greatlearning
Power Ahead

FRA

MILESTONE - 1

Pooja Kabadi
PGP-DSBA Online
Batch- A4
11-06-2022

Table of Contents:

List of Tables:

List of Figures:

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 Net worth of the company in the following year (2016) is provided which can be used to drive the labelled 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.

Data Dictionary:

Sl.no	Field Name	Description
1	Co_Code	Company Code
2	Co_Name	Company Name
3	Networth Next Year	Value of a company as on 2016 - Next Year(difference between the value of total assets and total liabilities)
4	Equity Paid Up	Amount that has been received by the company through the issue of shares to the shareholders
5	Networth	Value of a company as on 2015 - Current Year
6	Capital Employed	Total amount of capital used for the acquisition of profits by a company
7	Total Debt	The sum of money borrowed by the company and is due to be paid
8	Gross Block	Total value of all of the assets that a company owns
9	Net Working Capital	The difference between a company's current assets (cash, accounts receivable, inventories of raw materials and finished goods) and its current liabilities (accounts payable).
10	Current Assets	All the assets of a company that are expected to be sold or used as a result of standard business operations over the next year.

11	Current Liabilities and Provisions	Short-term financial obligations that are due within one year (includes amount that is set aside cover a future liability)
12	Total Assets/Liabilities	Ratio of total assets to liabilities of the company
13	Gross Sales	The grand total of sale transactions within the accounting period
14	Net Sales	Gross sales minus returns, allowances, and discounts
15	Other Income	Income realized from non-business activities (e.g. sale of long term asset)
16	Value Of Output	Product of physical output of goods and services produced by company and its market price
17	Cost of Production	Costs incurred by a business from manufacturing a product or providing a service
18	Selling Cost	Costs which are made to create the demand for the product (advertising expenditures, packaging and styling, salaries, commissions and travelling expenses of sales personnel, and the cost of shops and showrooms)
19	PBIDT	Profit Before Interest, Depreciation & Taxes
20	PBDT	Profit Before Depreciation and Tax
21	PBIT	Profit before interest and taxes
22	PBT	Profit before tax
23	PAT	Profit After Tax
24	Adjusted PAT	Adjusted profit is the best estimate of the true profit
26	CP	Commercial paper , a short-term debt instrument to meet short-term liabilities.
27	Revenue earnings in forex	Revenue earned in foreign currency
28	Revenue expenses in forex	Expenses due to foreign currency transactions
29	Capital expenses in forex	Long term investment in forex
30	Book Value (Unit Curr)	Net asset value
31	Book Value (Adj.) (Unit Curr)	Book value adjusted to reflect asset's true fair market value
32	Market Capitalisation	Product of the total number of a company's outstanding shares and the current market price of one share
33	CEPS (annualised) (Unit Curr)	Cash Earnings per Share, profitability ratio that measures the financial performance of a company by calculating cash flows on a per share basis
34	Cash Flow From Operating Activities	Use of cash from ongoing regular business activities
35	Cash Flow From Investing Activities	Cash used in the purchase of non-current assets–or long-term assets– that will deliver value in the future
36	Cash Flow From Financing Activities	Net flows of cash that are used to fund the company (transactions involving debt, equity, and dividends)
37	ROG-Net Worth (%)	Rate of Growth - Networth
38	ROG-Capital Employed (%)	Rate of Growth - Capital Employed
39	ROG-Gross Block (%)	Rate of Growth - Gross Block
40	ROG-Gross Sales (%)	Rate of Growth - Gross Sales
41	ROG-Net Sales (%)	Rate of Growth - Net Sales
42	ROG-Cost of Production (%)	Rate of Growth - Cost of Production

43	ROG-Total Assets (%)	Rate of Growth - Total Assets
44	ROG-PBIDT (%)	Rate of Growth- PBIDT
45	ROG-PBDT (%)	Rate of Growth- PBDT
46	ROG-PBIT (%)	Rate of Growth- PBIT
47	ROG-PBT (%)	Rate of Growth- PBT
48	ROG-PAT (%)	Rate of Growth- PAT
49	ROG-CP (%)	Rate of Growth- CP
50	ROG-Revenue earnings in forex (%)	Rate of Growth - Revenue earnings in forex
51	ROG-Revenue expenses in forex (%)	Rate of Growth - Revenue expenses in forex
52	ROG-Market Capitalisation (%)	Rate of Growth - Market Capitalisation
53	Current Ratio[Latest]	Liquidity ratio, company's ability to pay short-term obligations or those due within one year
54	Fixed Assets Ratio[Latest]	Solvency ratio, the capacity of a company to discharge its obligations towards long-term lenders indicating
55	Inventory Ratio[Latest]	Activity ratio, specifies the number of times the stock or inventory has been replaced and sold by the company
56	Debtors Ratio[Latest]	Measures how quickly cash debtors are paying back to the company
57	Total Asset Turnover Ratio[Latest]	The value of a company's revenues relative to the value of its assets
58	Interest Cover Ratio[Latest]	Determines how easily a company can pay interest on its outstanding debt
59	PBIDTM (%)[Latest]	Profit before Interest Depreciation and Tax Margin
60	PBITM (%)[Latest]	Profit Before Interest Tax Margin
61	PBDTM (%)[Latest]	Profit Before Depreciation Tax Margin
62	CPM (%)[Latest]	Cost per thousand (advertising cost)
63	APATM (%)[Latest]	After tax profit margin
64	Debtors Velocity (Days)	Average days required for receiving the payments
65	Creditors Velocity (Days)	Average number of days company takes to pay suppliers
66	Inventory Velocity (Days)	Average number of days the company needs to turn its inventory into sales
67	Value of Output/Total Assets	Ratio of Value of Output (market value) to Total Assets
68	Value of Output/Gross Block	Ratio of Value of Output (market value) to Gross Block

Table 1. Data dictionary

Objective of the report:

The financial statements of companies for the year 2015 and the net worth of companies for the year 2016 have been provided to us as a dataset. We must estimate credit risk based on the facts by estimating the companies' Net worth, which will assist investors in making the best decision possible.

Approach: To forecast net worth, we used Python to develop a Logistic regression model. We will pre-process the data to remove outliers and missing values before developing the model and also find out the significant variables by calculating the VIF score and building the models based on the features with VIF score from 0 to 5.

Exploratory Data Analysis:

Read and view data: Reading the dataset from the excel file and checking the head () of the dataset i.e., the first 5 rows of the dataset.

Checking for the information of features

	Co_Code	Co_Name	Networth_Next_Year	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block	Net_Working_Capital	Curr_Assets
0	16974	Hind.Cables	-8021.60	419.36	-7027.48	-1007.24	5936.03	474.30	-1076.34	40.50
1	21214	Tata Tele. Mah.	-3986.19	1954.93	-2968.08	4458.20	7410.18	9070.86	-1098.88	486.86
2	14852	ABG Shipyard	-3192.58	53.84	506.86	7714.68	6944.54	1281.54	4496.25	9097.64
3	2439	GTL	-3054.51	157.30	-623.49	2353.88	2326.05	1033.69	-2612.42	1034.12
4	23505	Bharati Defence	-2967.36	50.30	-1070.83	4675.33	5740.90	1084.20	1836.23	4685.81

5 rows × 67 columns

Figure 1. Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3586 entries, 0 to 3585
Data columns (total 67 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Co_Code                               3586 non-null   int64
 1   Co_Name                               3586 non-null   object
 2   Networth_Next_Year                    3586 non-null   float64
 3   Equity_Paid_Up                        3586 non-null   float64
 4   Networth                              3586 non-null   float64
 5   Capital_Employed                      3586 non-null   float64
 6   Total_Debt                            3586 non-null   float64
 7   Gross_Block                           3586 non-null   float64
 8   Net_Working_Capital                   3586 non-null   float64
 9   Curr_Assets                           3586 non-null   float64
 10  Curr_Liab_and_Prov                    3586 non-null   float64
 11  Total_Assets_to_Liab                  3586 non-null   float64
 12  Gross_Sales                           3586 non-null   float64
 13  Net_Sales                             3586 non-null   float64
 14  Other_Income                          3586 non-null   float64
 15  Value_Of_Output                       3586 non-null   float64
 16  Cost_of_Prod                          3586 non-null   float64
 17  Selling_Cost                          3586 non-null   float64
 18  PBIDT                                 3586 non-null   float64
 19  PBDT                                  3586 non-null   float64
 20  PBIT                                  3586 non-null   float64
 21  PBT                                   3586 non-null   float64
 22  PAT                                   3586 non-null   float64
 23  Adjusted_PAT                          3586 non-null   float64
 24  CP                                    3586 non-null   float64
 25  Rev_earn_in_forex                     3586 non-null   float64
 26  Rev_exp_in_forex                      3586 non-null   float64
 27  Capital_exp_in_forex                  3586 non-null   float64
```

