FRA MILESTONE – 1

Finance & Risk Analysis

PROJECT: CREDIT RISK ANALYSIS

A credit risk is risk of default on a debt that may arise from a borrower failing to make required payments. In the first resort, the risk is that of the lender and includes lost principal and interest, disruption to cash flows, and increased collection costs. The loss may be complete or partial.

MANISHA PARAKANDLA. PGP-DSBA-ONLINE 08-05-2022

Contents:

1.1 Outlier Treatment 6

1.2 Missing Value Treatment 9

1.3 Transform Target variable into 0 and 1 9

1.4 Univariate (4 marks) & Bivariate 10

1.5 Train Test Split 18

1.6 Build Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff. 19

1.7 Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model 21

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.

Dataset for the problem: Credit Risk Dataset

Data Dictionary: 'Credit Default Data Dictionary.xlsx'

Exploratory data Analysis:

→First, we import the necessary libraries and read the data for the further Analysis.

→SAMPLE OF THE DATA:

	Co_Code	Co_Name	Networth Next Year	Equity Paid Up	Networth	Capital Employed	Total Debt	Gross Block	Net Working Capital	Current Assets	...	PBIDTM (%) [Latest]	PBITM (%) [Latest]	PBDTM (%) [Latest]	CPM (%) [Latest]
0	16974	Hind.Cables	-8021.60	419.36	-7027.48	-1007.24	5936.03	474.30	-1076.34	40.50	...	0.00	0.00	0.00	0.00
1	21214	Tata Tele. Mah.	-3986.19	1954.93	-2968.08	4458.20	7410.18	9070.86	-1098.88	486.86	...	-10.30	-39.74	-57.74	-57.74
2	14852	ABG Shipyard	-3192.58	53.84	506.86	7714.68	6944.54	1281.54	4496.25	9097.64	...	-5279.14	-5516.98	-7780.25	-7723.67
3	2439	GTL	-3054.51	157.30	-623.49	2353.88	2326.05	1033.69	-2612.42	1034.12	...	-3.33	-7.21	-48.13	-47.70
4	23505	Bharati Defence	-2967.36	50.30	-1070.83	4675.33	5740.90	1084.20	1836.23	4685.81	...	-295.55	-400.55	-845.88	379.79
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3581	4987	HDFC Bank	72677.77	501.30	62009.42	590576.00	496009.19	8463.30	0.00	444633.50	...	0.00	0.00	0.00	0.00
3582	502	Vedanta	79162.19	296.50	34057.87	71906.06	37643.79	29848.44	2503.86	11554.45	...	39.92	32.17	29.81	30.52
3583	12002	I O C L	88134.31	2427.95	67969.97	140686.75	55245.01	121643.45	6376.84	89609.82	...	8.09	6.69	7.31	5.69
3584	12001	NTPC	91293.70	8245.46	81657.35	173099.14	85995.34	128477.59	11449.79	42353.59	...	28.12	20.55	23.39	19.55
3585	15542	Bharti Airtel	111729.10	1998.70	78270.80	104241.00	21569.70	100084.90	-12145.30	11947.10	...	42.47	22.88	34.04	25.97

3586 rows × 67 columns

The dataset consists of 3586 rows and 67 columns.

The data dictionary consists of the detailed explanation of each feature.

Data Dictionary

The 67 variables are listed below:

```
Index(['Co_Code', 'Co_Name', 'Networth_Next_Year', 'Equity_Paid_Up',
       'Networth', 'Capital_Employed', 'Total_Debt', 'Gross_Block',
       'Net_Working_Capital', 'Curr_Assets', 'Curr_Liab_and_Prov',
       'Total_Assets_to_Liab', 'Gross_Sales', 'Net_Sales', 'Other_Income',
       'Value_Of_Output', 'Cost_of_Prod', 'Selling_Cost', 'PBIDT', 'PBDT',
       'PBIT', 'PBT', 'PAT', 'Adjusted_PAT', 'CP', 'Rev_earn_in_forex',
       'Rev_exp_in_forex', 'Capital_exp_in_forex', 'Book_Value_Unit_Curr',
       'Book_Value_Adj_Unit_Curr', 'Market_Capitalisation',
       'CEPS_annualised_Unit_Curr', 'Cash_Flow_From_Opr', 'Cash_Flow_From_Inv'
       'Cash_Flow_From_Fin', 'ROG_Net_Worth_perc', 'ROG_Capital_Employed_perc'
       'ROG_Gross_Block_perc', 'ROG_Gross_Sales_perc', 'ROG_Net_Sales_perc',
       'ROG_Cost_of_Prod_perc', 'ROG_Total_Assets_perc', 'ROG_PBIDT_perc',
       'ROG_PBDT_perc', 'ROG_PBIT_perc', 'ROG_PBT_perc', 'ROG_PAT_perc',
       'ROG_CP_perc', 'ROG_Rev_earn_in_forex_perc',
       'ROG_Rev_exp_in_forex_perc', 'ROG_Market_Capitalisation_perc',
       'Curr_Ratio_Latest', 'Fixed_Assets_Ratio_Latest',
       'Inventory_Ratio_Latest', 'Debtors_Ratio_Latest',
       'Total_Asset_Turnover_Ratio_Latest', 'Interest_Cover_Ratio_Latest',
       'PBIDTM_perc_Latest', 'PBITM_perc_Latest', 'PBDTM_perc_Latest',
       'CPM_perc_Latest', 'APATM_perc_Latest', 'Debtors_Vel_Days',
       'Creditors_Vel_Days', 'Inventory_Vel_Days',
       'Value_of_Output_to_Total_Assets', 'Value_of_Output_to_Gross_Block'],
      dtype='object')
```

→There are no duplicate values present in the data.

→Let's check the info () of the data

```
<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
```

we can see that, the dataset has 67 variables of which 63 are of float data type, 3 are integer type and 1 is object type.

Descriptive summary of the data:

	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block	Net_Working_Capital	Curr_Assets	Curr_Liab_and_Prov	Total_Assets_to_Liab
count	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000
mean	13.994651	73.691121	152.489564	47.439152	85.099863	36.800227	91.366022	43.518483	208.542427
std	14.001442	112.937394	207.868131	68.217338	121.230109	59.347261	122.085815	60.878782	280.419359
min	0.000000	-166.215000	-320.901250	-0.720000	-41.190000	-89.406250	-0.910000	-0.230000	-4.510000
25%	3.750000	3.892500	7.602500	0.030000	0.570000	0.942500	4.000000	0.732500	10.555000
50%	8.290000	18.580000	39.090000	7.490000	15.870000	10.145000	24.540000	9.225000	52.010000
75%	19.517500	117.297500	226.605000	72.350000	131.895000	61.175000	135.277500	65.650000	310.540000
max	43.168750	287.405000	555.108750	180.830000	328.882500	151.523750	332.193750	163.026250	760.517500

8 rows × 65 columns

Gross_Sales	...	PBITM_perc_Latest	PBDTM_perc_Latest	CPM_perc_Latest	APATM_perc_Latest	Debtors_Vel_Days	Creditors_Vel_Days	Inventory_Vel_Days
3586.000000	...	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000
159.511180	...	7.213992	6.667837	5.320970	2.821170	75.286670	62.440742	61.223926
224.244869	...	15.281989	15.111930	12.291001	8.728962	81.861954	68.144543	73.203385
-62.590000	...	-21.435000	-21.165000	-17.085000	-11.115000	0.000000	0.000000	-144.000000
1.442500	...	0.000000	0.000000	0.000000	0.000000	8.000000	8.000000	0.000000
31.210000	...	5.230000	4.690000	3.890000	1.590000	49.000000	39.000000	35.000000
242.250000	...	14.285000	14.100000	11.387500	7.407500	106.000000	89.000000	93.000000
603.461250	...	35.725000	35.275000	28.475000	18.525000	253.000000	210.500000	240.000000

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
3586.000000	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000	3586.000000
2.821170	75.286670	62.440742	61.223926	0.730884	3.356559	0.107920
8.728962	81.861954	68.144543	73.203385	0.774996	4.101482	0.310322
-11.115000	0.000000	0.000000	-144.000000	-0.330000	-6.690000	0.000000
0.000000	8.000000	8.000000	0.000000	0.070000	0.270000	0.000000
1.590000	49.000000	39.000000	35.000000	0.480000	1.530000	0.000000
7.407500	106.000000	89.000000	93.000000	1.160000	4.910000	0.000000
18.525000	253.000000	210.500000	240.000000	2.795000	11.870000	1.000000

with the describe () function, we can understand the five number summary of the data for all the continuous numeric variables. The values of mean, standard deviation, minimum and maximum, 25th, 50th and 75th percentile mentioned in the above tables. In the next step I have checked the skewness of the data.

Dropping Co_Code & Co_Name columns as they do not add any value to the analysis.

	Networth_Next_Year	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block	Net_Working_Capital	Curr_Assets	Curr_Liab_and_Prov	T
0	-8021.60	419.36	-7027.48	-1007.24	5936.03	474.30	-1076.34	40.50	1116.85	
1	-3986.19	1954.93	-2968.08	4458.20	7410.18	9070.86	-1098.88	486.86	1585.74	
2	-3192.58	53.84	506.86	7714.68	6944.54	1281.54	4496.25	9097.64	4601.39	
3	-3054.51	157.30	-623.49	2353.88	2326.05	1033.69	-2612.42	1034.12	3646.54	
4	-2967.36	50.30	-1070.83	4675.33	5740.90	1084.20	1836.23	4685.81	2849.58	

