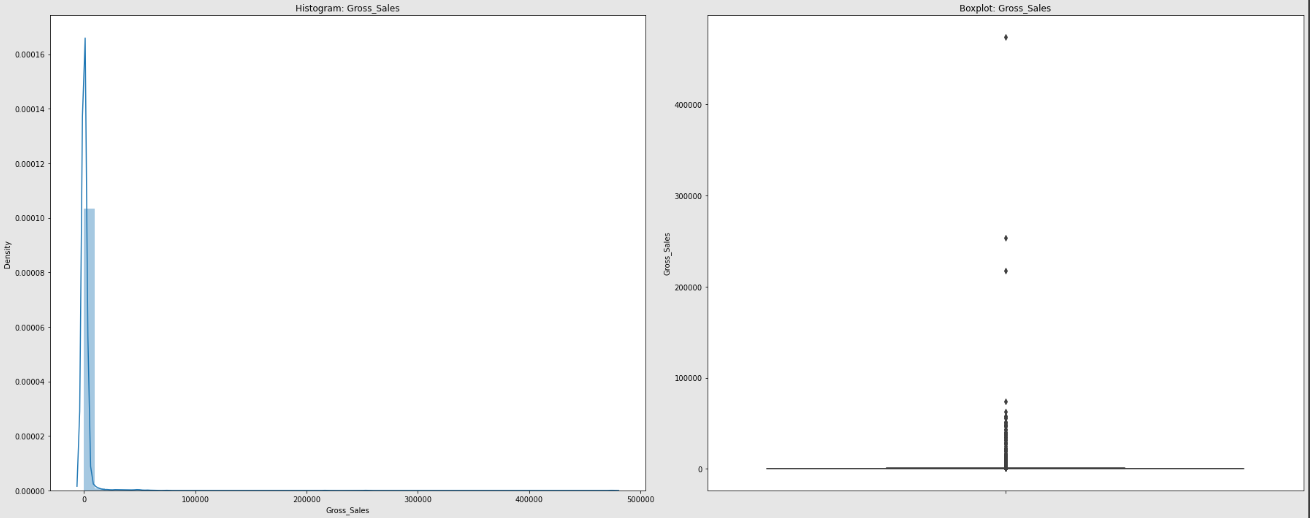
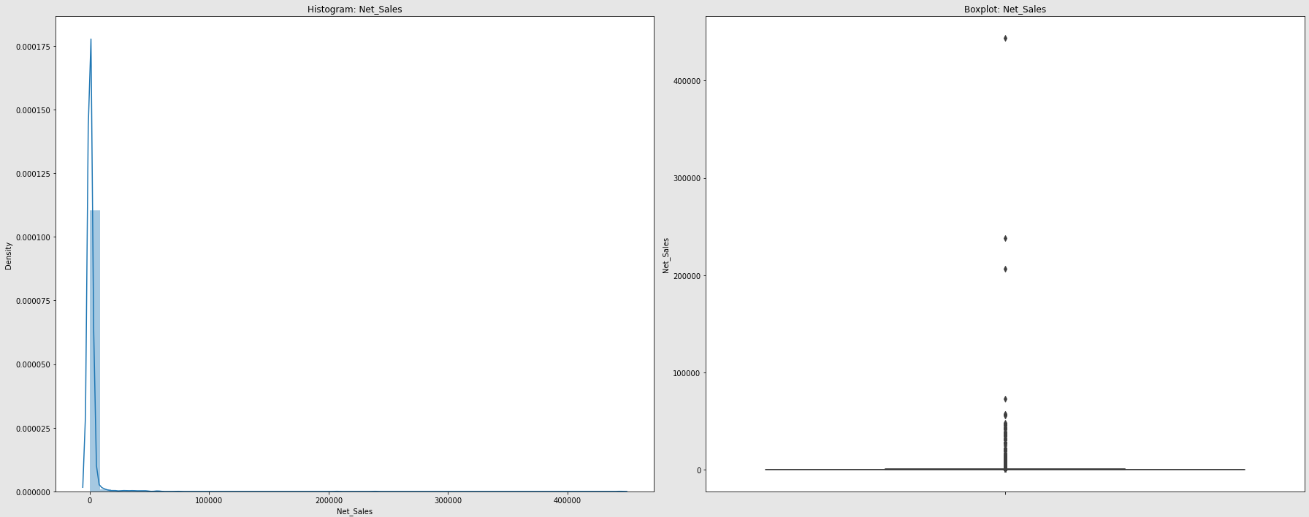