```
 28  Book_Value_Unit_Curr               3586 non-null   float64
 29  Book_Value_Adj_Unit_Curr           3582 non-null   float64
 30  Market_Capitalisation              3586 non-null   float64
 31  CEPS_annualised_Unit_Curr          3586 non-null   float64
 32  Cash_Flow_From_Opr                 3586 non-null   float64
 33  Cash_Flow_From_Inv                 3586 non-null   float64
 34  Cash_Flow_From_Fin                 3586 non-null   float64
 35  ROG_Net_Worth_perc                 3586 non-null   float64
 36  ROG_Capital_Employed_perc          3586 non-null   float64
 37  ROG_Gross_Block_perc               3586 non-null   float64
 38  ROG_Gross_Sales_perc               3586 non-null   float64
 39  ROG_Net_Sales_perc                 3586 non-null   float64
 40  ROG_Cost_of_Prod_perc              3586 non-null   float64
 41  ROG_Total_Assets_perc              3586 non-null   float64
 42  ROG_PBIDT_perc                     3586 non-null   float64
 43  ROG_PBDT_perc                      3586 non-null   float64
 44  ROG_PBIT_perc                      3586 non-null   float64
 45  ROG_PBT_perc                       3586 non-null   float64
 46  ROG_PAT_perc                       3586 non-null   float64
 47  ROG_CP_perc                        3586 non-null   float64
 48  ROG_Rev_earn_in_forex_perc         3586 non-null   float64
 49  ROG_Rev_exp_in_forex_perc          3586 non-null   float64
 50  ROG_Market_Capitalisation_perc     3586 non-null   float64
 51  Curr_Ratio_Latest                  3585 non-null   float64
 52  Fixed_Assets_Ratio_Latest          3585 non-null   float64
 53  Inventory_Ratio_Latest             3585 non-null   float64
 54  Debtors_Ratio_Latest               3585 non-null   float64
 55  Total_Asset_Turnover_Ratio_Latest  3585 non-null   float64
 56  Interest_Cover_Ratio_Latest        3585 non-null   float64
 57  PBIDTM_perc_Latest                 3585 non-null   float64
 58  PBITM_perc_Latest                  3585 non-null   float64
 59  PBDTM_perc_Latest                  3585 non-null   float64
 60  CPM_perc_Latest                    3585 non-null   float64
 61  APATM_perc_Latest                  3585 non-null   float64
 62  Debtors_Vel_Days                   3586 non-null   int64
 63  Creditors_Vel_Days                 3586 non-null   int64
 64  Inventory_Vel_Days                 3483 non-null   float64
 65  Value_of_Output_to_Total_Assets    3586 non-null   float64
 66  Value_of_Output_to_Gross_Block     3586 non-null   float64
dtypes: float64(63), int64(3), object(1)
memory usage: 1.8+ MB
```

Observations:

- Number of rows: 3587
- Number of columns: 67
- Data has 67 variables of which 63 are float type, 3 are integer and 1 is object data type.
- There are no missing and duplicate values in the dataset.
- The target variable is Net worth which is further used to segregate into values of 0 and 1, value of 1 when net worth next year is negative & 0 when net worth next year is positive.

Checking the description of dataset:

Dropping company code variable as its not useful for modelling. And later checking the 5-point statistical memory of all the variables.

	Networth_Next_Year	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block	Net_Working_Capital	Curr_Assets
count	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000
mean	725.045251	62.966584	649.746299	2799.611054	1994.823779	594.178829	410.809665	1960.349172
std	4769.681004	778.761744	4091.988792	26975.135385	23652.842746	4871.547802	6301.218546	22577.570829
min	-8021.600000	0.000000	-7027.480000	-1824.750000	-0.720000	-41.190000	-13162.420000	-0.910000
25%	3.985000	3.750000	3.892500	7.602500	0.030000	0.570000	0.942500	4.000000
50%	19.015000	8.290000	18.580000	39.090000	7.490000	15.870000	10.145000	24.540000
75%	123.802500	19.517500	117.297500	226.605000	72.350000	131.895000	61.175000	135.277500
max	111729.100000	42263.460000	81657.350000	714001.250000	652823.810000	128477.590000	223257.560000	721166.000000

Curr_Liab_and_Prov	Total_Assets_to_Liab	...	PBIDTM_perc_Latest	PBITM_perc_Latest	PBDTM_perc_Latest	CPM_perc_Latest	APATM_perc_Latest
3586.000000	3586.000000	...	3585.000000	3585.000000	3585.000000	3585.000000	3585.000000
391.992078	1778.453751	...	-51.162890	-109.213414	-311.570357	-307.005632	-365.056187
2675.001631	11437.574690	...	1795.131025	3057.635870	10921.592639	10676.149629	12500.051387
-0.230000	-4.510000	...	-78870.450000	-141600.000000	-590500.000000	-572000.000000	-688600.000000
0.732500	10.555000	...	0.000000	0.000000	0.000000	0.000000	0.000000
9.225000	52.010000	...	8.070000	5.230000	4.690000	3.890000	1.590000
65.650000	310.540000	...	18.990000	14.290000	14.110000	11.390000	7.410000
83232.980000	254737.220000	...	19233.330000	19195.700000	15640.000000	15640.000000	15266.670000

Debtors_Vel_Days	Creditors_Vel_Days	Inventory_Vel_Days	Value_of_Output_to_Total_Assets	Value_of_Output_to_Gross_Block
3586.000000	3.586000e+03	3483.000000	3586.000000	3586.000000
603.894032	2.057855e+03	79.644559	0.819757	61.884548
10636.759580	5.416948e+04	137.847792	1.201400	976.824352
0.000000	0.000000e+00	-199.000000	-0.330000	-61.000000
8.000000	8.000000e+00	0.000000	0.070000	0.270000
49.000000	3.900000e+01	35.000000	0.480000	1.530000
106.000000	8.900000e+01	96.000000	1.160000	4.910000
514721.000000	2.034145e+06	996.000000	17.630000	43404.000000

Figure 2. Description of data

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 real effect and leads to faulty predictions

In our case we employee **Inter-Quartile Range (IQR) Treatment** for Outliers

IQR range – Interquartile Range is just a mathematical way to find outliers. Box plots are based on these calculations only. All the data points from first quartile to third quartile are said to lie in the interquartile range.

We subtract 1.5*IQR to find the minimum value below which all data points are considered as outliers whereas, we add 1.5*IQR to find the maximum value above which all the data points are considered as outliers.

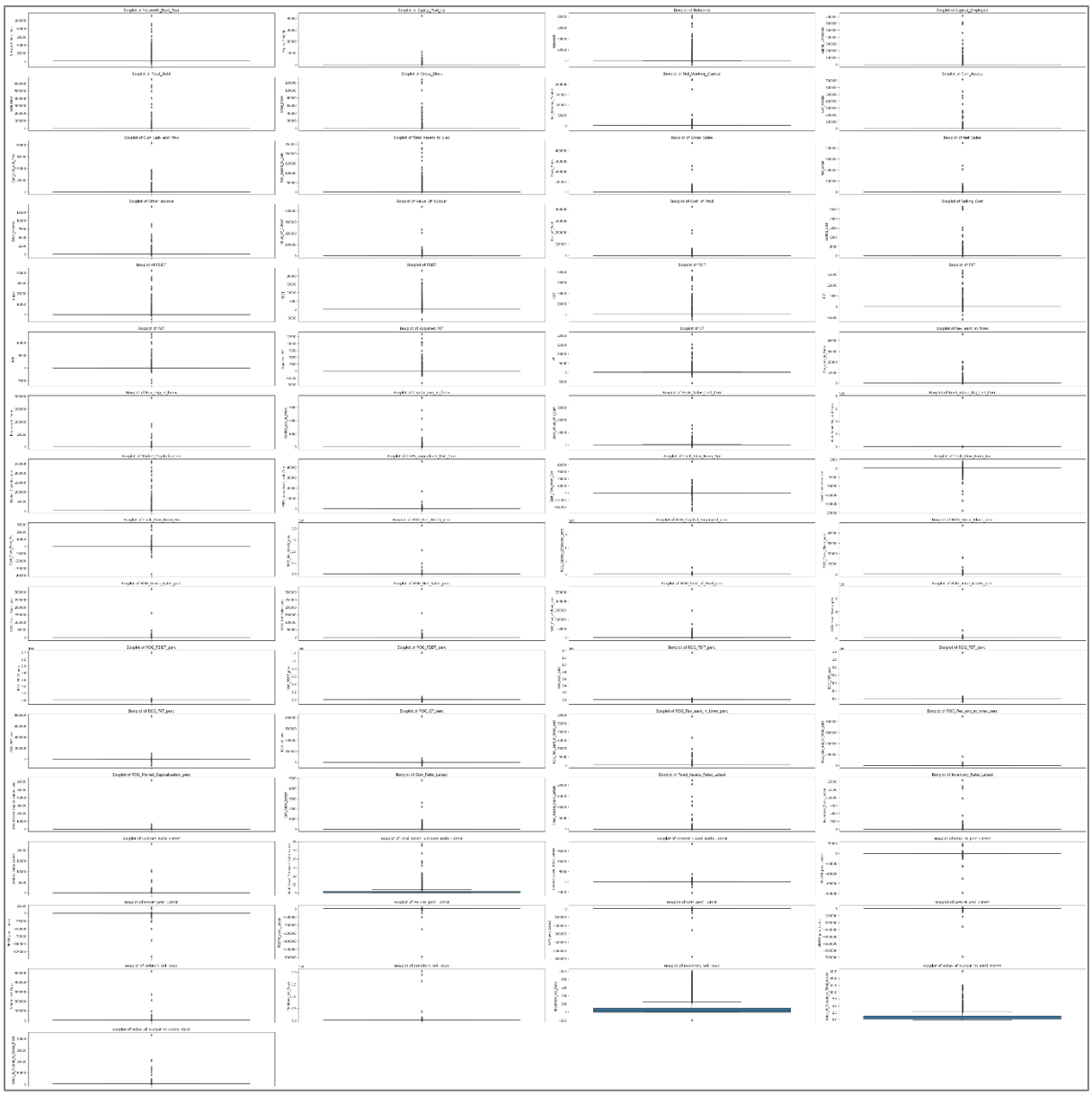
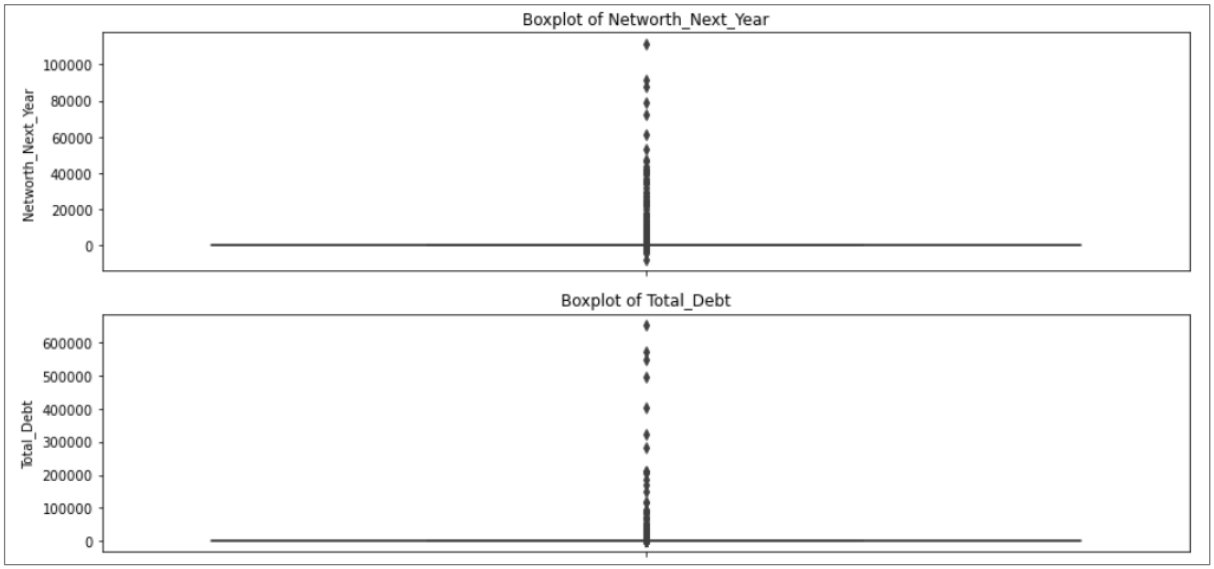


Figure 3. Boxplot before outlier treatment

The boxplot of few of the variables are zoomed and shown below:



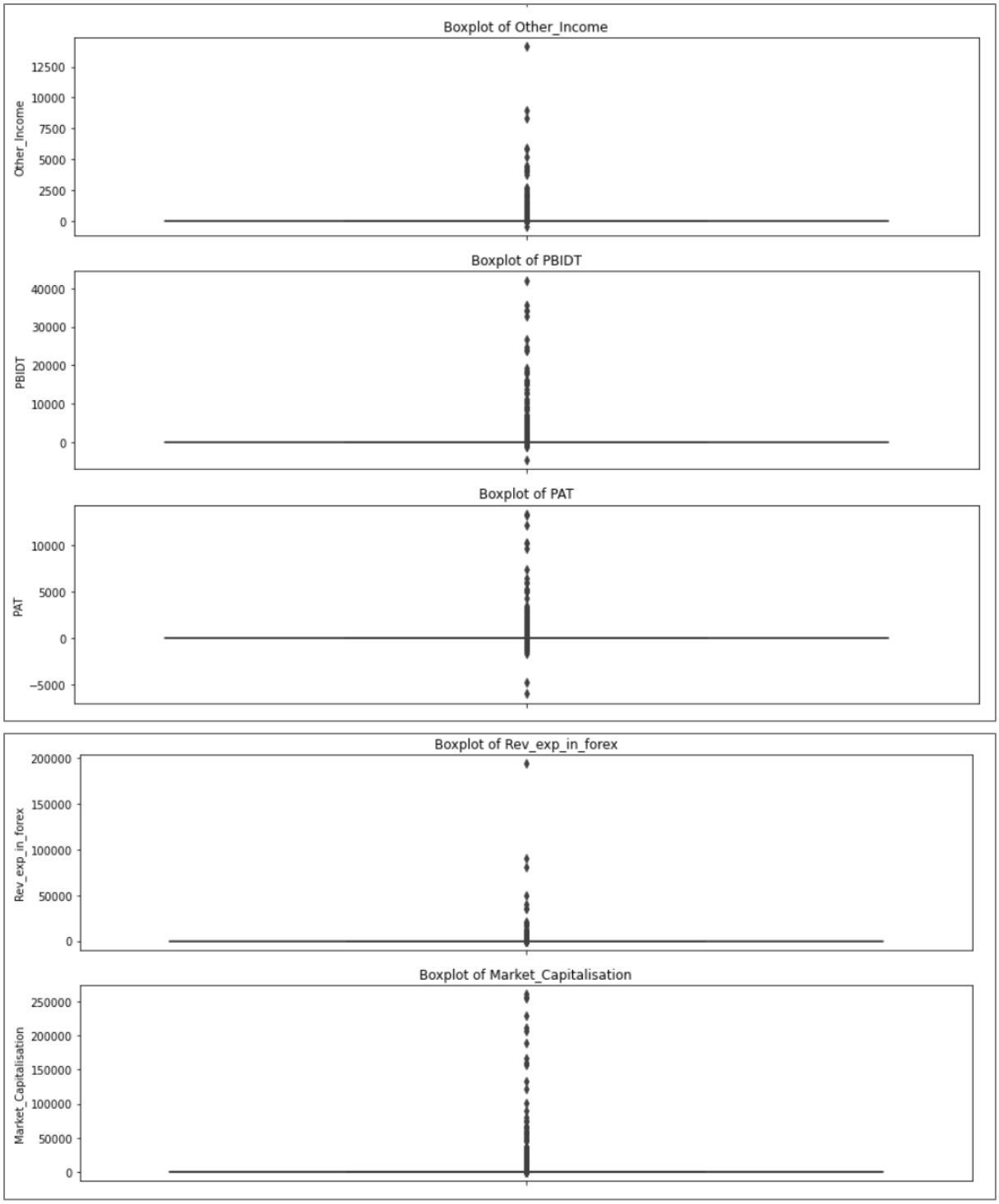


Figure 4. Zoomed Boxplot of variables to check outliers.

Observations:

- It is very evident that there are outliers present in almost all the variables.
- Significant number of outliers are present for almost all the variables. Captured the actual percentage of data which are above and below the third and first quantiles and capped the outliers accordingly.

The data is divided in to X and Y as:

default_X = default_num.drop('Networth_Next_Year', axis = 1)

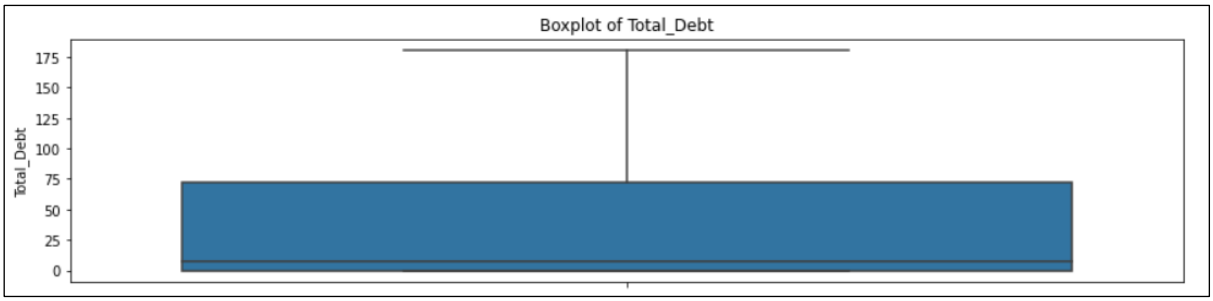
default_Y = default_num['Networth_Next_Year']

- The outlier treatment is applied to all the variables excluding Networth next year variable.

Plotting Boxplot after treating the outlier:

Figure 5. Boxplot after Outlier treatment.

The boxplot of few of the variables are zoomed and shown below:



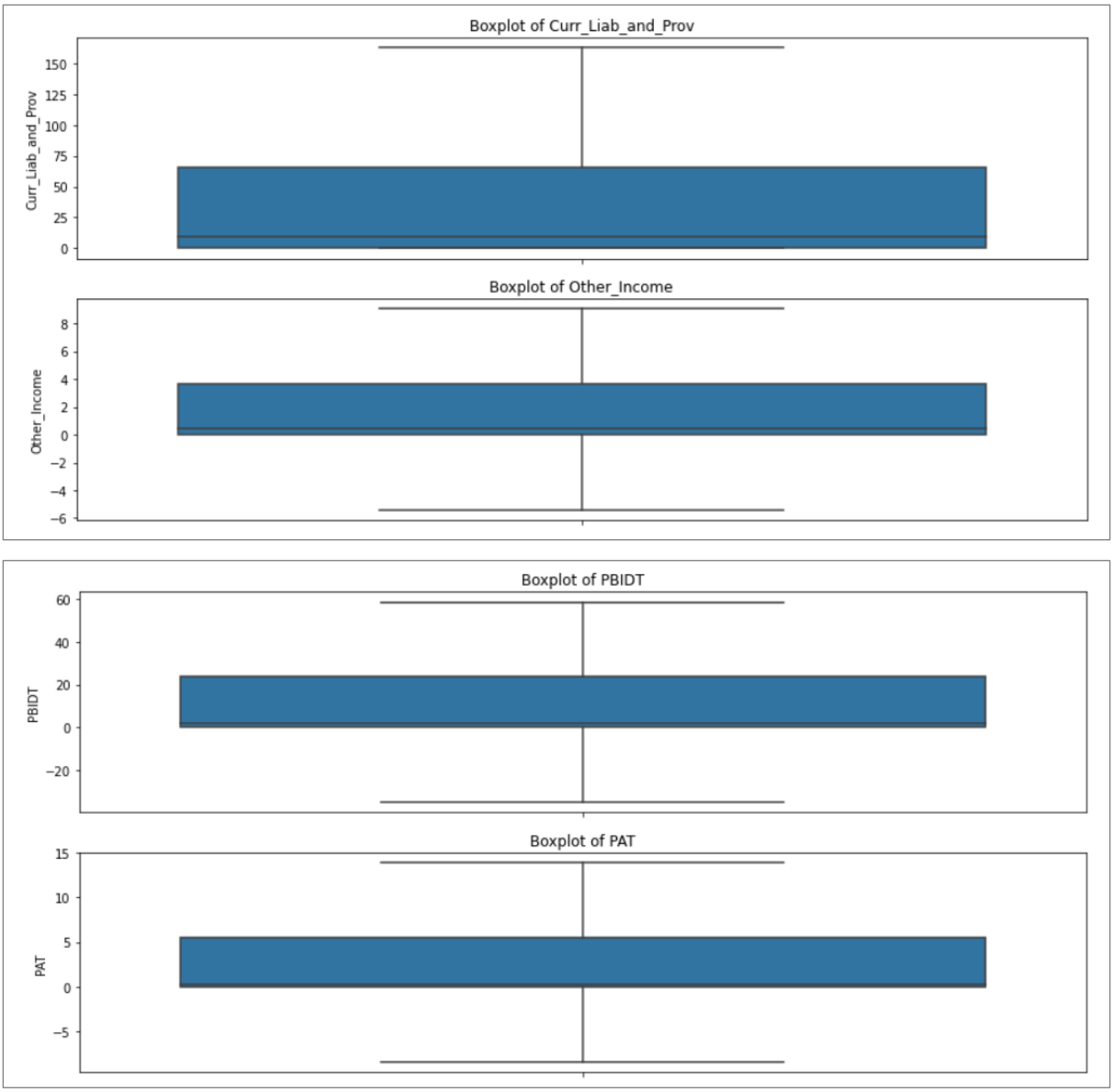


Figure 6. Zoomed Boxplot of variables after outlier treatment

1.2 Missing Value Treatment:

There are not many missing values to begin with, given the size of the data set, which are 3586 rows. In the complete dataset, there were a total of 118 missing entries.

Below is the output of missing values of all variables:

Co_Name	0
Networth_Next_Year	0
Equity_Paid_Up	0
Networth	0
Capital_Employed	0
Total_Debt	0
Gross_Block	0
Net_Working_Capital	0
Curr_Assets	0
Curr_Liab_and_Prov	0
Total_Assets_to_Liab	0
Gross_Sales	0