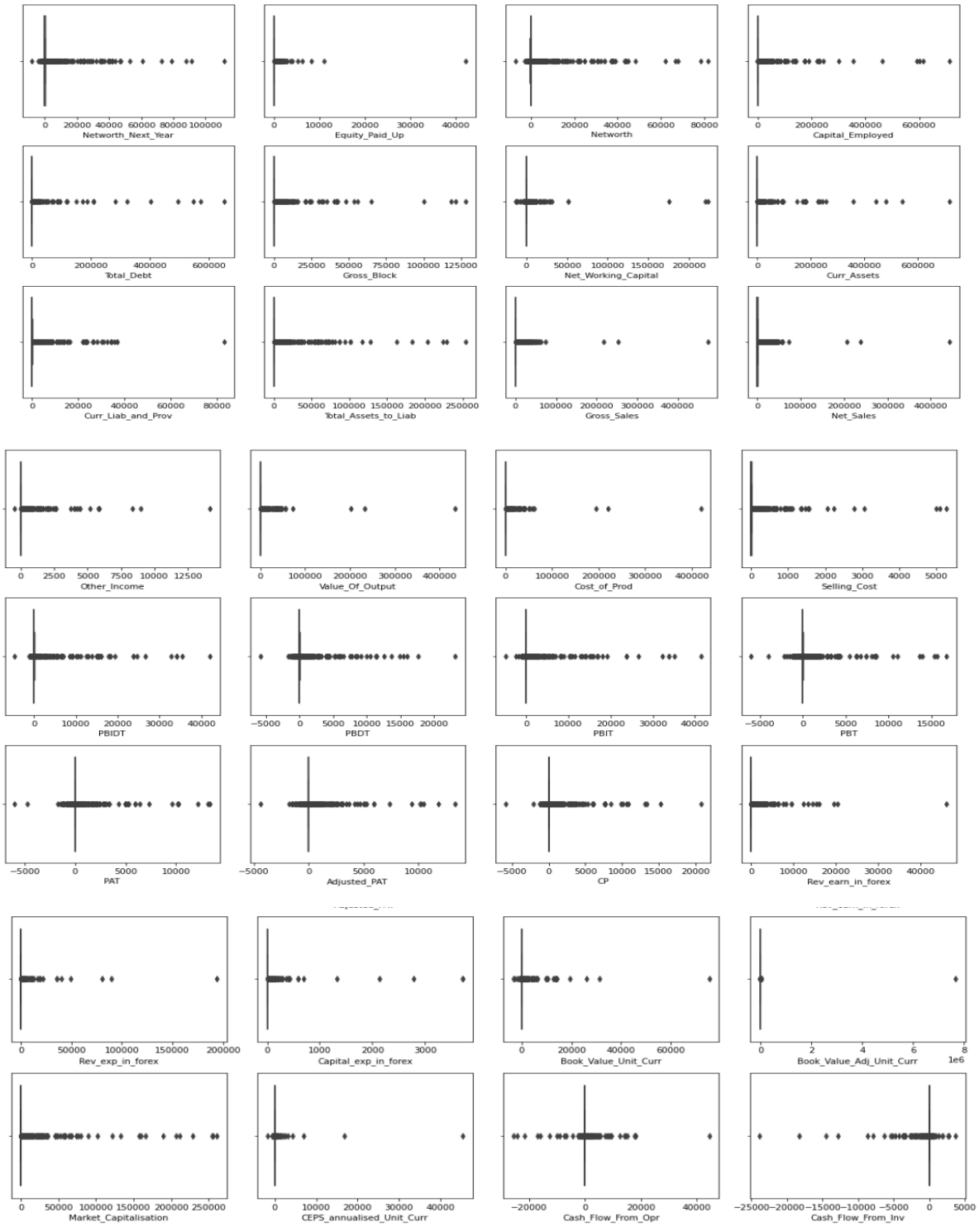
5 rows × 65 columns

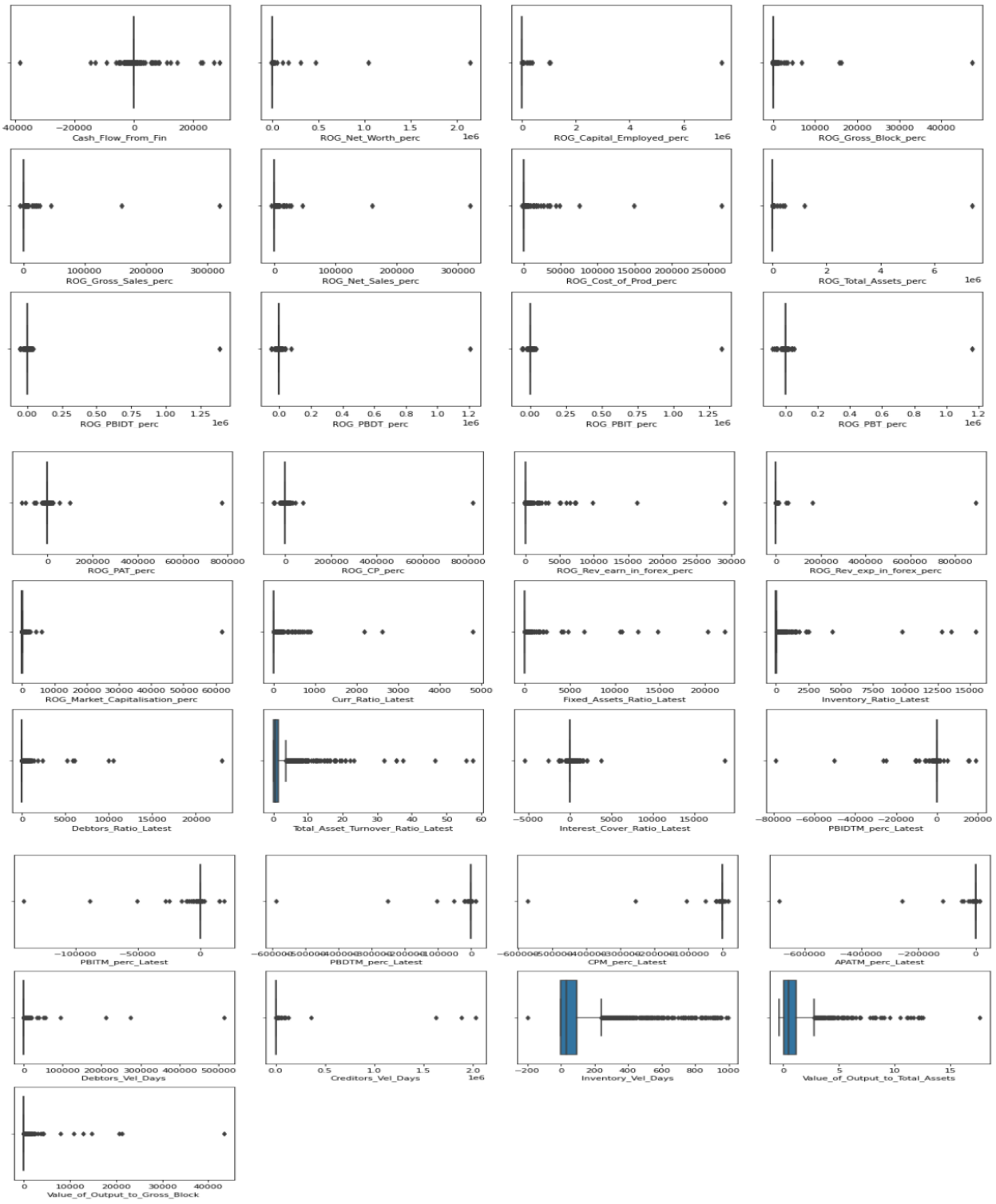
The values of mean, standard deviation, minimum and maximum, 25th, 50th and 75th percentile mentioned in the above tables.

In the next step I have checked the skewness of the data.

1.1 Outlier Treatment

Let's check for outliers in the data using boxplot:

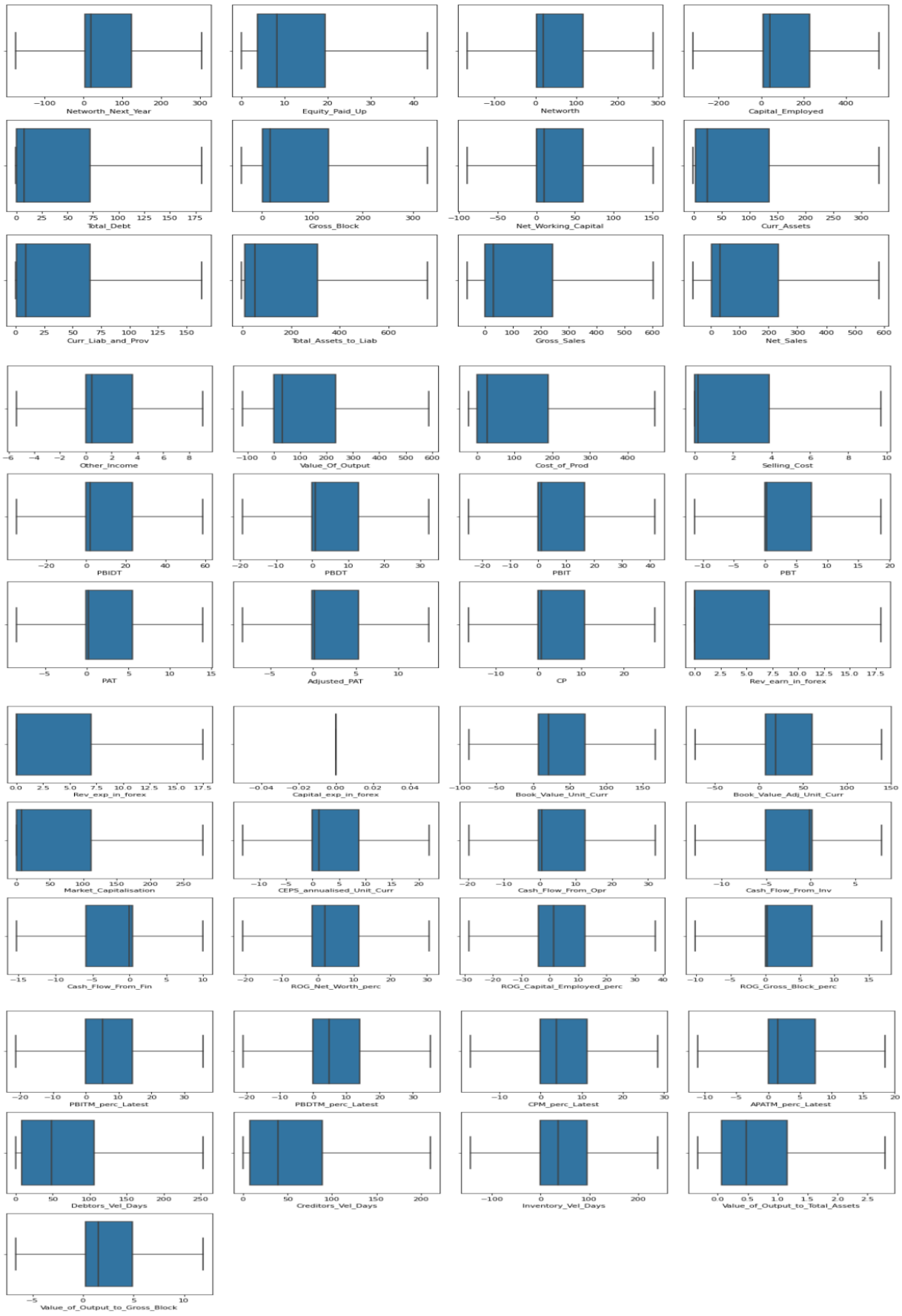




We can see that there are outliers present in almost each feature.