Fig 20. Histogram and Boxplot for Current\_Liabilities\_and\_Provisions\_

Fig 21. Histogram and Boxplot for Total\_Assets\_to\_Liabilities

Fig 22. Histogram and Boxplot for Gross\_Sales\_

Fig 23. Histogram and Boxplot for Net\_Sales\_

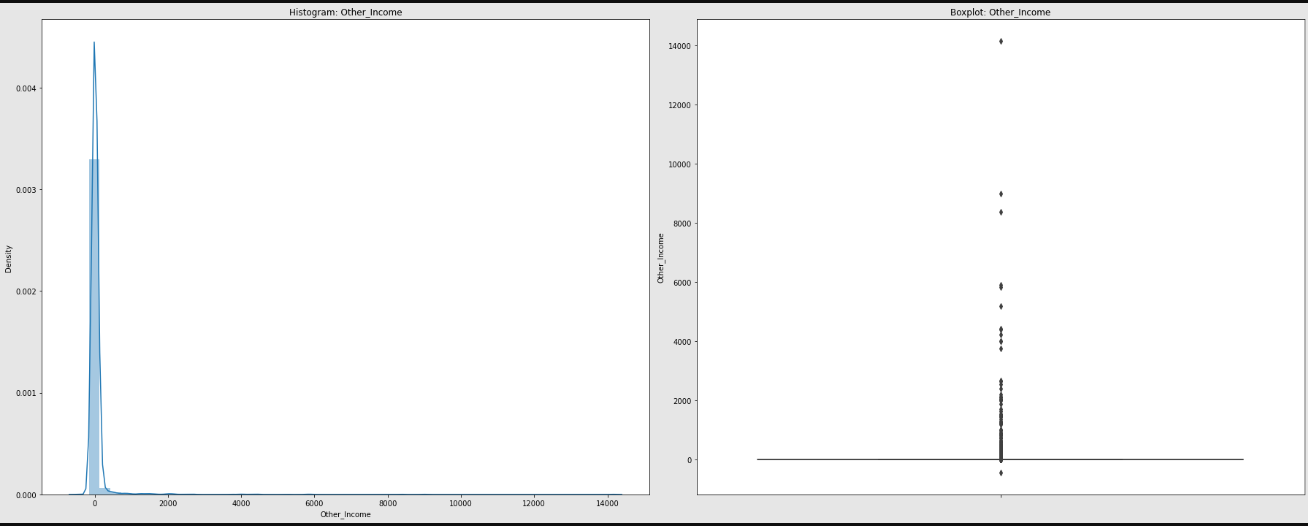


Fig 24. Histogram and Boxplot for Other\_Income\_

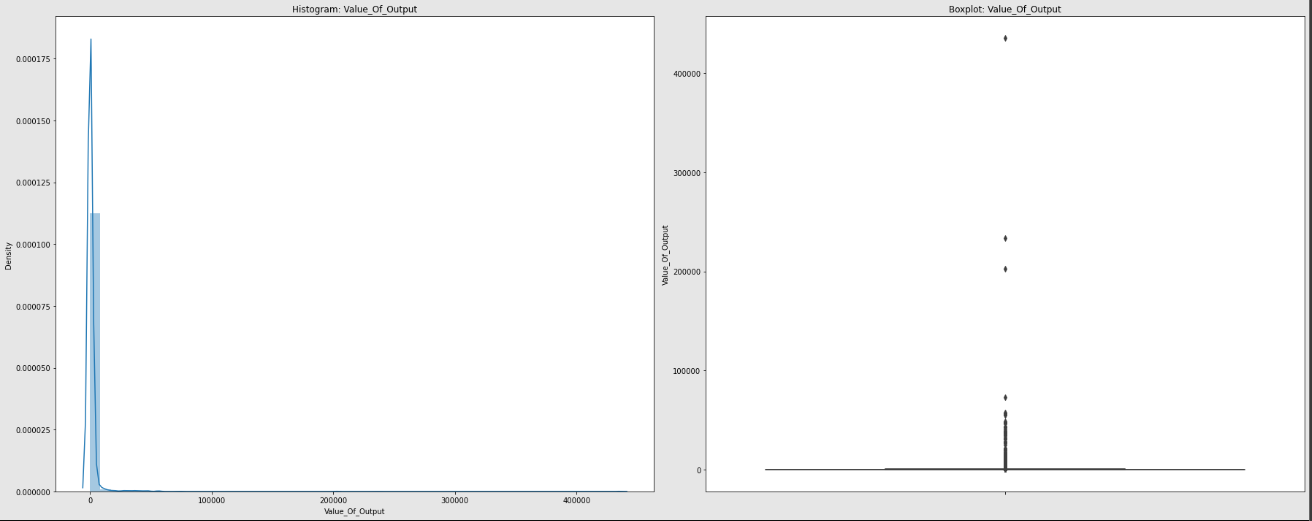


Fig 25. Histogram and Boxplot for Value\_Of\_Output

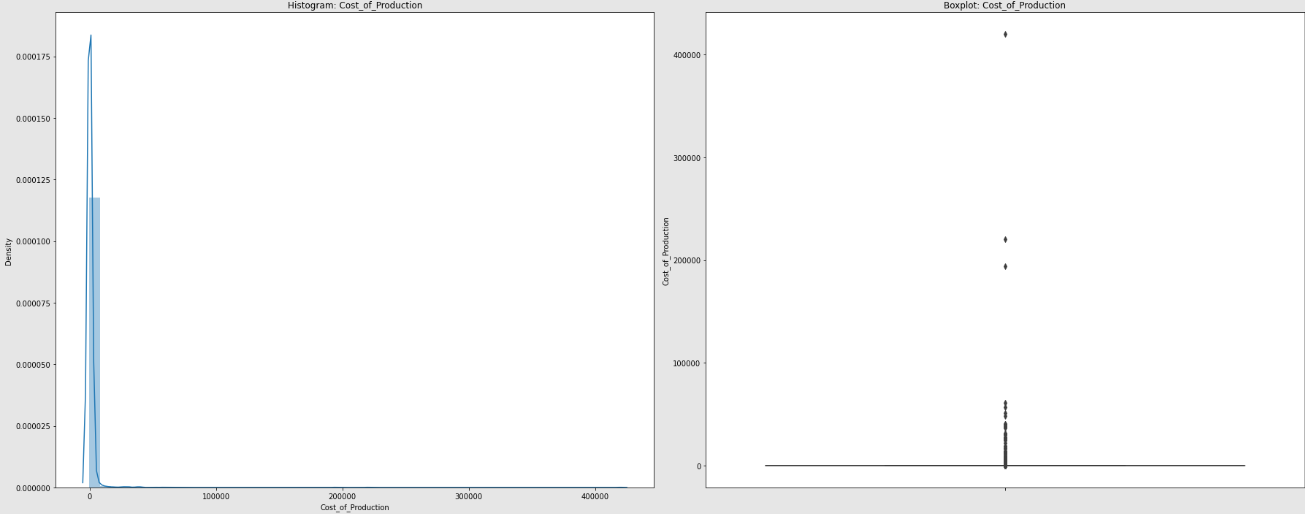


Fig 26. Histogram and Boxplot for Cost\_Of\_Production

Correlation Heatmap:

Mostly correlation is positive in nature among variables with very few negative correlations.

‘default’ is the target field created based on Networth\_Next\_Year and hence not considering it in this plot.

There is very strong correlation between the following variables:

* Networth & Networth\_Next\_Year
* Capital\_Employed & Total\_Debt/Current\_Assets
* Net\_Sales & Gross\_Sales
* Value\_of\_Output & Net\_Sales/Gross\_Sales
* Revenue\_expenses\_in\_forex & Gross\_Sales/Net\_Sales/Value\_Of\_Output/Cost\_of\_Production
* Book\_Value\_Unit\_Curr has very less correlation to other variables.