```
Net_Sales                              0
Other_Income                           0
Value_Of_Output                        0
Cost_of_Prod                           0
Selling_Cost                           0
PBIDT                                  0
PBDT                                   0
PBIT                                   0
PBT                                    0
PAT                                    0
Adjusted_PAT                           0
CP                                     0
Rev_earn_in_forex                      0
Rev_exp_in_forex                       0
Capital_exp_in_forex                   0
Book_Value_Unit_Curr                   0
Book_Value_Adj_Unit_Curr               4
Market_Capitalisation                  0
CEPS_annualised_Unit_Curr              0
Cash_Flow_From_Opr                     0
Cash_Flow_From_Inv                     0
Cash_Flow_From_Fin                     0
ROG_Net_Worth_perc                     0
ROG_Capital_Employed_perc              0
ROG_Gross_Block_perc                   0
ROG_Gross_Sales_perc                   0
ROG_Net_Sales_perc                     0
ROG_Cost_of_Prod_perc                  0
ROG_Total_Assets_perc                  0
ROG_PBIDT_perc                         0
ROG_PBDT_perc                          0
ROG_PBIT_perc                          0
ROG_PBT_perc                           0
ROG_PAT_perc                           0
ROG_CP_perc                            0
ROG_Rev_earn_in_forex_perc             0
ROG_Rev_exp_in_forex_perc              0
ROG_Market_Capitalisation_perc         0
Curr_Ratio_Latest                      1
Fixed_Assets_Ratio_Latest              1
Inventory_Ratio_Latest                 1
Debtors_Ratio_Latest                   1
Total_Asset_Turnover_Ratio_Latest      1
Interest_Cover_Ratio_Latest            1
PBIDTM_perc_Latest                     1
PBITM_perc_Latest                      1
PBDTM_perc_Latest                      1
CPM_perc_Latest                        1
APATM_perc_Latest                      1
Debtors_Vel_Days                       0
Creditors_Vel_Days                     0
Inventory_Vel_Days                   103
Value_of_Output_to_Total_Assets        0
Value_of_Output_to_Gross_Block         0
dtype: int64
```

- Null values were present in many columns, however significant number was present in
- "Inventory_Vel_Days" column.
- Records with missing value in "Inventory_Vel_Days" column and other columns are imputed with the median value using simple imputer of sklearn.
- No more missing values were present after treatment.

Below is the final output after imputation:

```
Co_Name                          0
Networth_Next_Year               0
Equity_Paid_Up                   0
Networth                         0
Capital_Employed                 0
Total_Debt                       0
Gross_Block                      0
Net_Working_Capital              0
Curr_Assets                      0
Curr_Liab_and_Prov               0
Total_Assets_to_Liab             0
Gross_Sales                      0
Net_Sales                        0
Other_Income                     0
Value_Of_Output                  0
Cost_of_Prod                     0
Selling_Cost                     0
PBIDT                            0
PBDT                             0
PBIT                             0
PBT                              0
PAT                              0
Adjusted_PAT                     0
CP                               0
Rev_earn_in_forex                0
Rev_exp_in_forex                 0
Capital_exp_in_forex             0
Book_Value_Unit_Curr             0
Book_Value_Adj_Unit_Curr         0
Market_Capitalisation            0
CEPS_annualised_Unit_Curr        0
Cash_Flow_From_Opr               0
Cash_Flow_From_Inv               0
Cash_Flow_From_Fin               0
ROG_Net_Worth_perc               0
ROG_Capital_Employed_perc        0
ROG_Gross_Block_perc             0
ROG_Gross_Sales_perc             0
ROG_Net_Sales_perc               0
ROG_Cost_of_Prod_perc            0
ROG_Total_Assets_perc            0
ROG_PBIDT_perc                   0
ROG_PBDT_perc                    0
ROG_PBIT_perc                    0
ROG_PBT_perc                     0
ROG_PAT_perc                     0
ROG_CP_perc                      0
ROG_Rev_earn_in_forex_perc       0
ROG_Rev_exp_in_forex_perc        0
```

```
ROG_Market_Capitalisation_perc        0
Curr_Ratio_Latest                     0
Fixed_Assets_Ratio_Latest             0
Inventory_Ratio_Latest                0
Debtors_Ratio_Latest                  0
Total_Asset_Turnover_Ratio_Latest     0
Interest_Cover_Ratio_Latest           0
PBIDTM_perc_Latest                    0
PBITM_perc_Latest                     0
PBDTM_perc_Latest                     0
CPM_perc_Latest                       0
APATM_perc_Latest                     0
Debtors_Vel_Days                      0
Creditors_Vel_Days                    0
Inventory_Vel_Days                    0
Value_of_Output_to_Total_Assets       0
Value_of_Output_to_Gross_Block        0
dtype: int64
```

1.3 Transform Target variable into 0 and 1

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.

```
conditions = [ (default['Networth_Next_Year'] < 0),
(default['Networth_Next_Year'] > 0) ]
values = ['1', '0']
default['Default'] = np.select(conditions, values)
```

We have created a new column named default to predict with output as 0 or 1.
We have transformed the value as follows,

- **1 - If the Net Worth of 2016 is less than or 0 for the company**
- **0 - If the Net Worth Next of 2016 is greater than 0 for the company**

The proportion of the dependent variable are as follows:

```
0  -  3199  -  0.89208  -  89%
1  -  387   -  0.10792  -  11%
```

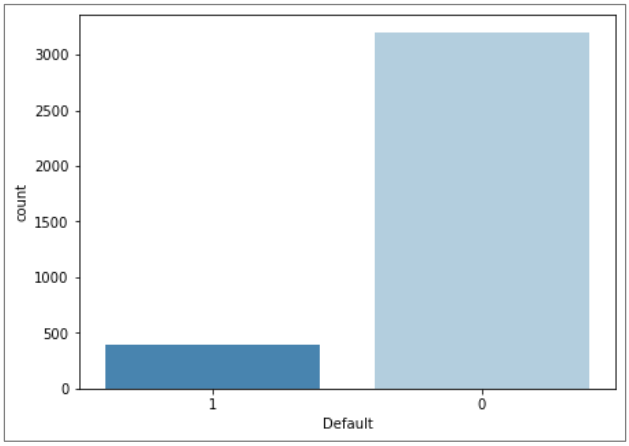


Figure 7. Bar plot of dependent variable- Default

Below is the output of dataset after appending the new variable 'Default' which is derived from the Net worth next year:

APATM_perc_Latest	Debtors_Vel_Days	Creditors_Vel_Days	Inventory_Vel_Days	Value_of_Output_to_Total_Assets	Value_of_Output_to_Gross_Block	Default
0.00	0.0	0.0	45.0	0.00	0.00	1
-87.18	29.0	101.0	2.0	0.31	0.24	1
-7961.51	97.0	210.5	0.0	-0.03	-0.26	1
-51.58	93.0	63.0	2.0	0.24	1.90	1
274.79	253.0	210.5	0.0	0.01	0.05	1

Figure 8. Head of dataset after appending default variable

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)

Data Visualization:

Let us define a function 'Univariate Analysis numeric' to display information as part of univariate analysis of numeric variables. The function will accept column name and number of bins as arguments. The function will display the statistical description of the numeric variable, histogram or distplot to view the distribution and the box plot to view 5-point summary of the variables which are significant and used for models which are selected through the VIF scores.

Univariate Analysis:

Univariate is analysis is done on the few of the variables which are significant for the model building.

Market_Capitalisation

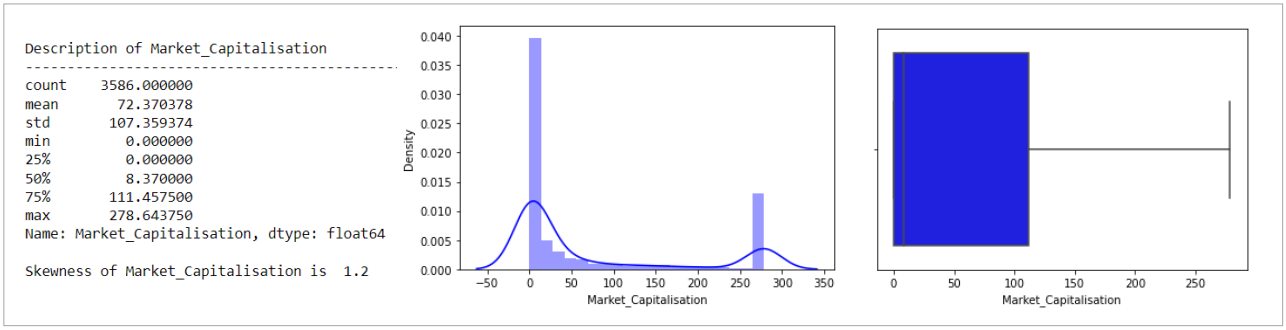


Figure 9. Boxplot and Distplot of 'Market_Capitalisation'

- From the above graphs, we can infer that mean 'Market_captilisation' is right skewed with skewness value of 1.2.
- The distplot shows the distribution of most of data from 0 to 200.
- Outliers of the 'Market_captilisation' variable have been treated.
- Median is towards the lower end of the distribution indicating the skewness.
- Mean is approximately 72 whereas median is approximately 8, showing the significant difference in this range.

Total_Debt:

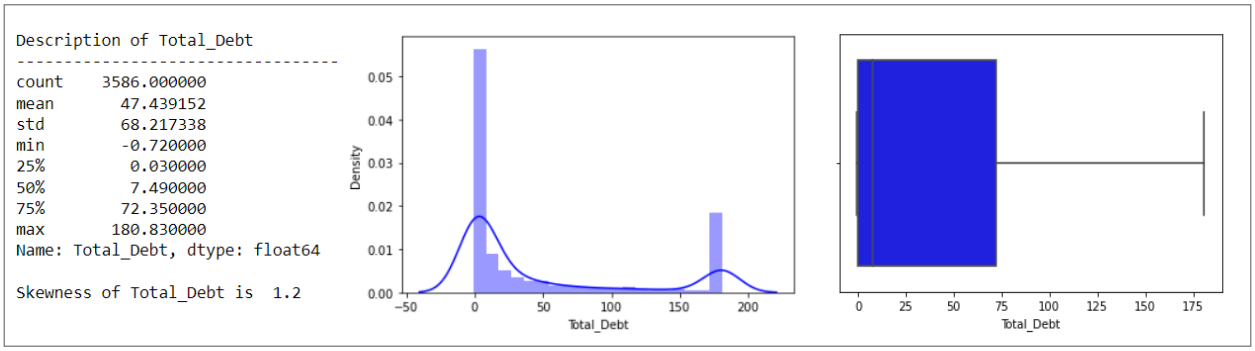


Figure 10. Boxplot and Distplot of 'Total_Debt'

- From the above graphs, we can infer that mean 'Total_Debt' is right skewed with skewness value of 1.2.
- The distplot shows the distribution of most of data from 0 to 150.
- Outliers of the 'Total_Debt' variable have been treated.
- Median is towards the lower end of the distribution indicating the skewness.
- Mean is approximately 72 whereas median is approximately 7.5, showing the significant difference in this range.

Equity_Paid_Up

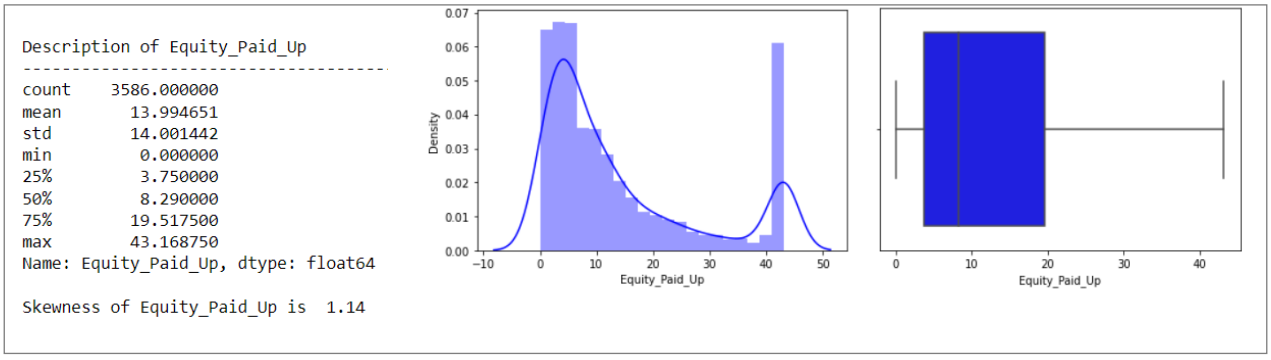


Figure 11. Boxplot and Distplot of 'Equity_Paid_Up'

- From the above graphs, we can infer that mean 'Equity_Paod_Up' is right skewed with skewness value of 1.14.
- The distplot shows the distribution of most of data from 0 to 40.
- Outliers of the 'Equity_Paod_Up' variable have been treated.
- Mean is approximately 14 whereas median is approximately 8.

ROG_Net_Worth_perc

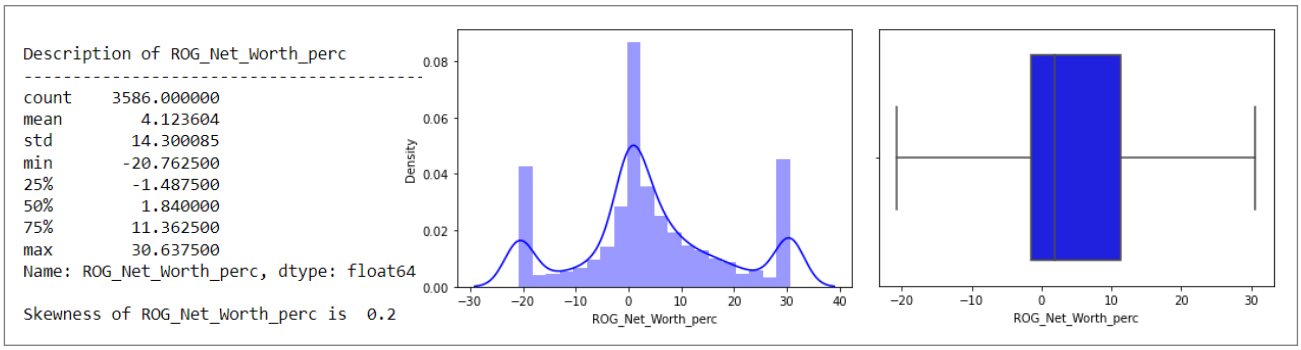


Figure 12. Boxplot and Distplot of 'ROG_Net_Worth_perc'

- From the above graphs, we can infer that mean 'ROG_Net_Worth_perc' is right skewed with skewness value of 0.2.
- The distplot shows the distribution of most of data from -20 to 30.
- The distribution is almost normally distributed.
- Outliers of the ROG_Net_Worth_perc' variable have been treated.
- Mean is approximately 4 whereas median is approximately 2, which does not have that significant difference in this range.

Cash_Flow_From_Opr

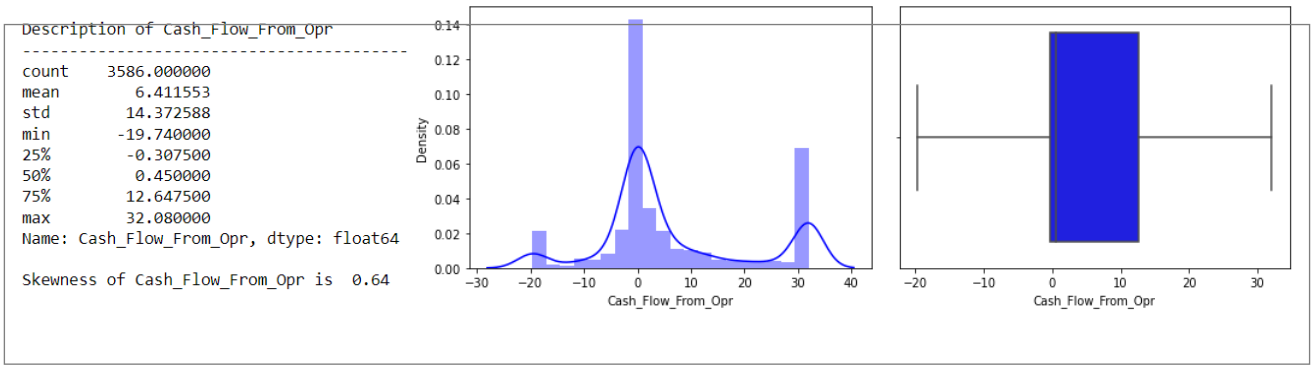


Figure 13.Boxplot and Distplot of 'Cash_Flow_From_Orp'

- From the above graphs, we can infer that mean 'Cash_Flow_From_Orp' is slightly right skewed with skewness value of 0.64.
- The distplot shows the distribution of most of data from -20 to 30.
- The distribution is almost normally distributed.
- Outliers of the 'Cash_Flow_From_Orp' variable have been treated.
- Mean is approximately 6.5 whereas median is approximately 0.5, showing the significant difference in this range.

Bivariate Analysis:

For our Bivariate analysis, we would be checking some of the important variables with respect to our target variable 'default'

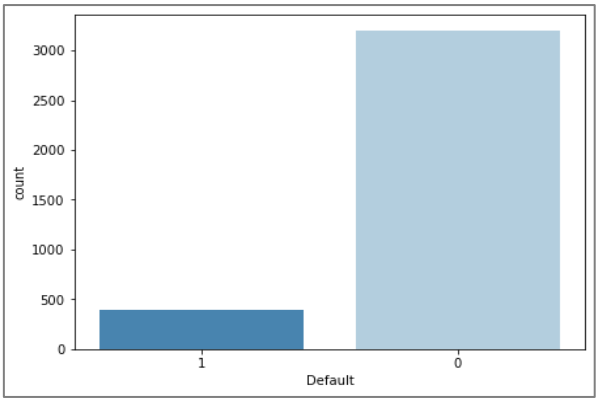


Figure 15. Count plot of dependent variable

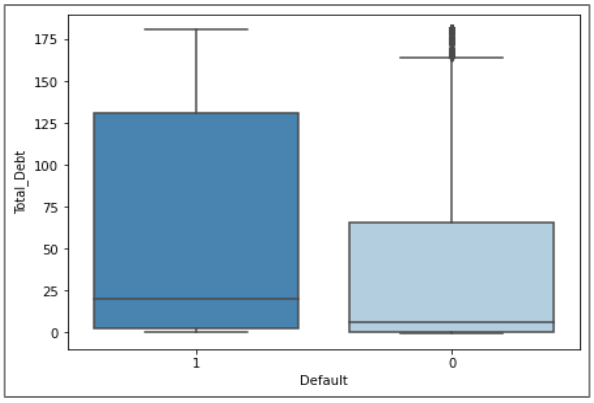


Figure 14. Boxplot of Total_debt with Default vs Default

- we can see that defaulters are very less as compared to non-defaulters. There are around 387 defaulters in comparison to around 3199 non-defaulters in the dataset.
- Checking distribution of Total Debt with respect to Default categorization
- Despite the fact that the number of defaults is lower, the overall debt of defaulters is significantly higher than that of non-defaulters providing proof for defaulting.

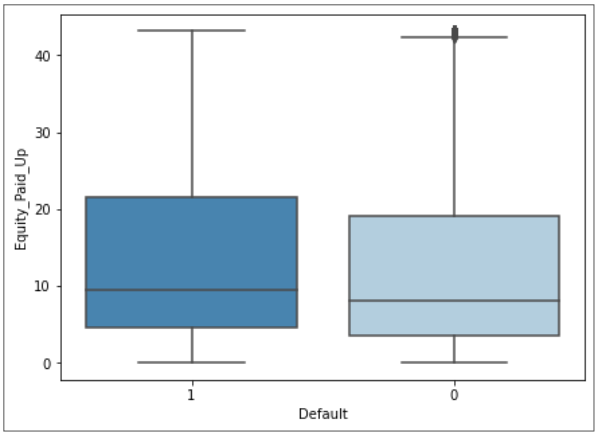


Figure 17.Boxplot of Equity_Paid_Up vs Default

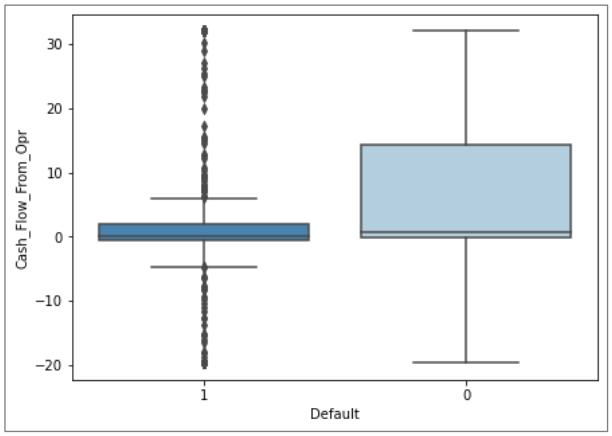


Figure 16. Boxplot of Cash_Flow_From_Opr vs Default

- Defaulters pay more equity, whereas non-defaulters pay less. Despite the fact that the number of defaulters is far less in proportion than the number of non-defaulters, the disparity is still significant.
- The Cash flow from the operations for non-defaulters is more and also, we can see many outliers in the cash flow in non-defaulters

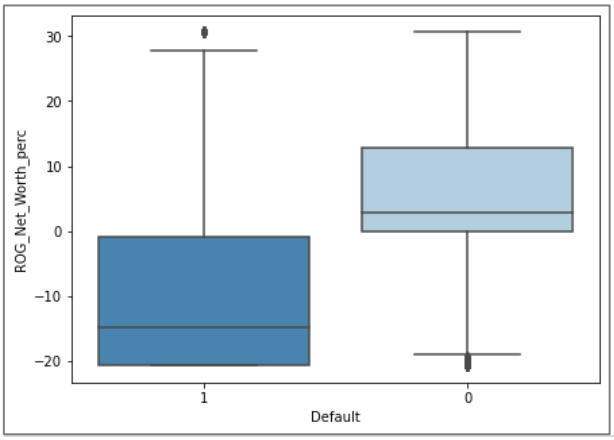


Figure 19. Boxplot of ROG_Net_Worth_perc vs Default

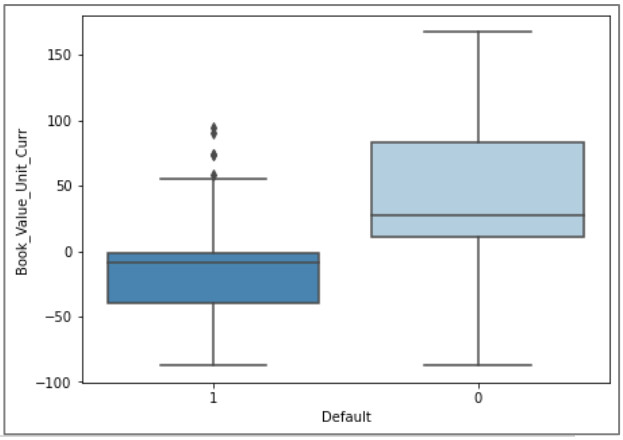


Figure 18. Boxplot of Book_Value_Unit_Curr

- The rate of growth of defaulters is in negative rate while the rate of non-defaulters is positive. Also see few of the outliers with negative rate of growth for non – defaulters as well
- Net asset value of for defaulters is mostly negative. We can also see that the net asset value for few of non-defaulters is negative as well.