There are 18 % datapoints that are outliers in the dataset.

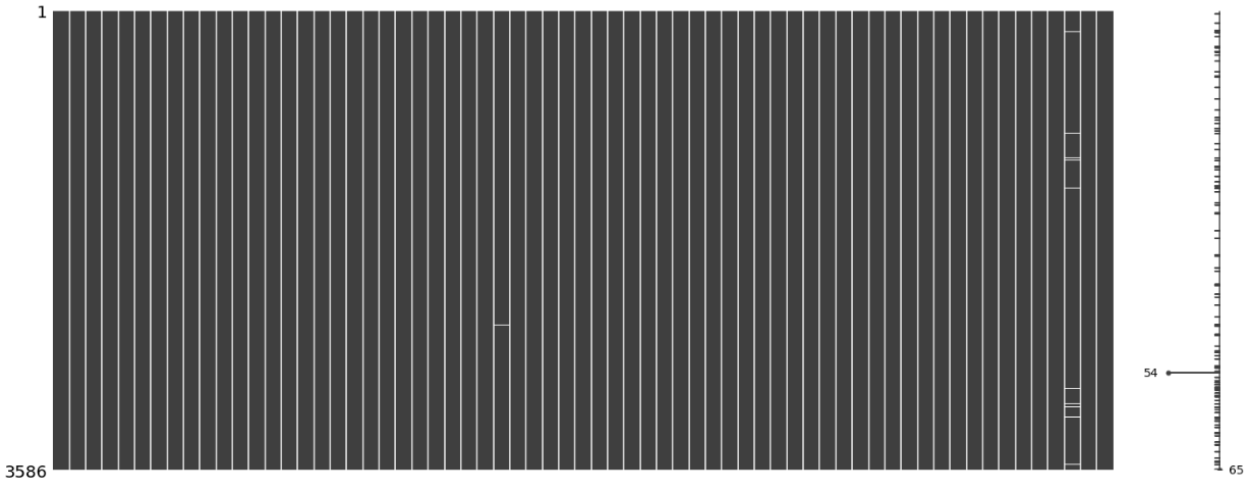
The total number of outliers present in the data are 42031. I have treated the outliers and removed them from the data.



The outliers are removed from the data now.

1.2 Missing Value Treatment

Let's check for the missing values in the data and treat them.



From the above snapshot from missingno library,

The size of the data is 3586 rows. There are 0.05% null values are present in the data.

The number of missing values in the data are 118.

Null values are imputed with Median.

Hence, there are no more missing values present in the data.

1.3 Transform Target variable into 0 and 1

A new dependent variable named "Default" was created based on the criteria given in the project instructions.

Criteria:

1 - If the Net Worth Next Year is negative for the company

0 - If the Net Worth Next Year is positive for the company

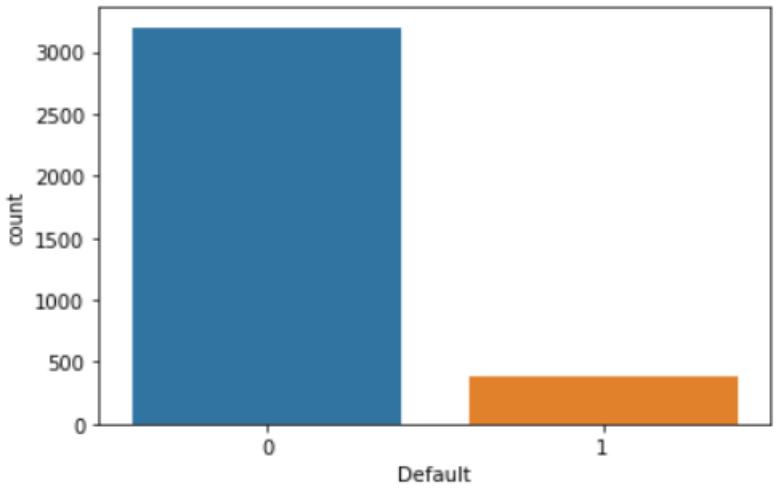
Making use of np.where function to achieve this.

```
0    3199
1     387
Name: Default, dtype: int64
```

Proportion of 'Default' Variable:

```
0    0.89208
1    0.10792
Name: Default, dtype: float64
```

After generating the dependent column, we checked for the split of data based on this dependent variable. Below is a bar plot showing the same.



1.4 Univariate (4 marks) & Bivariate

Univariate Analysis:

	count	mean	std	min	25%	50%	75%	max
Equity_Paid_Up	3586.0	13.994651	14.001442	0.00000	3.7500	8.290	19.5175	43.16875
Networth	3586.0	73.691121	112.937394	-166.21500	3.8925	18.580	117.2975	287.40500
Capital_Employed	3586.0	152.489564	207.868131	-320.90125	7.6025	39.090	226.6050	555.10875
Total_Debt	3586.0	47.439152	68.217338	-0.72000	0.0300	7.490	72.3500	180.83000
Gross_Block	3586.0	85.099863	121.230109	-41.19000	0.5700	15.870	131.8950	328.88250
Net_Working_Capital	3586.0	36.800227	59.347261	-89.40625	0.9425	10.145	61.1750	151.52375
Curr_Assets	3586.0	91.366022	122.085815	-0.91000	4.0000	24.540	135.2775	332.19375
Curr_Liab_and_Prov	3586.0	43.518483	60.878782	-0.23000	0.7325	9.225	65.6500	163.02625
Total_Assets_to_Liab	3586.0	208.542427	280.419359	-4.51000	10.5550	52.010	310.5400	760.51750
Gross_Sales	3586.0	159.511180	224.244869	-62.59000	1.4425	31.210	242.2500	603.46125
Net_Sales	3586.0	154.121439	216.700894	-62.59000	1.4400	30.440	234.4400	583.94000
Other_Income	3586.0	2.434778	3.440992	-5.40250	0.0200	0.450	3.6350	9.05750
Value_Of_Output	3586.0	154.889446	217.848047	-119.10000	1.4125	30.895	235.8375	587.47500
Cost_of_Prod	3586.0	124.595792	174.932864	-22.65000	0.9400	25.990	189.5500	472.46500
Selling_Cost	3586.0	2.478154	3.720671	0.00000	0.0000	0.160	3.8825	9.70625
PBIDT	3586.0	14.161569	23.752052	-35.18750	0.0400	2.045	23.5250	58.75250
PBDT	3586.0	7.013376	14.217803	-19.41750	0.0000	0.795	12.9450	32.36250
PBIT	3586.0	9.393051	17.456131	-25.00125	0.0000	1.150	16.6675	41.66875
PBT	3586.0	3.494844	8.714200	-11.28375	-0.0600	0.310	7.4225	18.64625
PAT	3586.0	2.583009	6.563022	-8.46000	-0.0600	0.255	5.5400	13.94000
Adjusted_PAT	3586.0	2.407965	6.361349	-8.23875	-0.0900	0.210	5.3425	13.49125
CP	3586.0	5.935719	11.963796	-16.36500	0.0000	0.740	10.9100	27.27500
Rev_earn_in_forex	3586.0	4.492794	7.380730	0.00000	0.0000	0.000	7.2000	18.00000
Rev_exp_in_forex	3586.0	4.370090	7.024458	0.00000	0.0000	0.000	6.9875	17.46875
Capital_exp_in_forex	3586.0	0.000000	0.000000	0.00000	0.0000	0.000	0.0000	0.00000
Book_Value_Unit_Curr	3586.0	45.460531	60.086473	-87.59500	7.9625	21.665	71.6675	167.22500
Book_Value_Adj_Unit_Curr	3586.0	38.138709	50.129294	-72.36500	7.0650	18.925	59.9600	139.43500
Market_Capitalisation	3586.0	72.370378	107.359374	0.00000	0.0000	8.370	111.4575	278.64375
CEPS_annualised_Unit_Curr	3586.0	4.816161	8.991216	-13.15875	0.0000	1.145	8.7725	21.93125
Cash_Flow_From_Opr	3586.0	6.411553	14.372588	-19.74000	-0.3075	0.450	12.6475	32.08000
Cash_Flow_From_Inv	3586.0	-2.453100	5.998657	-12.97375	-5.1175	-0.120	0.1200	7.97625
Cash_Flow_From_Fin	3586.0	-2.237539	7.585471	-15.30500	-5.8475	0.000	0.4575	9.91500
ROG_Net_Worth_perc	3586.0	4.123604	14.300085	-20.76250	-1.4875	1.840	11.3625	30.63750
ROG_Capital_Employed_perc	3586.0	4.351926	16.916967	-28.46875	-3.8350	1.375	12.5875	37.22125