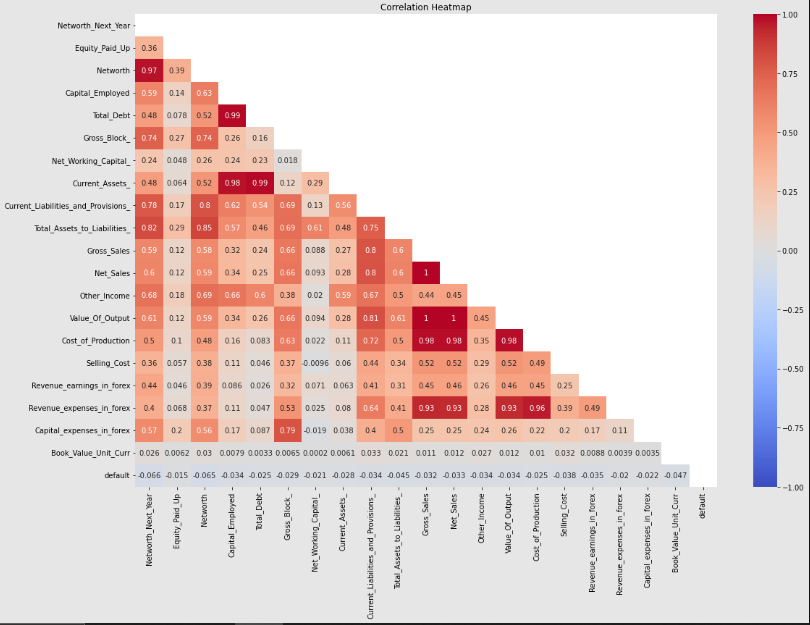


Fig 27. Correlation Heatmap

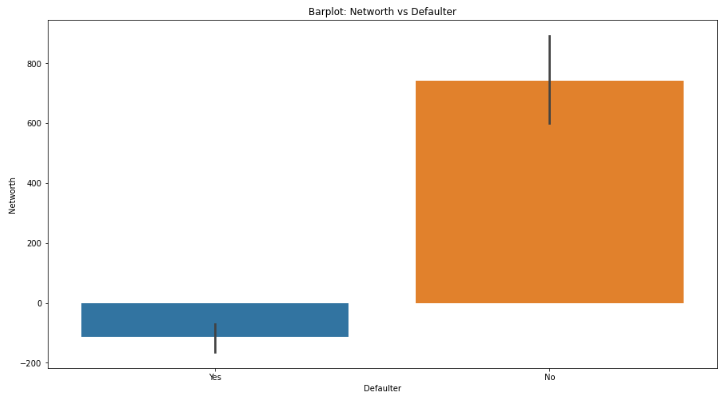


Fig 28. BarPlot – Networth vs Defaulter

Created a variable Defaulter corresponding to default column with 0 as ‘No’ and 1 as ‘Yes’.

We can see for Defaulters the average Networth is around -100 whereas for non-defaulters its around 700.

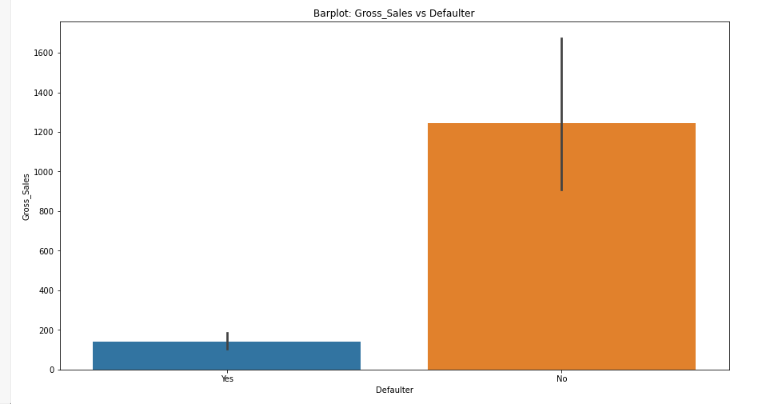


Fig 28. BarPlot – Gross Sales vs Defaulter

We can see for Defaulters the Average Gross\_Sales is around 150 whereas for non-defaulters its around 1200.

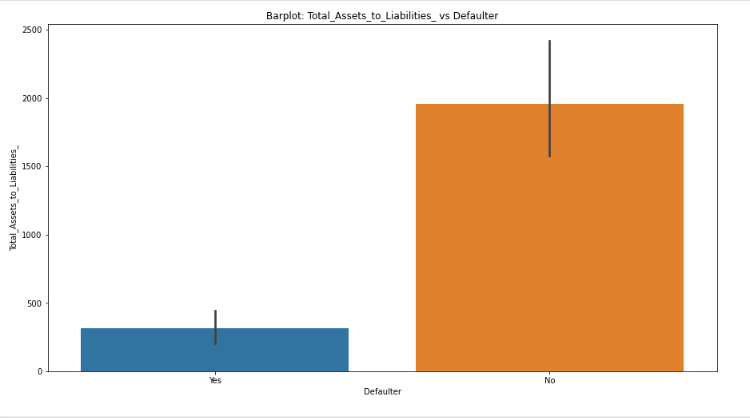


Fig 29. BarPlot – Total\_Assets\_to\_Liabilities\_ vs Defaulter

We can see for Defaulters the Average Total\_Assets\_to\_Liabilities is around 400 whereas for non-defaulters its around 1900.