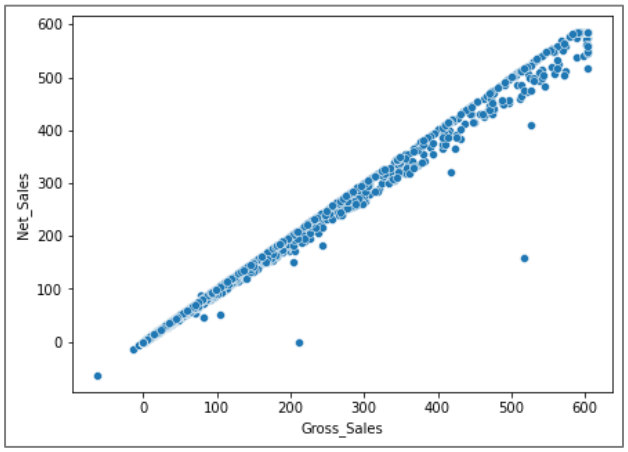


Figure 21. Scatter plot of Gross sales and Net sales

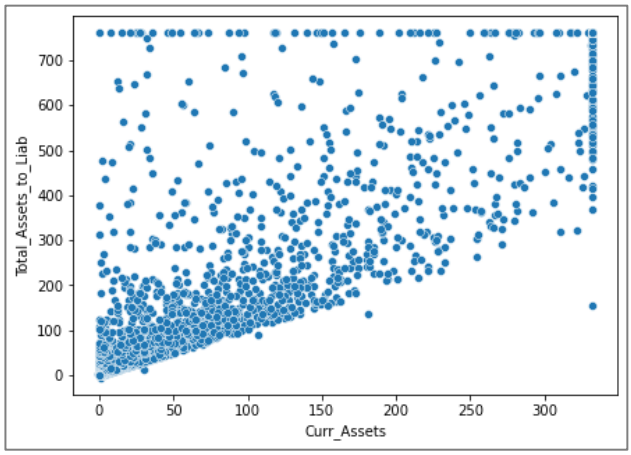
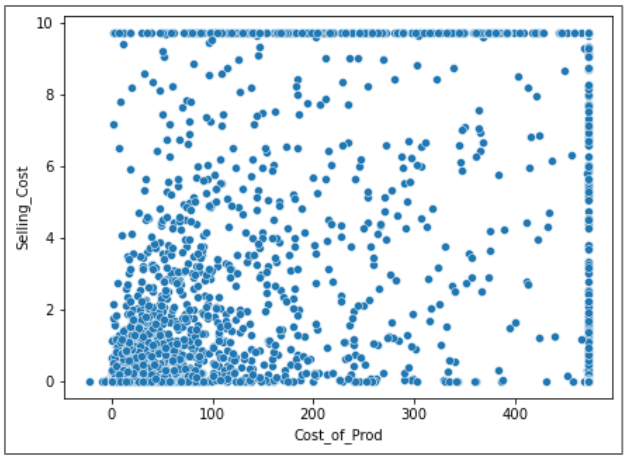
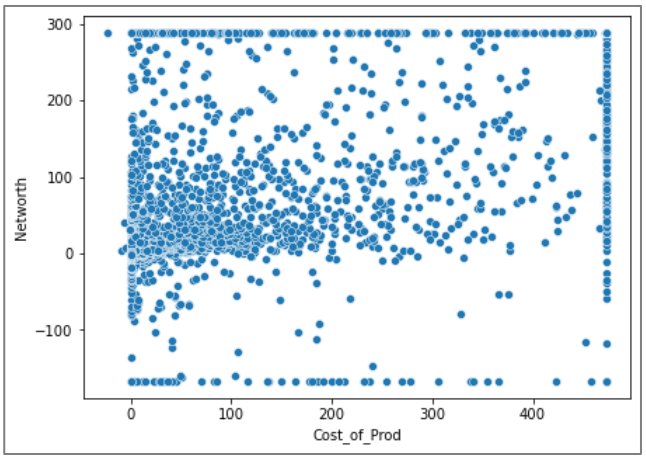


Figure 20. Scatter plot of Current Assets vs Total Assets to Liab

- The grand total of sale transactions within the accounting period is highly correlated to the net sales.
- 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 is positively correlated to the current assets of company





- The relationship between selling cost and manufacturing cost is difficult to determine. The points are dispersed in such a way that there isn't much of a connection between them. The majority of the points are distributed between 0 and 100 in terms of manufacturing cost and 0 to 4 in terms of selling cost.
- The relationship between cost of production and networth is also similar. Looks bit positively correlated but yet the data is dispersed like a cloud.

Multi-variate Analysis:

After looking at the bivariate analysis, we'll look at the dataset's multivariate analysis. Making a heat map is the best technique to perform a multivariate analysis. A heatmap depicts the relationship between all continuous variables.

Heatmap

Each square shows the correlation between the variables on each axis. Correlation ranges from -1 to +1. Values closer to zero indicate that there is no linear trend between the two variables. Closer to 1 the correlation is, more positively correlated are the variables that is as one increases so does the other. A correlation closer to -1 is similar, but instead of both increasing one variable will decrease as the other increases

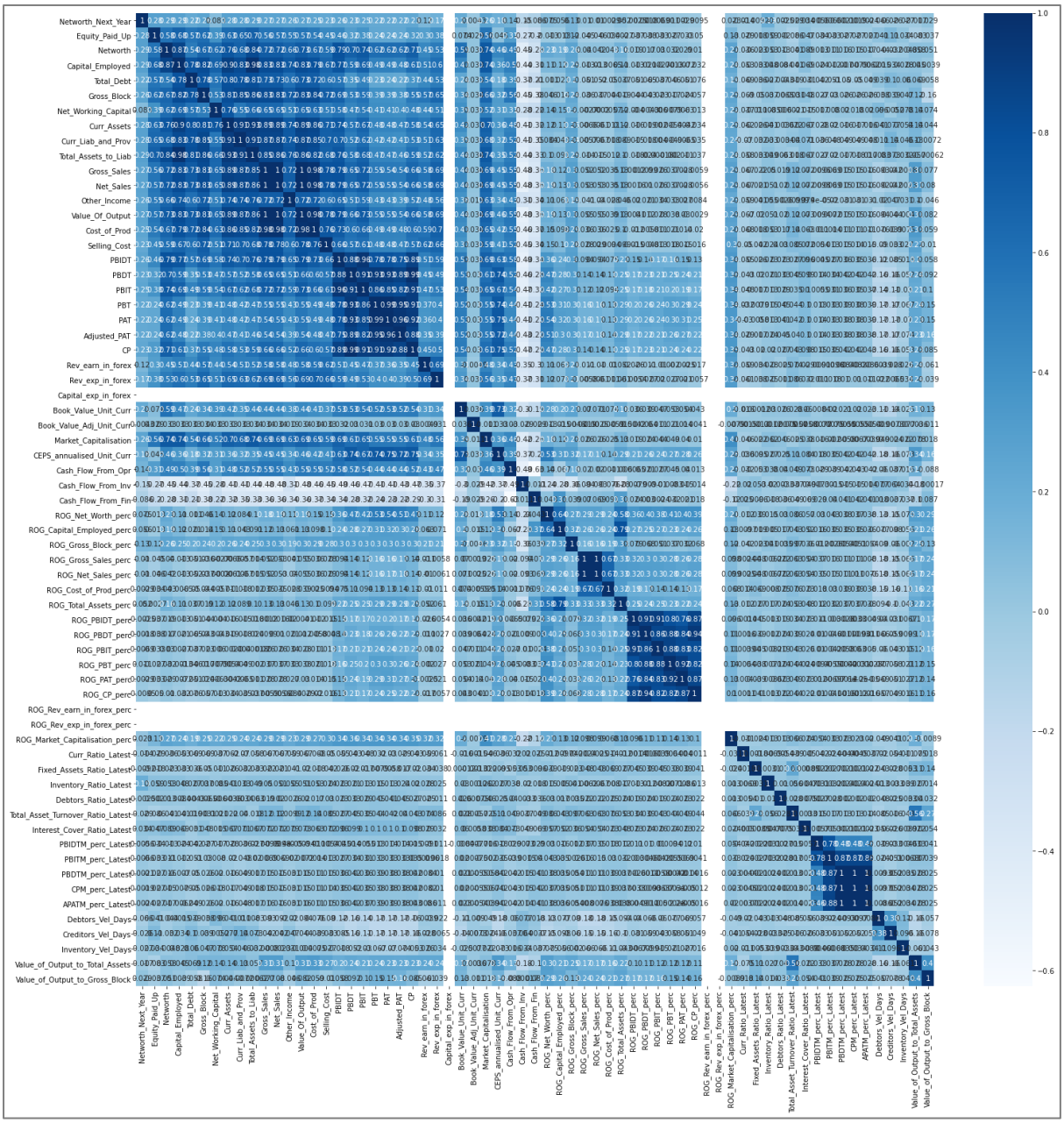


Figure 22. Heatmap

- A few variables are highly correlated, others are moderately connected, and some are negatively correlated, according to the heatmap. The blank lines indicate that the variable is empty. Before we create our models, we need remove these variables.
- Multi-Collinearity is caused by highly correlated features, and it affects the interpretability of a Logistic Regression model. It's preferable to get rid of them.
- We use Variance Inflation factor method and remove all variables with VIF > 5. This is done recursively, one-by-one.

Below is the heatmap of top significant variable:

Figure 23. Heatmap of significant variables.

1.5 Train Test Split

Capture the target column into separate vectors for training set and test set. We split the data into train and test set in the ratio 67:33 where 67% of our data (i.e., 2402 observations) will be used for training purposes and 33% (i.e., 1184 observations) will be used for testing purposes.

We are now starting the process of dividing our dataset into train and test groups. The first step is to retain the dependent variable 'Net worth next year' in 'Y' by removing it from the dataset. The rest of the data is recorded in the 'X' variable.

Only the variables that are important in developing the machine learning model are included in the X part of the dataset. By computing the VIF scores of all variables in the dataset, we can determine which variables are significant.

A VIF score of 5 or above indicates that the variable is not relevant in model construction. As a result, we'll eliminate variables with high VIF scores one by one until we're left with variables with VIF values of less than 5.

The final set of significant variables along with their VIF scores are as below:

variables	VIF	
10	Market_Capitalisation	4.112414
4	Selling_Cost	4.070446
15	ROG_Capital_Employed_perc	3.857521
1	Total_Debt	3.733792
3	Other_Income	3.662748
7	Rev_exp_in_forex	3.623937
11	Cash_Flow_From_Opr	3.584558
0	Equity_Paid_Up	3.543609
34	Value_of_Output_to_Total_Assets	3.542013
19	ROG_Total_Assets_perc	3.232673
20	ROG_PBIT_perc	3.227564
21	ROG_CP_perc	3.224693
6	Rev_earn_in_forex	3.038739
5	Adjusted_PAT	2.891579
13	Cash_Flow_From_Fin	2.846712
2	Net_Working_Capital	2.760406
14	ROG_Net_Worth_perc	2.683281
12	Cash_Flow_From_Inv	2.421594
8	Book_Value_Unit_Curr	2.370061
35	Value_of_Output_to_Gross_Block	2.347341

31	Debtors_Vel_Days	2.139876
17	ROG_Net_Sales_perc	2.124237
32	Creditors_Vel_Days	2.120980
18	ROG_Cost_of_Prod_perc	2.017257
27	Total_Asset_Turnover_Ratio_Latest	1.919566
22	ROG_Market_Capitalisation_perc	1.625425
16	ROG_Gross_Block_perc	1.535765
33	Inventory_Vel_Days	1.336250
30	CPM_perc_Latest	1.331267
29	PBIDTM_perc_Latest	1.328937
24	Fixed_Assets_Ratio_Latest	1.097535
28	Interest_Cover_Ratio_Latest	1.045214
25	Inventory_Ratio_Latest	1.022330
23	Curr_Ratio_Latest	1.021412
26	Debtors_Ratio_Latest	1.019020
9	Book_Value_Adj_Unit_Curr	1.011783

Table 2. Final VIF score table of significant variables

So, we've got our X and y datasets ready, and we're going to do the train-test split.

The dataset is divided into train and test groups in a 67-33 ratio, with training data accounting for 67 percent and test data for 33 percent. The number 42 represents the random situation indicated below.

- Train dataset has 2402 records (67% of the total dataset)
- Test dataset has 1184 records (33% of the total dataset)

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

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. The classification algorithm Logistic Regression is used to predict the likelihood of a categorical dependent variable.

By estimating probabilities using a logistic regression equation, it is used in statistical software to comprehend the connection between the dependent variable and one or more independent variables. This form of analysis can aid in predicting the chance of an occurrence or a decision occurring.