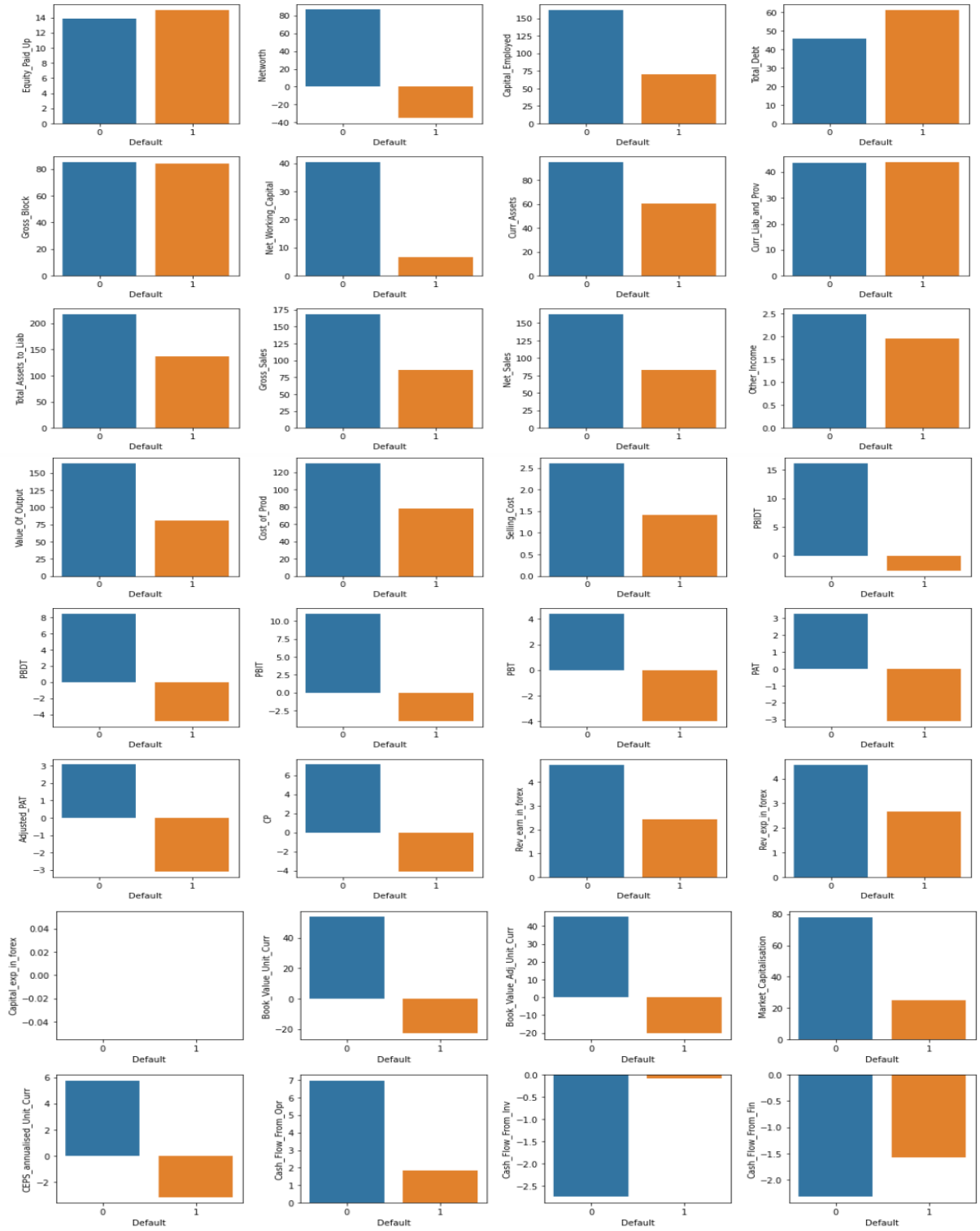
ROG_Gross_Block_perc	3586.0	2.946049	7.652767	-10.08000	0.0000	0.250	6.7200	16.80000
ROG_Gross_Sales_perc	3586.0	6.635307	32.007348	-52.48125	-8.0775	3.310	21.5250	65.92875
ROG_Net_Sales_perc	3586.0	6.634232	32.056429	-52.64500	-8.1175	3.205	21.5675	66.09500
ROG_Cost_of_Prod_perc	3586.0	7.889492	33.097493	-52.79000	-7.2425	4.415	23.1225	68.67000
ROG_Total_Assets_perc	3586.0	4.274319	16.373758	-28.68125	-3.9725	1.475	12.5000	37.20875
ROG_PBIDT_perc	3586.0	12.605595	75.031601	-130.21875	-23.3625	4.570	47.8750	154.73125
ROG_PBDT_perc	3586.0	11.284433	87.304064	-155.86625	-30.5975	3.365	52.9150	178.18375
ROG_PBIT_perc	3586.0	9.955165	85.416812	-153.59500	-31.3525	2.130	50.1425	172.38500
ROG_PBT_perc	3586.0	8.286174	106.668670	-196.02375	-41.2350	0.025	61.9575	216.74625
ROG_PAT_perc	3586.0	9.353691	111.888935	-207.35250	-43.7325	0.000	65.3475	228.96750
ROG_CP_perc	3586.0	12.059554	87.096692	-153.12375	-29.5050	4.615	52.9075	176.52625
ROG_Rev_earn_in_forex_perc	3586.0	0.000000	0.000000	0.00000	0.0000	0.000	0.0000	0.00000
ROG_Rev_exp_in_forex_perc	3586.0	0.000000	0.000000	0.00000	0.0000	0.000	0.0000	0.00000
ROG_Market_Capitalisation_perc	3586.0	23.411672	48.960825	-71.27250	0.0000	0.000	47.5150	118.78750
Curr_Ratio_Latest	3586.0	2.084084	1.806351	0.00000	0.8800	1.360	2.7700	5.60500
Fixed_Assets_Ratio_Latest	3586.0	3.310121	3.970312	0.00000	0.2700	1.560	4.7400	11.44500
Inventory_Ratio_Latest	3586.0	6.071224	7.311440	0.00000	0.0000	3.560	8.9375	22.35000
Debtors_Ratio_Latest	3586.0	5.990296	6.626035	0.00000	0.4200	3.820	8.5175	20.67000
Total_Asset_Turnover_Ratio_Latest	3586.0	0.990337	1.110387	0.00000	0.0700	0.600	1.5500	3.77000
Interest_Cover_Ratio_Latest	3586.0	2.078465	3.912223	-5.56500	0.0000	1.080	3.7100	9.27500
PBIDTM_perc_Latest	3586.0	10.548339	18.935335	-28.48500	0.0000	8.070	18.9875	47.47500
PBITM_perc_Latest	3586.0	7.213992	15.281989	-21.43500	0.0000	5.230	14.2850	35.72500
PBDTM_perc_Latest	3586.0	6.667837	15.111930	-21.16500	0.0000	4.690	14.1000	35.27500
CPM_perc_Latest	3586.0	5.320970	12.291001	-17.08500	0.0000	3.890	11.3875	28.47500
APATM_perc_Latest	3586.0	2.821170	8.728962	-11.11500	0.0000	1.590	7.4075	18.52500
Debtors_Vel_Days	3586.0	75.286670	81.861954	0.00000	8.0000	49.000	106.0000	253.00000
Creditors_Vel_Days	3586.0	62.440742	68.144543	0.00000	8.0000	39.000	89.0000	210.50000
Inventory_Vel_Days	3586.0	61.223926	73.203385	-144.00000	0.0000	35.000	93.0000	240.00000
Value_of_Output_to_Total_Assets	3586.0	0.730884	0.774996	-0.33000	0.0700	0.480	1.1600	2.79500
Value_of_Output_to_Gross_Block	3586.0	3.356559	4.101482	-6.69000	0.2700	1.530	4.9100	11.87000