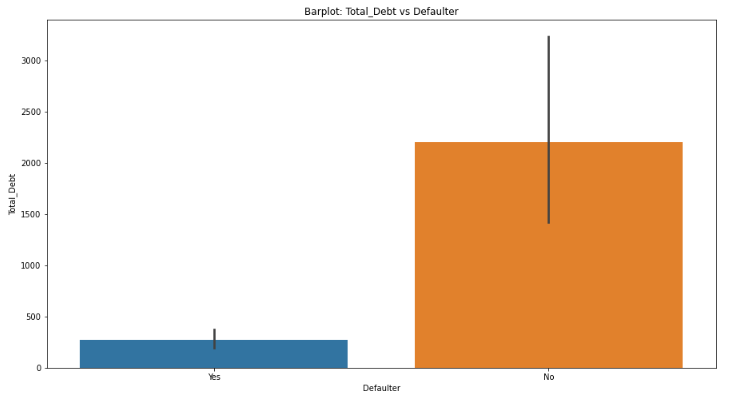
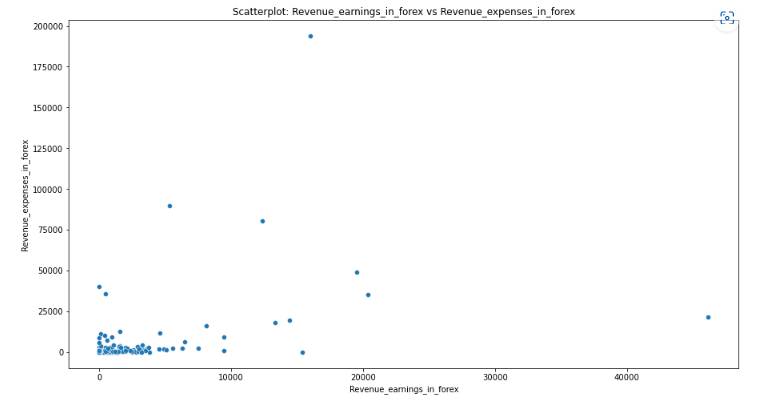


Fig 30. BarPlot – Total\_Debt vs Defaulter

We can see for Defaulters the Average Total\_Debt is around 250 whereas for non-defaulters its around 2100. The value of current debt is not a good indicator for identifying defaulters as non-defaulters has higher debts than defaulters.7

Fig 31. ScatterPlot – Revenue\_Earnings\_in\_Forex vs Revenue\_Expenses\_in\_Forex

We can see that most of the datapoints in Revenue\_expenses\_in\_forex at 15,000 has much lesser value of Revenue\_earnings\_in\_forex. Seems like customers who deals with large amount of forex are mostly in loss.

**1.5 Train Test Split**

Used train\_test\_split module from sklearn library.

The split will be done using test\_size of 0.33 to get a train/test split of 67,33 percentage.

The split will be executed with stratify option to ensure that the defaulters/non-defaulters’ ratio is maintained in both train and test dataset.

After the split we can see that both train and test dataset have the ratio of 89:11.

**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**

The first step towards model building, should be feature selection. So let us look at the features to see if there are multicollinear features, which could possibly undermine the significance of an independent variable.

The collinearity between features through a heatmap (Fig. 10) derived above shows us that there are many fields that have strong positive and negative correlation with each other. Hence we need to identify features with low correlation between each other.

Determining the VIF (Variance Inflation Factor) of the features will help to choose only features where VIF value is low as high VIF indicates that particular feature is highly collinear with other variables in the model.

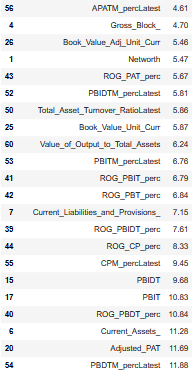


Fig 32. Variation\_Inflation\_Factor

The above table shows the top 50 features ordered as per VIF value. As a norm VIF values above 5 can be not considered as they indicate that these features introduce collinearity among the other features.

We will consider the first 30 features in the above figure for model creation.

We will be creating the models using logistic regression function logit from statsmodel library.

**Model 1:**

In this model the first 30 features from VIF are considered as predictors and default as the target variable.