The form of regression analysis used to determine the likelihood of a specific event occurring is known as logistic regression. When we have a categorical dependent variable that can only take discrete values, this is the optimal sort of regression to use.

The logit() method is used to create a logistic regression model, which is then fitted using statsmodel.

We create the first model by combining all of the factors that were deemed to be important following the VIF score computations. The first model's summary is as follows:

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	2402
Model:	Logit	**Df Residuals:**	2365
Method:	MLE	**Df Model:**	36
Date:	Sat, 11 Jun 2022	**Pseudo R-squ.:**	0.6321
Time:	19:02:46	**Log-Likelihood:**	-291.12
converged:	True	**LL-Null:**	-791.34
Covariance Type:	nonrobust	**LLR p-value:**	1.270e-186

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.2150	0.244	-0.881	0.378	-0.693	0.263
Market_Capitalisation	-0.0040	0.002	-1.660	0.097	-0.009	0.001
Selling_Cost	0.0281	0.070	0.399	0.690	-0.110	0.166
ROG_Capital_Employed_perc	0.0127	0.010	1.315	0.189	-0.006	0.032
Total_Debt	0.0114	0.004	3.104	0.002	0.004	0.019
Other_Income	0.0161	0.057	0.281	0.779	-0.096	0.128
Rev_exp_in_forex	0.0330	0.031	1.052	0.293	-0.028	0.094
Equity_Paid_Up	-0.0269	0.012	-2.198	0.028	-0.051	-0.003
Cash_Flow_From_Opr	-0.0381	0.021	-1.851	0.064	-0.078	0.002
Value_of_Output_to_Total_Assets	-0.1207	0.221	-0.545	0.586	-0.555	0.313
ROG_Total_Assets_perc	-0.0181	0.010	-1.789	0.074	-0.038	0.002
ROG_PBIT_perc	0.0004	0.002	0.208	0.835	-0.004	0.004
ROG_CP_perc	-0.0013	0.002	-0.653	0.514	-0.005	0.003

Rev_earn_in_forex	-0.0028	0.029	-0.097	0.923	-0.060	0.054
Adjusted_PAT	-0.0419	0.040	-1.046	0.296	-0.120	0.037
Cash_Flow_From_Fin	7.516e-05	0.036	0.002	0.998	-0.071	0.071
Net_Working_Capital	-0.0005	0.004	-0.152	0.880	-0.007	0.006
ROG_Net_Worth_perc	-0.0185	0.010	-1.829	0.067	-0.038	0.001
Cash_Flow_From_Inv	-0.0406	0.039	-1.039	0.299	-0.117	0.036
Book_Value_Unit_Curr	-0.0971	0.041	-2.377	0.017	-0.177	-0.017
Value_of_Output_to_Gross_Block	-0.1172	0.048	-2.419	0.016	-0.212	-0.022
Debtors_Vel_Days	-0.0029	0.001	-2.141	0.032	-0.006	-0.000
Creditors_Vel_Days	0.0019	0.001	1.341	0.180	-0.001	0.005
ROG_Net_Sales_perc	-0.0024	0.004	-0.565	0.572	-0.011	0.006
ROG_Cost_of_Prod_perc	0.0004	0.004	0.090	0.929	-0.008	0.008
Total_Asset_Turnover_Ratio_Latest	0.0449	0.042	1.074	0.283	-0.037	0.127
ROG_Market_Capitalisation_perc	-0.0020	0.003	-0.737	0.461	-0.007	0.003
ROG_Gross_Block_perc	-0.0128	0.020	-0.642	0.521	-0.052	0.026
Inventory_Vel_Days	0.0008	0.001	1.062	0.288	-0.001	0.002
CPM_perc_Latest	-5.728e-05	2.92e-05	-1.959	0.050	-0.000	1.67e-08
PBIDTM_perc_Latest	-5.273e-05	0.000	-0.456	0.648	-0.000	0.000
Fixed_Assets_Ratio_Latest	-3.236e-05	0.001	-0.042	0.966	-0.002	0.001
Inventory_Ratio_Latest	-0.0009	0.001	-0.738	0.461	-0.003	0.002
Interest_Cover_Ratio_Latest	-0.0016	0.001	-1.305	0.192	-0.004	0.001
Curr_Ratio_Latest	-0.3393	0.083	-4.070	0.000	-0.503	-0.176
Debtors_Ratio_Latest	-0.0093	0.004	-2.575	0.010	-0.016	-0.002
Book_Value_Adj_Unit_Curr	-0.0206	0.041	-0.505	0.614	-0.101	0.059

Possibly complete quasi-separation: A fraction 0.38 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

We check the p values of all variables used for this model's development using this summary. The level of significance (alpha) has been set at 0.05. So, if a variable's p value is greater than 0.05, we can state with 95 percent confidence that the variable isn't significant in the model we're building.

We'll start by removing the variables with the highest p values from the model one by one , until we have all variables with p values less than 0.05.

Recursively building one by one model by eliminating the variables with p-value > 0.05. Consequently built 27 models one by one and arrived with all the variables less than 0.05 and which are significant for the model building.

Below is the summary of the final model with all the significant variables:

Table 3. Final summary of model

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	2402
Model:	Logit	**Df Residuals:**	2391
Method:	MLE	**Df Model:**	10
Date:	Sat, 11 Jun 2022	**Pseudo R-squ.:**	0.6187
Time:	19:03:21	**Log-Likelihood:**	-301.73
converged:	True	**LL-Null:**	-791.34
Covariance Type:	nonrobust	**LLR p-value:**	5.555e-204

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.3291	0.184	-1.786	0.074	-0.690	0.032
Market_Capitalisation	-0.0041	0.002	-1.970	0.049	-0.008	-2.18e-05
Total_Debt	0.0152	0.003	5.391	0.000	0.010	0.021
Equity_Paid_Up	-0.0229	0.011	-2.030	0.042	-0.045	-0.001
Cash_Flow_From_Opr	-0.0330	0.011	-2.942	0.003	-0.055	-0.011
ROG_Net_Worth_perc	-0.0272	0.007	-3.646	0.000	-0.042	-0.013
Book_Value_Unit_Curr	-0.1214	0.010	-11.705	0.000	-0.142	-0.101
Value_of_Output_to_Gross_Block	-0.1300	0.042	-3.111	0.002	-0.212	-0.048
CPM_perc_Latest	-6.248e-05	2.01e-05	-3.103	0.002	-0.000	-2.3e-05
Curr_Ratio_Latest	-0.2681	0.046	-5.800	0.000	-0.359	-0.178
Debtors_Ratio_Latest	-0.0087	0.004	-2.470	0.014	-0.016	-0.002

Possibly complete quasi-separation: A fraction 0.37 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

The top 10 important variables on which our final model is built are as follows:

1) Market Capitalisation
2) Total Debt

3) Equity Paid Up
4) Cash Flow from Opr
5) ROG Net Worth perc
6) Book Value Unit Curr
7) Value of Output to gross Block
8) CPM perc Latest
9) Curr Ratio Latest
10) Debtors Ratio Latest

The probability of an event occurring is calculated using a logistic regression model. Now we need to figure out what the best cut-off value is for predicting whether a company would default (1) or not (0).