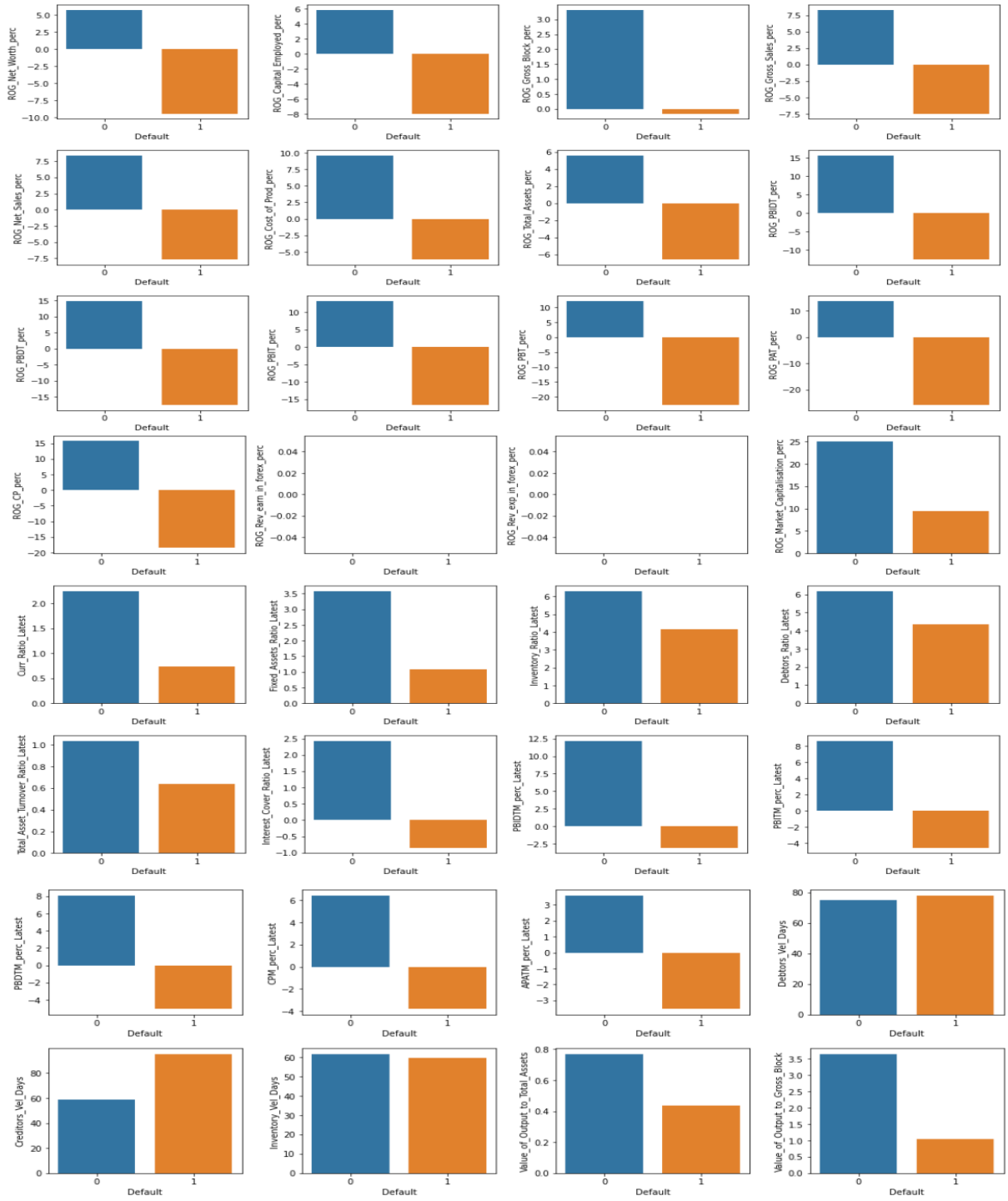
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3586 entries, 0 to 3585
Data columns (total 65 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Equity_Paid_Up                  3586 non-null   float64
 1   Networth                        3586 non-null   float64
 2   Capital_Employed                3586 non-null   float64
 3   Total_Debt                      3586 non-null   float64
 4   Gross_Block                     3586 non-null   float64
 5   Net_Working_Capital             3586 non-null   float64
 6   Curr_Assets                     3586 non-null   float64
 7   Curr_Liab_and_Prov              3586 non-null   float64
 8   Total_Assets_to_Liab            3586 non-null   float64
 9   Gross_Sales                     3586 non-null   float64
 10  Net_Sales                       3586 non-null   float64
 11  Other_Income                    3586 non-null   float64
 12  Value_Of_Output                 3586 non-null   float64
 13  Cost_of_Prod                    3586 non-null   float64
 14  Selling_Cost                    3586 non-null   float64
 15  PBIDT                           3586 non-null   float64
 16  PBDT                            3586 non-null   float64
 17  PBIT                            3586 non-null   float64
 18  PBT                             3586 non-null   float64
 19  PAT                             3586 non-null   float64
 20  Adjusted_PAT                    3586 non-null   float64
 21  CP                              3586 non-null   float64
 22  Rev_earn_in_forex               3586 non-null   float64
 23  Rev_exp_in_forex                3586 non-null   float64
 24  Capital_exp_in_forex            3586 non-null   float64
 25  Book_Value_Unit_Curr            3586 non-null   float64
 26  Book_Value_Adj_Unit_Curr        3586 non-null   float64
 27  Market_Capitalisation           3586 non-null   float64
 28  CEPS_annualised_Unit_Curr       3586 non-null   float64
 29  Cash_Flow_From_Opr              3586 non-null   float64
 30  Cash_Flow_From_Inv              3586 non-null   float64
 31  Cash_Flow_From_Fin              3586 non-null   float64
 32  ROG_Net_Worth_perc              3586 non-null   float64
 33  ROG_Capital_Employed_perc       3586 non-null   float64
 34  ROG_Gross_Block_perc            3586 non-null   float64
 35  ROG_Gross_Sales_perc            3586 non-null   float64
 36  ROG_Net_Sales_perc              3586 non-null   float64
 37  ROG_Cost_of_Prod_perc           3586 non-null   float64
 38  ROG_Total_Assets_perc           3586 non-null   float64
 39  ROG_PBIDT_perc                  3586 non-null   float64
 40  ROG_PBDT_perc                   3586 non-null   float64
 41  ROG_PBIT_perc                   3586 non-null   float64
 42  ROG_PBT_perc                    3586 non-null   float64
 43  ROG_PAT_perc                    3586 non-null   float64
 44  ROG_CP_perc                     3586 non-null   float64
 45  ROG_Rev_earn_in_forex_perc      3586 non-null   float64
 46  ROG_Rev_exp_in_forex_perc       3586 non-null   float64
 47  ROG_Market_Capitalisation_perc  3586 non-null   float64
 48  Curr_Ratio_Latest               3586 non-null   float64
 49  Fixed_Assets_Ratio_Latest       3586 non-null   float64
 50  Inventory_Ratio_Latest          3586 non-null   float64
 51  Debtors_Ratio_Latest            3586 non-null   float64
 52  Total_Asset_Turnover_Ratio_Latest 3586 non-null float64
 53  Interest_Cover_Ratio_Latest     3586 non-null   float64