Logit Regression Results:

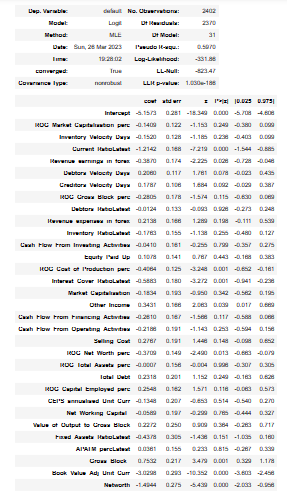


Fig 33. Model 1 Summary

The below predictors have significance greater than 0.05 and hence are less effective in explaining the variance of data determining the target variable.

|  |  |
| --- | --- |
| ROG\_Market\_Capitalisation\_perc | Equity\_Paid\_Up |
| Inventory\_Velocity\_Days | Market\_Capitalisation |
| Debtors\_Velocity\_Days | Cash\_Flow\_From\_Financing\_Activities |
| Creditors\_Velocity\_Days | Cash\_Flow\_From\_Operating\_Activities |
| ROG\_Gross\_Block\_perc | Selling\_Cost |
| Debtors\_RatioLatest | ROG\_Total\_Assets\_perc |
| Revenue\_expenses\_in\_forex | Total\_Debt |
| Inventory\_RatioLatest | ROG\_Capital\_Employed\_perc |
| Cash\_Flow\_From\_Investing\_Activities | CEPS\_annualised\_Unit\_Curr |
| Fixed\_Assets\_RatioLatest | Net\_Working\_Capital\_ |
| APATM\_percLatest | Value\_of\_Output\_to\_Gross\_Block |

Let us look at the confusion matrix and classification report of both Train and Test set

Train:

[[2112 30]

[ 87 173]]

^

precision recall f1-score support

0.0 0.96 0.99 0.97 2142

1.0 0.85 0.67 0.75 260

accuracy 0.95 2402

macro avg 0.91 0.83 0.86 2402

weighted avg 0.95 0.95 0.95 2402

The recall is 0.67 and is very low. Hence we need to further fine tune the model.

**Model 2:**

This model is an iteration of above model 1, where we will use only the pertinent features and avoid the predictors with p-value greater than 0.05. Hence the features considered for this model are:

|  |  |
| --- | --- |
| Current\_RatioLatest | ROG\_Net\_Worth\_perc |
| Revenue\_earnings\_in\_forex | Gross\_Block\_ |
| ROG\_Cost\_of\_Production\_perc | Book\_Value\_Adj\_Unit\_Curr |
| Interest\_Cover\_RatioLatest | Networth |
| Other\_Income |  |

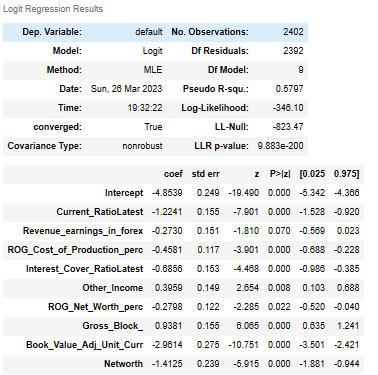


Fig 34. Model 2 Summary

Confusion matrinx and classification report on Train and Test set in Model 2:

Train:

[[2116 26]

[ 87 173]]

precision recall f1-score support

0.0 0.96 0.99 0.97 2142

1.0 0.87 0.67 0.75 260

accuracy 0.95 2402

macro avg 0.91 0.83 0.86 2402

weighted avg 0.95 0.95 0.95 2402

The recall scores are still low at 67 % that is of all the defaulter’s model can predict 67% of them, which is a low benchmark. We need to better this score by further fine tuning this model.

**Model 3:**

In this model we will again consider the first 30 features in table 1.11 as predictors and default as the target variable.

On creating the model and training with the smote-train dataset we get the below summary:

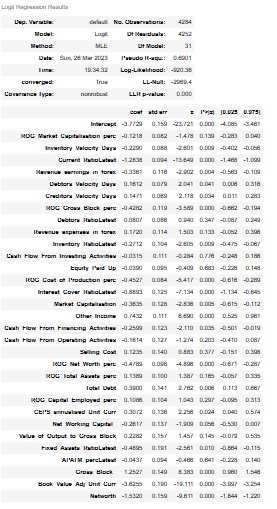


Fig 35. Model 3 Summary

We can see that the below predictors have significance greater than 0.05 and hence are less effective in explaining the variance of data determining the target variable.

|  |  |
| --- | --- |
| ROG\_Market\_Capitalisation\_perc | Selling\_Cost |
| Debtors\_RatioLatest | ROG\_Total\_Assets\_perc |
| Revenue\_expenses\_in\_forex | ROG\_Capital\_Employed\_perc |
| Cash\_Flow\_From\_Investing\_Activities | Net\_Working\_Capital\_ |
| Equity\_Paid\_Up | Value\_of\_Output\_to\_Gross\_Block |
| Cash\_Flow\_From\_Operating\_Activities | APATM\_percLatest |