The roc curve function is used to determine the threshold value. The optimal threshold, as estimated by the roc curve () method which is 0.21472.

Value of the Optimal Threshold = 0.21472

We examine the probabilities obtained using the logistic function and compare them to the roc curve threshold value (). If the likelihood is more than 0.21 (the threshold value), the company is classed as 'default' (1); if the probability is less than the threshold, the company is classified as 'non-default' (0). On the train and test datasets, this is how we obtain predictions.

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

Model performance helps to understand how good the model that we have trained using the dataset is so that we have confidence in the performance of the model for future predictions.

We evaluate our models' performance on train and test datasets once they've been constructed. We try to determine if the model is underfitting or overfitting by checking for accuracy, precision, and other factors. We have specific scores and matrices for our model's performance. Following are the methods used to evaluate the model performance:

- Confusion Matrix
- Classification Report
    - Accuracy
    - Precision
    - Recall
    - F1 Score

1. Confusion Matrix:

This gives us how many zeros (0s) i.e. (class = No claim) and ones (1s) i.e. (class = Yes claim) were correctly predicted by our model and how many were wrongly predicted.

	Predicted Class		
Actual class		Class = No	Class = Yes
	Class = No	True Negative	False Positive
	Class = yes	False Negative	True Positive

True positive and true negatives are the observations that are correctly predicted and therefore shown in green. We want to minimize false positives and false negatives so they are shown in red color.

2. Accuracy :

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

3. Precision:

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision = TP/(TP + FP)

4. Recall (Sensitivity):

Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

Recall = TP/(TP + FN)

5. F1 Score:

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. That is, a good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1 , while the model is a total failure when it's 0

F1 score = 2 x [(Precision x Recall) / (Precision + Recall)]

Companies that default are represented by 1 in our dependent variable 'Default,' whereas companies that do not default are represented by 0. In this scenario, True Positives are companies that are indeed defaulters, which our model accurately predicted, and True Negatives are companies that are not defaulters, which our model also correctly anticipated.

False positives, on the other hand, are non-defaulters who were predicted to default by the model. False Negatives, on the other side, are those who do default but are predicted as non-defaulters by our model.

In this case, if the model predicts a defaulter company as a non-defaulter, it will have a greater impact than if it predicts the opposite. As a result, we'll have to concentrate more on the recall value for this situation.

Checking the model performance of Logistic Regression Model at optimum threshold:

Model performance of train data:

Confusion Matrix:

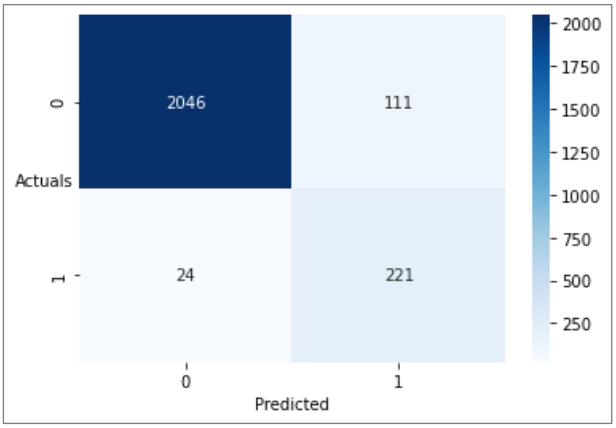


Figure 24. Confusion matrix for train data

Classification Report:

	precision	recall	f1-score	support
0	0.988	0.949	0.968	2157
1	0.666	0.902	0.766	245
accuracy			0.944	2402
macro avg	0.827	0.925	0.867	2402
weighted avg	0.955	0.944	0.947	2402

Figure 25. Classification report for train data

Model performance of test data:

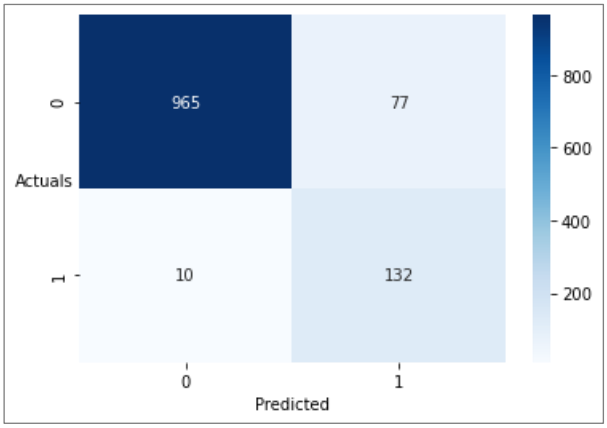


Figure 26. Confusion matrix for test data

	precision	recall	f1-score	support
0	0.990	0.926	0.957	1042
1	0.632	0.930	0.752	142
accuracy			0.927	1184
macro avg	0.811	0.928	0.855	1184
weighted avg	0.947	0.927	0.932	1184

Figure 27. Classification report for test data

Model Performance					
Sl. No		**Train Data**		**Test Data**	
1.	True Positive	211		132	
2.	True Negative	2046		965	
3.	False Positive	111		77	
4.	False Negative	24		10	
6.	Accuracy	94.4%		92.7%	
		Defaulter (1)	**Non-Defaulters (0)**	**Defaulter (1)**	**Non-Defaulters (0)**
7.	Precision	67%	99%	63%	99%
8.	Recall	90%	95%	93%	93%
9.	F1 score	77%	97%	75%	96%

Table 4. Model Performance

The accuracy of the test dataset is less than 10% of that of the train dataset, indicating that the dataset is not overfitting. The model is a decent classification model in general. Furthermore, the derived metrics have high values and are well-fitting.

Interpretation from the model:

- Recall of 95% means - 95% of Actual Defaults were Predicted Correctly
- Precision of 77% means - 77% of Predicted Defaults were Actual
- For this modelling, we needed to predict as many of Actual Defaults as possible and minimise Type 2 errors for predictions
- Hence Recall and Precision are considered in choosing the best model.
- In Table 3 above, coefficients of all variables indicate the weightage of that variable in predictions of Defaulters and non-defaulters.
- Positive coefficient means, higher value of the variable will lead to higher likelihood for default.
- Negative coefficient means, higher value of this variable will lead to lower likelihood for default.
- We note that highest negative value is in our variables is ‘Curr_Ratio_Latest’ suggesting if ‘Curr_Ratio_Latest’ - Liquidity ratio, company's ability to pay short-term obligations or those due within one year increases then the Probability of Default by that Company decreases.

- Also, positive coefficient is of ‘Total debt’ suggesting if total debt of a Company increases, then the Probability of Default by that Company increases.