 54  PBIDTM_perc_Latest              3586 non-null   float64
 55  PBITM_perc_Latest               3586 non-null   float64
 56  PBDTM_perc_Latest               3586 non-null   float64
 57  CPM_perc_Latest                 3586 non-null   float64
 58  APATM_perc_Latest               3586 non-null   float64
 59  Debtors_Vel_Days                3586 non-null   float64
 60  Creditors_Vel_Days              3586 non-null   float64
 61  Inventory_Vel_Days              3586 non-null   float64
 62  Value_of_Output_to_Total_Assets 3586 non-null   float64
 63  Value_of_Output_to_Gross_Block  3586 non-null   float64
 64  Default                         3586 non-null   object
```

Shape of the data (3586, 65).
Maximum Total_debt = 180.83000000000004

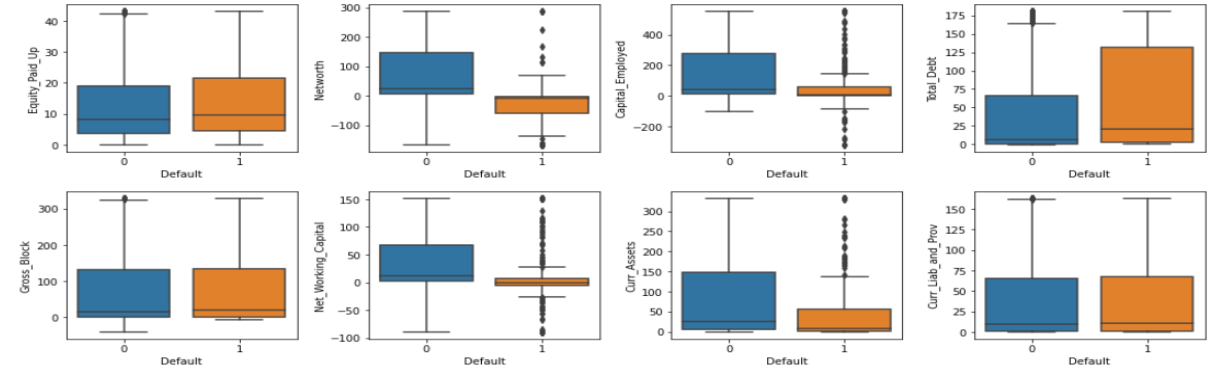
Bivariate Analysis:

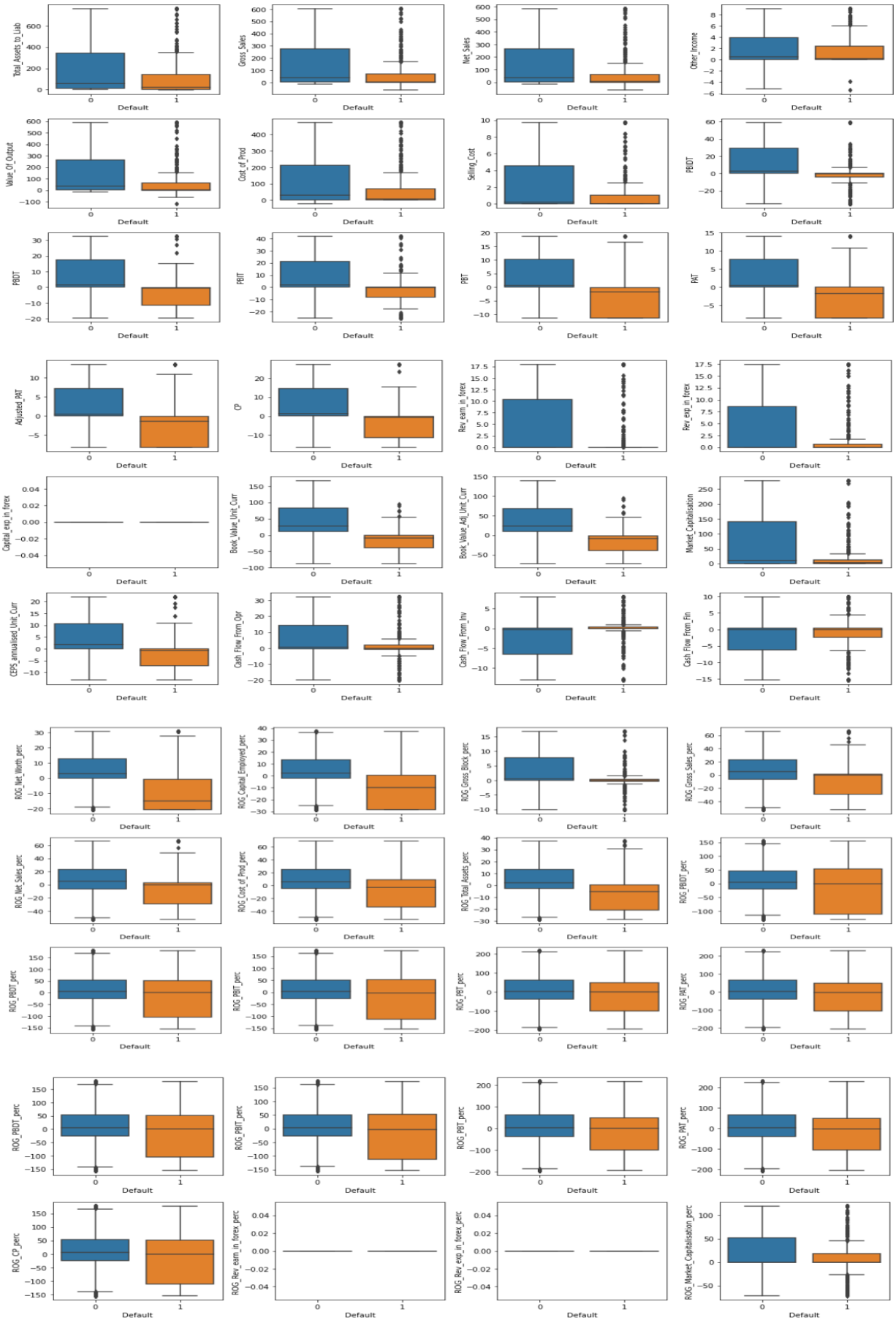
Bar plots:

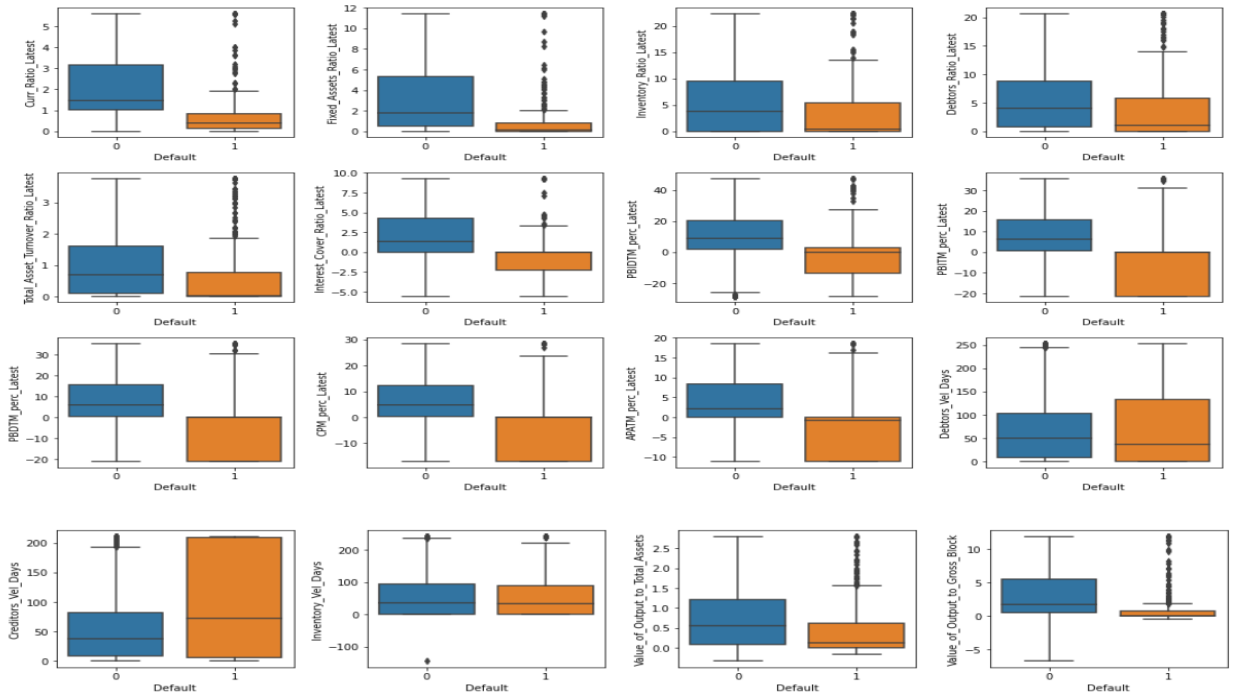



ROG_Net_Worth_perc
ROG_Capital_Employed_perc
ROG_Gross_Block_perc
ROG_Gross_Sales_perc
ROG_Net_Sales_perc
ROG_Cost_of_Prod_perc
ROG_Total_Assets_perc
ROG_PBIDT_perc
ROG_PBDT_perc
ROG_PBIT_perc
ROG_PBT_perc
ROG_PAT_perc
ROG_CP_perc
ROG_Rev_earn_in_forex_perc
ROG_Rev_exp_in_forex_perc
ROG_Market_Capitalisation_perc
Curr_Ratio_Latest
Fixed_Assets_Ratio_Latest
Inventory_Ratio_Latest
Debtors_Ratio_Latest
Total_Asset_Turnover_Ratio_Latest
Interest_Cover_Ratio_Latest
PBIDTM_perc_Latest
PBITM_perc_Latest
PBDTM_perc_Latest
CPM_perc_Latest
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

Box Plots:

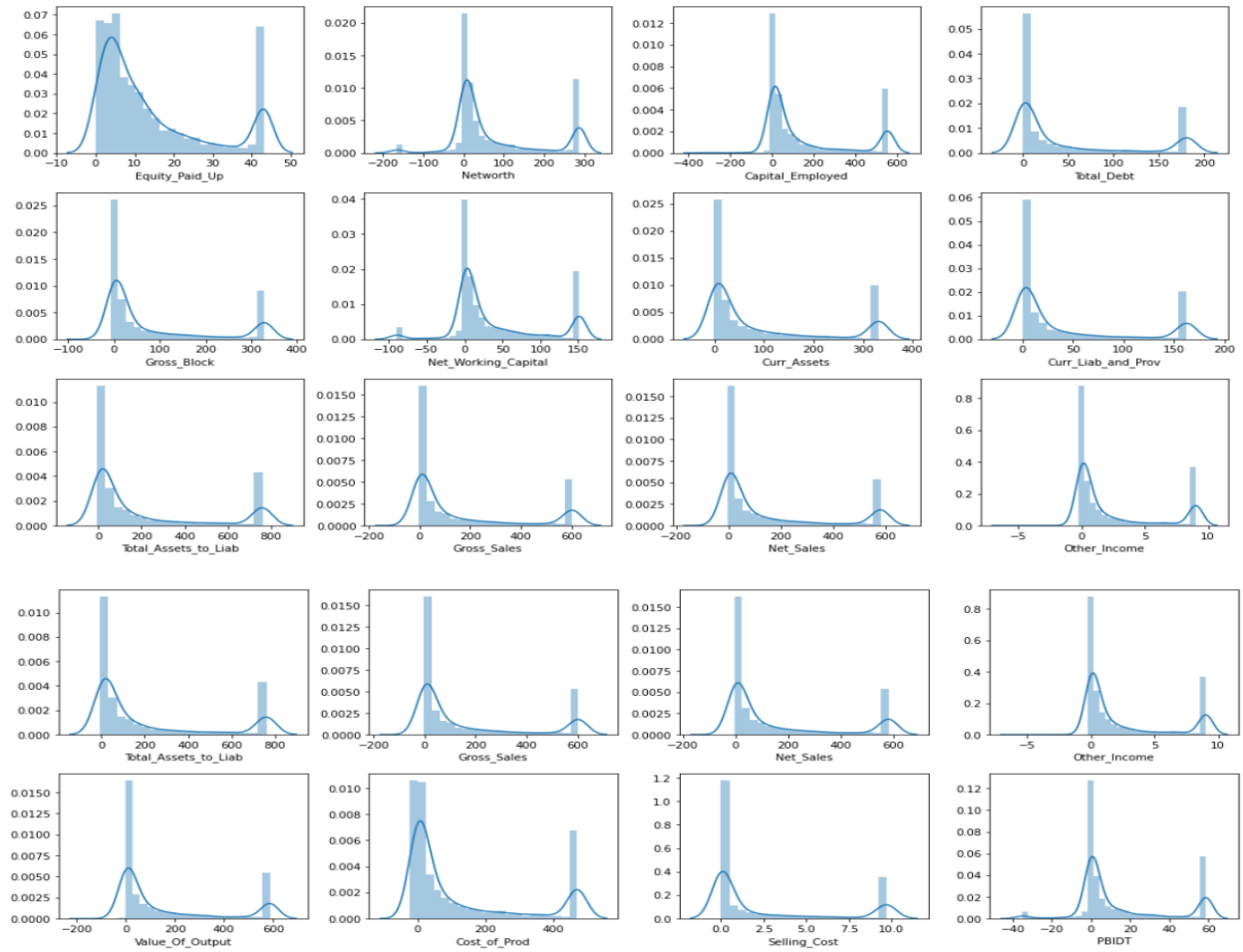

Equity_Paid_Up
Networth
Capital_Employed
Total_Debt
Gross_Block
Net_Working_Capital
Curr_Assets
Curr_Liab_and_Prov
Default

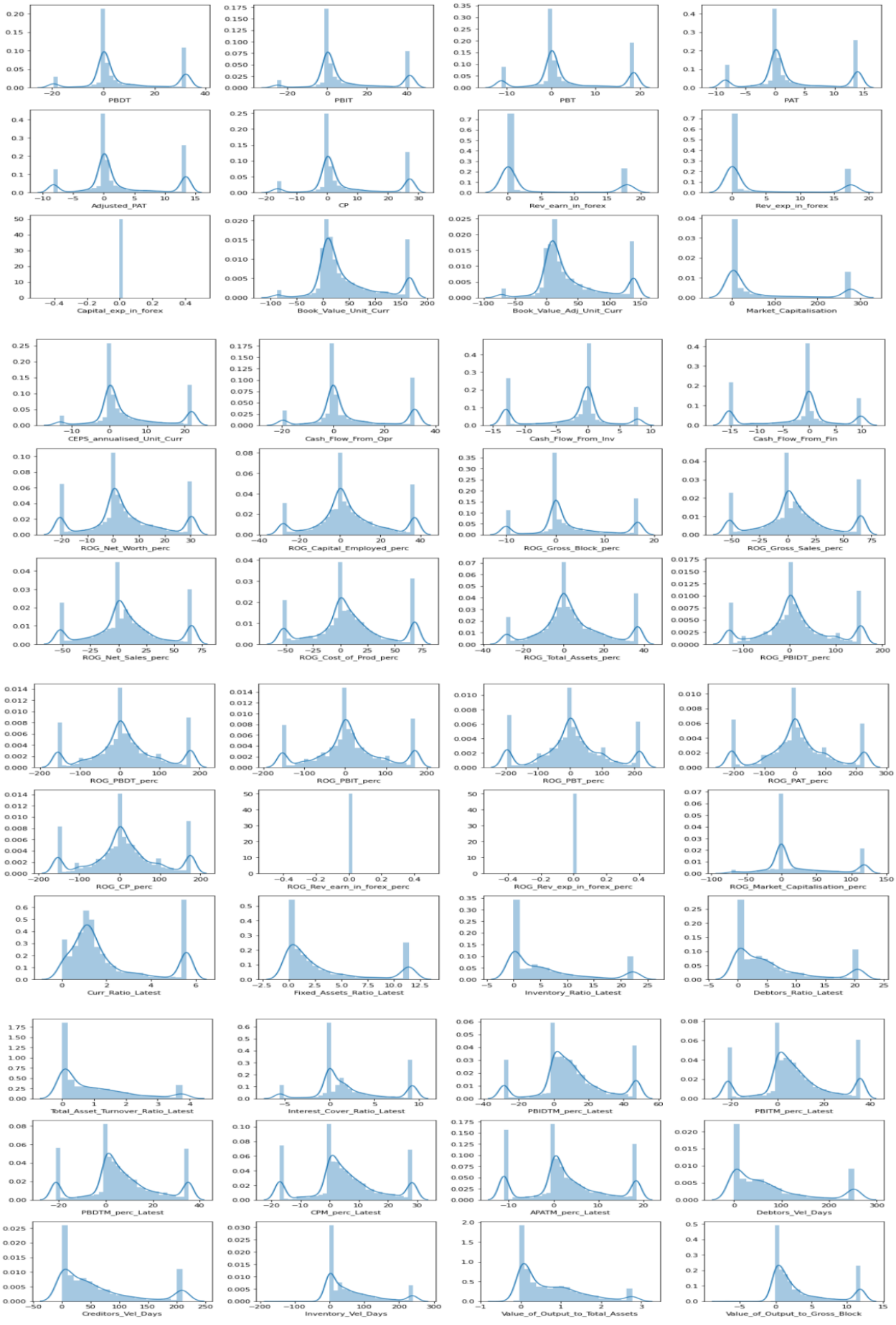

Total_Assets_to_Liab
Gross_Sales
Net_Sales
Other_Income
Value_Of_Output
Cost_of_Prod
Selling_Cost
PBDT
PBDT
PBIT
PBT
PAT
Adjusted_PAT
CP
Rev_earn_in_forex
Rev_exp_in_forex
Capital_exp_in_forex
Book_Value_Unit_Curr
Book_Value_Adj_Unit_Curr
Market_Capitalisation
CEPS_annualised_Unit_Curr
Cash_Flow_From_Opr
Cash_Flow_From_Inv
Cash_Flow_From_Fin
ROC_Net_Worth_perc
ROC_Capital_Employed_perc
ROC_Gross_Block_perc
ROC_Gross_Sales_perc
ROC_Net_Sales_perc
ROC_Cost_of_Prod_perc
ROC_Total_Assets_perc
ROC_PBDT_perc
ROC_PBDT_perc
ROC_PBIT_perc
ROC_PBT_perc
ROC_PAT_perc
ROC_PBDT_perc
ROC_PBIT_perc
ROC_PBT_perc
ROC_PAT_perc
ROC_CP_perc
ROC_Rev_earn_in_forex_perc
ROC_Rev_exp_in_forex_perc
ROC_Market_Capitalisation_perc
Default



We can see some outliers are present in the default feature.

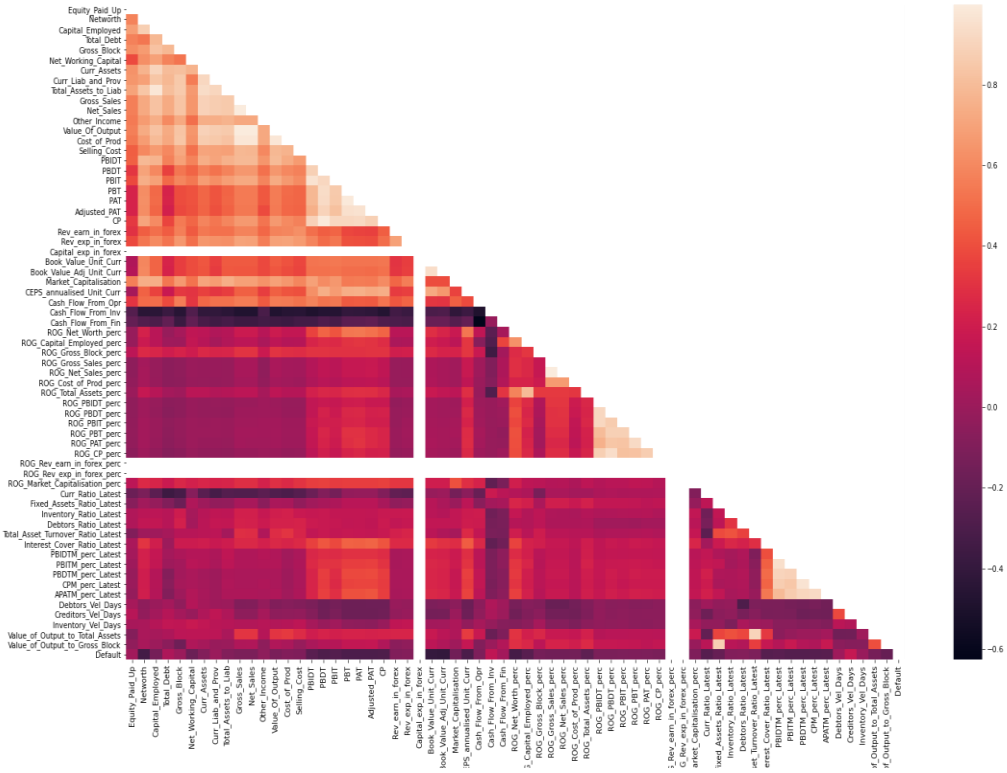
Histogram:




PBDT
PBIT
PBT
PAT
Adjusted_PAT
CP
Rev_earn_in_forex
Rev_exp_in_forex
Capital_exp_in_forex
Book_Value_Unit_Curr
Book_Value_Adj_Unit_Curr
Market_Capitalisation
CEPS_annualised_Unit_Curr
Cash_Flow_From_Opr
Cash_Flow_From_Inv
Cash_Flow_From_Fin
ROG_Net_Worth_perc
ROG_Capital_Employed_perc
ROG_Gross_Block_perc
ROG_Gross_Sales_perc
ROG_Net_Sales_perc
ROG_Cost_of_Prod_perc
ROG_Total_Assets_perc
ROG_PBIDT_perc
ROG_PBDT_perc
ROG_PBIT_perc
ROG_PBT_perc
ROG_PAT_perc
ROG_CP_perc
ROG_Rev_earn_in_forex_perc
ROG_Rev_exp_in_forex_perc
ROG_Market_Capitalisation_perc
Curr_Ratio_Latest
Fixed_Assets_Ratio_Latest
Inventory_Ratio_Latest
Debtors_Ratio_Latest
Total_Asset_Turnover_Ratio_Latest
Interest_Cover_Ratio_Latest
PBIDTM_perc_Latest
PBITM_perc_Latest
PBDTM_perc_Latest
CPM_perc_Latest
APATM_perc_Latest
Debtors_Vel_Days
Creditors_Vel_Days
Inventory_Vel_Days
Value_of_Output_to_Total_Assets