Confusion matrix and classification report on Train and Test set in Model 3:

Train:

[[1935 207]

[ 129 2013]]

precision recall f1-score support

0.0 0.94 0.90 0.92 2142

1.0 0.91 0.94 0.92 2142

accuracy 0.92 4284

macro avg 0.92 0.92 0.92 4284

weighted avg 0.92 0.92 0.92 4284

As we can see from the above values, the precision for defaulters is 0.91 and recall is 0.94.

The recall for this model is high and the model performance looks good.

**Model 4:**

This model is an iteration of above model 3, where we will use only the pertinent features and avoid the predictors with p-value greater than 0.05. Hence the features considered for this model are:

|  |  |
| --- | --- |
| Inventory\_Velocity\_Days | Other\_Income |
| Current\_RatioLatest | Cash\_Flow\_From\_Financing\_Activities |
| Revenue\_earnings\_in\_forex | ROG\_Net\_Worth\_perc |
| Debtors\_Velocity\_Days | Total\_Debt |
| Creditors\_Velocity\_Days | CEPS\_annualised\_Unit\_Curr |
| ROG\_Gross\_Block\_perc | Fixed\_Assets\_RatioLatest |
| Inventory\_RatioLatest | Gross\_Block\_ |
| ROG\_Cost\_of\_Production\_perc | Book\_Value\_Adj\_Unit\_Curr |
| Interest\_Cover\_RatioLatest | Networth |
| Market\_Capitalisation |  |

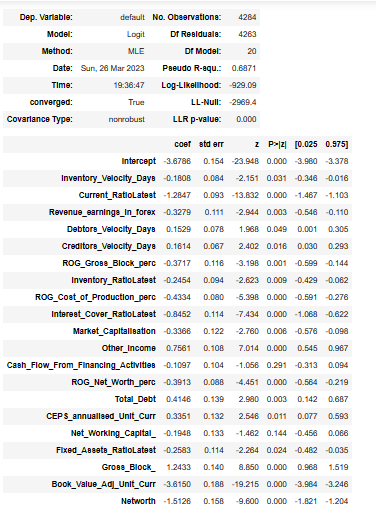


Fig 36. Model 4 Summary

We can see that all predictors have p-values lesser than 0.05 and is significant in explaining the deviance towards target variable.

Confusion matrix and classification report on Train and Test set in Model 4:

Train:

[[1948 194]

[ 129 2013]]

precision recall f1-score support

0.0 0.94 0.91 0.92 2142

1.0 0.91 0.94 0.93 2142

accuracy 0.92 4284

macro avg 0.92 0.92 0.92 4284

weighted avg 0.92 0.92 0.92 4284

So, we can see that the precision for defaulters is 0.91 and recall is 0.94.

The recall for this model is high and the model performance looks good.

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

Model 1 Test:

[[1038 18]

[ 52 76]]

precision recall f1-score support

0.0 0.95 0.98 0.97 1056

1.0 0.81 0.59 0.68 128

accuracy 0.94 1184

macro avg 0.88 0.79 0.83 1184

weighted avg 0.94 0.94 0.94 1184

Model 2 Test:

[[1040 16]

[ 49 79]]

precision recall f1-score support

0.0 0.96 0.98 0.97 1056

1.0 0.83 0.62 0.71 128

accuracy 0.95 1184

macro avg 0.89 0.80 0.84 1184

weighted avg 0.94 0.95 0.94 1184

Model 3 Test:

[[936 120]

[ 18 110]]

precision recall f1-score support

0.0 0.98 0.89 0.93 1056

1.0 0.48 0.86 0.61 128

accuracy 0.88 1184

macro avg 0.73 0.87 0.77 1184

weighted avg 0.93 0.88 0.90 1184

Model 4 Test:

[[934 122]

[ 18 110]]

precision recall f1-score support

0.0 0.98 0.88 0.93 1056

1.0 0.47 0.86 0.61 128

accuracy 0.88 1184

macro avg 0.73 0.87 0.77 1184

weighted avg 0.93 0.88 0.90 1184

Let us also look at ROC\_AUC score and ROC curves:

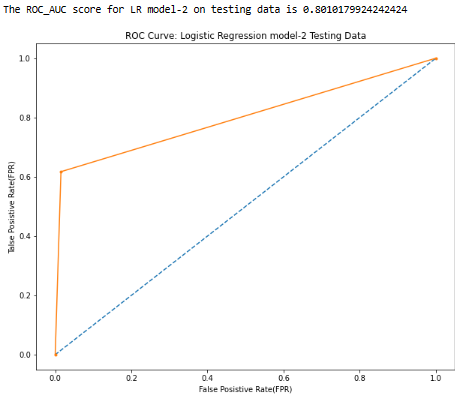
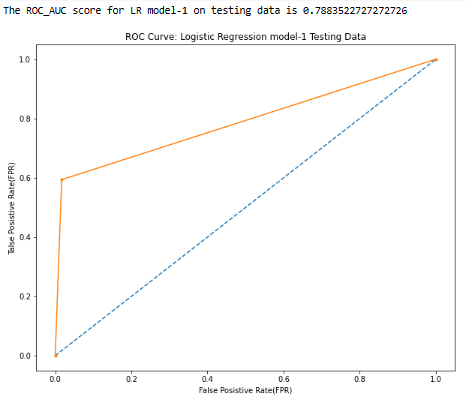


Fig 37: Model 1 ROC\_AUC score and ROC curve Fig 38: Model 2 ROC\_AUC score and ROC curve

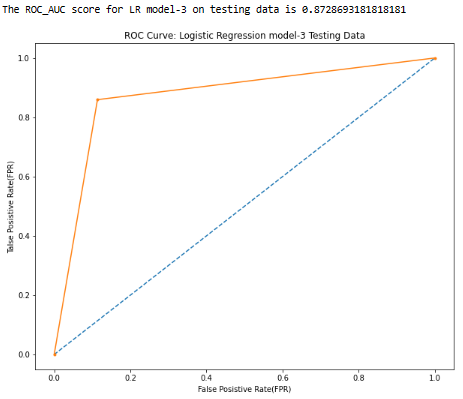
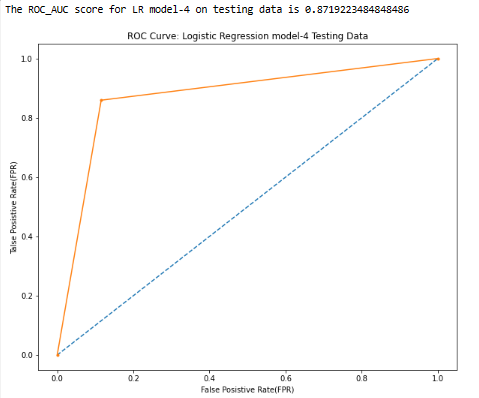
 

Fig 39: Model 3 ROC\_AUC score and ROC curve Fig 40: Model 4 ROC\_AUC score and ROC curve

For Model 1, we can see that the precision for defaulters is 0.81 and recall is 0.59 and ROC\_AUC score is 0.79

The recall for this model is 59%, that is of all the defaulters, model can predict 59% of them, which is a low benchmark and not acceptable model.

For Model 2, we can see that the precision for defaulters is 0.83 and recall is 0.62 and ROC\_AUC score is 0.8.

The recall for this model is 62%, that is of all the defaulters, model can predict 62% of them, which is a low benchmark and not acceptable model, though it is better than model 1.

For Model 3, we can see that the precision for defaulters is 0.48 and recall is 0.86 and ROC\_AUC score is 0.87.

The recall for this model is 86%, that is of all the defaulters, model can predict 86% of them, which is an acceptable benchmark. The recall for the training dataset is 94% and hence the model is neither overfit nor underfit as the recall percent difference is within the threshold of 10%

For Model 4, we can see that the precision for defaulters is 0.47 and recall is 0.86, and ROC\_AUC score is 0.87.

The recall for this model is 86%, that is of all the defaulters, model can predict 86% of them, which is an acceptable benchmark. The recall for the training dataset is 94% and hence the model is neither overfit or underfit as the recall percent difference is within the threshold of 10%.

**Interpretations from the model:**

We can choose model 4 as the best model as it has comparable recall with Model 3 and uses lesser features, thus being cost effective.

Since the input dataset is imbalanced, SMOTE has helped in balancing the learning dataset and thus improve the recall statistics as required in this business scenario.

We have achieved a decent recall value without overfit and ROC\_AUC score is a decent 0.87. Considering the opportunities such as outliers, missing values, and correlated features this is a fairly good model. It can be improved if we get better quality data where the features explaining the default are not missing from the dataset.