Value_of_Output_to_Gross_Block

None of the variables show perfect normal distribution. Few of the variables have skewness in data. There are no duplicate values in the data.Skewness was observed in almost all the variables. Most of the variables were right skewed while a few were also found to be left skewed.

Multivariate Analysis:



We also performed multi variate analysis on the data to see if there are any correlation that are observed within the data. Correlation function was used and seaborn heatmap was used to plot the correlations and to make better sense of the data. We observed that networth and networth next year were highly correlated. Apart from this, we also found various Rate of Growth variables were highly correlated. This analysis tells us that there is a problem of collinearity with this data set.

Encoding the data before train test split:

	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block	Net_Working_Capital	Curr_Assets	Curr_Liab_and_Prov	Total_Assets_to_Liab
0	43.16875	-166.215	-320.90125	180.83	328.8825	-89.40625	40.50000	163.02625	109.6000
1	43.16875	-166.215	555.10875	180.83	328.8825	-89.40625	332.19375	163.02625	760.5175
2	43.16875	287.405	555.10875	180.83	328.8825	151.52375	332.19375	163.02625	760.5175
3	43.16875	-166.215	555.10875	180.83	328.8825	-89.40625	332.19375	163.02625	760.5175
4	43.16875	-166.215	555.10875	180.83	328.8825	151.52375	332.19375	163.02625	760.5175
...	...	...	...	...	...	...	...	...	...
3581	43.16875	287.405	555.10875	180.83	328.8825	0.00000	332.19375	163.02625	760.5175
3582	43.16875	287.405	555.10875	180.83	328.8825	151.52375	332.19375	163.02625	760.5175
3583	43.16875	287.405	555.10875	180.83	328.8825	151.52375	332.19375	163.02625	760.5175
3584	43.16875	287.405	555.10875	180.83	328.8825	151.52375	332.19375	163.02625	760.5175
3585	43.16875	287.405	555.10875	180.83	328.8825	-89.40625	332.19375	163.02625	760.5175

3586 rows × 65 columns

1.5 Train Test Split

Train Test Split - We split the data into Train and Test dataset in a ratio of 67:33 and used random_state =42. Model Building is done on Train Dataset and Model Validation is done on Test Dataset.

First five rows of Train data:

	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block	Net_Working_Capital	Curr_Assets	Curr_Liab_and_Prov	Total_Assets_to_Liab
662	3.00	2.680	2.72000	0.00	1.0100	1.15000	1.17000	0.01000	2.7300
1373	2.09	9.740	9.83000	0.00	1.0100	2.35000	2.71000	0.36000	10.1900
3268	25.70	287.405	555.10875	180.83	328.8825	151.52375	332.19375	163.02625	760.5175
3246	35.03	287.405	555.10875	180.83	328.8825	151.52375	332.19375	163.02625	760.5175
1456	2.50	8.270	15.14000	6.87	7.5400	8.80000	9.67000	0.88000	16.0200

5 rows × 64 columns

First five rows of Test data:

	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block	Net_Working_Capital	Curr_Assets	Curr_Liab_and_Prov	Total_Assets_to_Liab	Gro
3163	16.79000	287.405	555.10875	180.83	328.8825	151.52375	332.19375	163.02625	760.5175	6
3133	43.16875	287.405	555.10875	180.83	328.8825	10.08000	157.02000	146.94000	760.5175	6
937	5.19000	4.390	4.39000	0.00	0.0400	4.39000	4.43000	0.04000	4.4300	
196	3.81000	-10.210	17.17000	15.16	25.8400	-4.07000	7.94000	12.00000	29.1700	
2852	25.88000	194.120	271.36000	53.59	283.7700	85.21000	122.50000	37.29000	308.6500	2

5 rows × 64 columns

Feature Selection: From sklearn.feature_selection RFE , I have performed the feature selection and selected the important features based on the result.

```
['Equity_Paid_Up',
 'APATM_perc_Latest',
 'PBDTM_perc_Latest',
 'PBITM_perc_Latest',
 'Interest_Cover_Ratio_Latest',
 'Curr_Ratio_Latest',
 'Cash_Flow_From_Opr',
 'CEPS_annualised_Unit_Curr',
 'Book_Value_Adj_Unit_Curr',
 'Book_Value_Unit_Curr',
 'Adjusted_PAT',
 'PAT',
 'PBT',
 'PBIT',
 'PBDT',
 'Value_of_Output_to_Gross_Block',
 'Curr_Assets',
 'Total_Debt',
 'Gross_Block',
 'Networth',
 'PBIDT',
 'Capital_Employed']
```

1.6 Build Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff.

Logistic regression is performed on the selected features and reduced the features based on the test results.

Formula='Default~Value_of_Output_to_Gross_Block+CEPS_annualised_Unit_Curr +Adjusted_PAT+Value_of_Output_to_Total_Assets+PBT+PBIT+PBIDT+Curr_Ratio_Latest+Inventory_Ratio_Latest+Value_Of_Output+Book_Value_Adj_Unit_Curr+Other_Income+Interest_Cover_Ratio_Latest+Total_Assets_to_Liab+Curr_Liab_and_Prov+PBITM_perc_Latest+PBDTM_perc_Latest+APATM_perc_Latest+Total_Debt+Capital_Employed+Networth+Book_Value_Unit_Curr'

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	2402
Model:	Logit	**Df Residuals:**	2379
Method:	MLE	**Df Model:**	22
Date:	Mon, 09 May 2022	**Pseudo R-squ.:**	0.6465
Time:	21:25:25	**Log-Likelihood:**	-279.72
converged:	True	**LL-Null:**	-791.34
Covariance Type:	nonrobust	**LLR p-value:**	2.214e-202

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.3258	0.577	-12.692	0.000	-8.457	-6.194
Value_of_Output_to_Gross_Block	-0.5071	0.194	-2.611	0.009	-0.888	-0.126
CEPS_annualised_Unit_Curr	-0.7253	0.385	-1.883	0.060	-1.480	0.030
Adjusted_PAT	0.4207	0.370	1.136	0.256	-0.305	1.147
Value_of_Output_to_Total_Assets	0.2254	0.170	1.330	0.184	-0.107	0.558
PBT	-0.5358	0.462	-1.160	0.246	-1.441	0.369
PBIT	0.7008	0.723	0.970	0.332	-0.716	2.117
PBIDT	-0.8265	0.664	-1.244	0.213	-2.128	0.475
Curr_Ratio_Latest	-0.9548	0.167	-5.707	0.000	-1.283	-0.627
Inventory_Ratio_Latest	-0.3574	0.160	-2.232	0.026	-0.671	-0.044
Value_Of_Output	-0.0386	0.412	-0.094	0.925	-0.845	0.768
Book_Value_Adj_Unit_Curr	-1.3083	2.411	-0.543	0.587	-6.034	3.418
Other_Income	0.1097	0.216	0.509	0.611	-0.313	0.532
Interest_Cover_Ratio_Latest	-0.2649	0.176	-1.504	0.133	-0.610	0.080
Total_Assets_to_Liab	-0.4968	1.895	-0.262	0.793	-4.211	3.217
Curr_Liab_and_Prov	0.4293	0.588	0.730	0.466	-0.724	1.582
PBITM_perc_Latest	-0.7235	0.344	-2.106	0.035	-1.397	-0.050
PBDTM_perc_Latest	-0.5772	0.424	-1.362	0.173	-1.408	0.253
APATM_perc_Latest	0.9029	0.370	2.442	0.015	0.178	1.628
Total_Debt	0.9651	0.364	2.648	0.008	0.251	1.679
Capital_Employed	-0.1023	1.599	-0.064	0.949	-3.236	3.031
Networth	-0.6646	0.397	-1.674	0.094	-1.443	0.114
Book_Value_Unit_Curr	-4.5085	2.835	-1.590	0.112	-10.065	1.048

Possibly complete quasi-separation: A fraction 0.31 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

Based on the summary results, I have reduced the features step by step and selected the important features

After that, I have performed VIF and selected the features to perform optimum model building.

	variable	vif_score
6	Total_Assets_to_Liab	5.662524
8	Total_Debt	4.562947
5	Book_Value_Adj_Unit_Curr	3.087592
2	Value_of_Output_to_Total_Assets	2.833183
1	CEPS_annualised_Unit_Curr	2.798254
0	Value_of_Output_to_Gross_Block	2.106397
4	Inventory_Ratio_Latest	1.984655
3	Curr_Ratio_Latest	1.740738
7	PBDTM_perc_Latest	1.493384

Dropped Total_Assests_to_Liab, because vif score is: 5.662524

	variable	vif_score
0	Value_of_Output_to_Gross_Block	2.095976
1	CEPS_annualised_Unit_Curr	2.776581
2	Value_of_Output_to_Total_Assets	2.794841
3	Curr_Ratio_Latest	1.740130
4	Inventory_Ratio_Latest	1.970294
5	Book_Value_Adj_Unit_Curr	2.844684
6	PBDTM_perc_Latest	1.479251
7	Total_Debt	1.475968

Final model is performed with the 7 features selected.

Formula='Default~Value_of_Output_to_Gross_Block+CEPS_annualised_Unit_Curr+Curr_Ratio_Latest+Inventory_Ratio_Latest+Book_Value_Adj_Unit_Curr+PBDTM_perc_Latest+Total_Debt'

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	2402
Model:	Logit	**Df Residuals:**	2394
Method:	MLE	**Df Model:**	7
Date:	Mon, 09 May 2022	**Pseudo R-squ.:**	0.6233
Time:	22:31:57	**Log-Likelihood:**	-298.08
converged:	True	**LL-Null:**	-791.34
Covariance Type:	nonrobust	**LLR p-value:**	9.816e-209

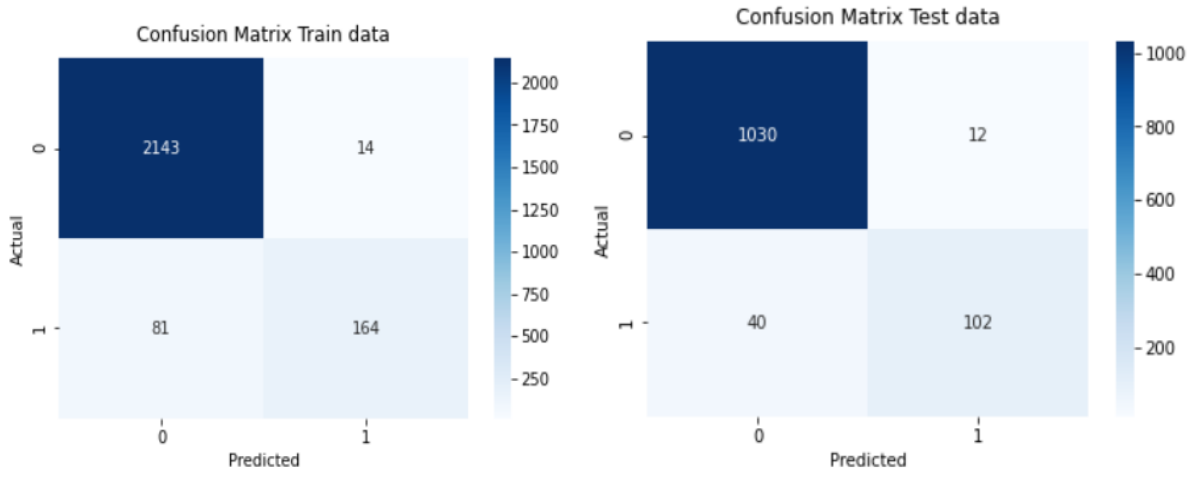
	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.0888	0.444	-15.977	0.000	-7.958	-6.219
Value_of_Output_to_Gross_Block	-0.4179	0.160	-2.616	0.009	-0.731	-0.105
CEPS_annualised_Unit_Curr	-0.9005	0.257	-3.502	0.000	-1.404	-0.396
Curr_Ratio_Latest	-0.9590	0.158	-6.075	0.000	-1.268	-0.650
Inventory_Ratio_Latest	-0.3094	0.131	-2.366	0.018	-0.566	-0.053
Book_Value_Adj_Unit_Curr	-6.0172	0.534	-11.269	0.000	-7.064	-4.971
PBDTM_perc_Latest	-0.5190	0.127	-4.087	0.000	-0.768	-0.270
Total_Debt	0.4095	0.133	3.083	0.002	0.149	0.670

Possibly complete quasi-separation: A fraction 0.29 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

P-values of all the variables are less than 0.05 and thus all the coefficients are relevant.

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

Confusion Matrix:



Classification Report for train data:

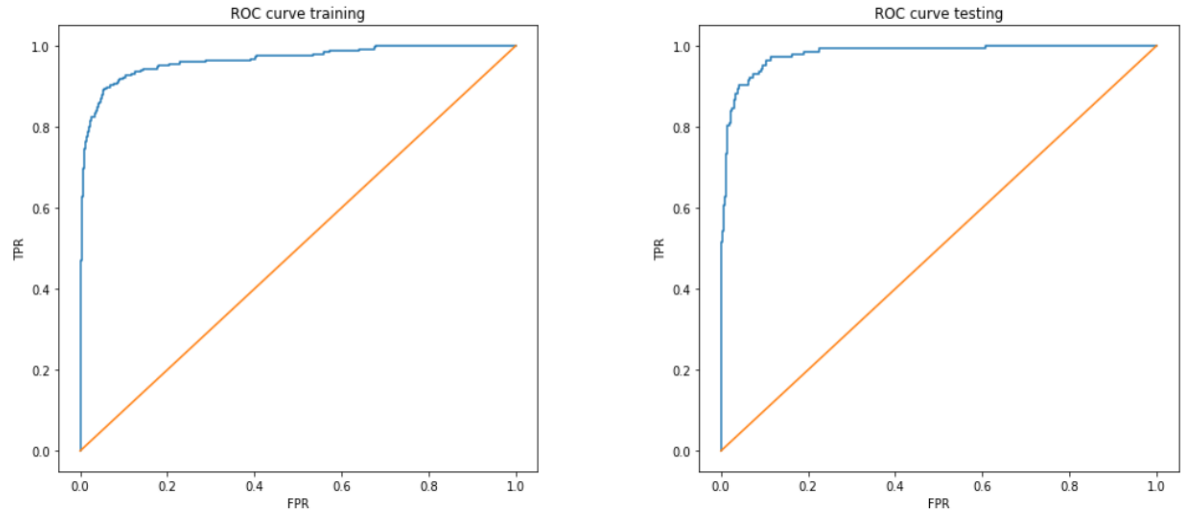
	precision	recall	f1-score	support
0	0.96	0.99	0.98	2157
1	0.92	0.67	0.78	245
accuracy			0.96	2402
macro avg	0.94	0.83	0.88	2402
weighted avg	0.96	0.96	0.96	2402

Classification Report for test data:

	precision	recall	f1-score	support
0	0.96	0.99	0.98	1042
1	0.89	0.72	0.80	142
accuracy			0.96	1184
macro avg	0.93	0.85	0.89	1184
weighted avg	0.95	0.96	0.95	1184

AUC & ROC:

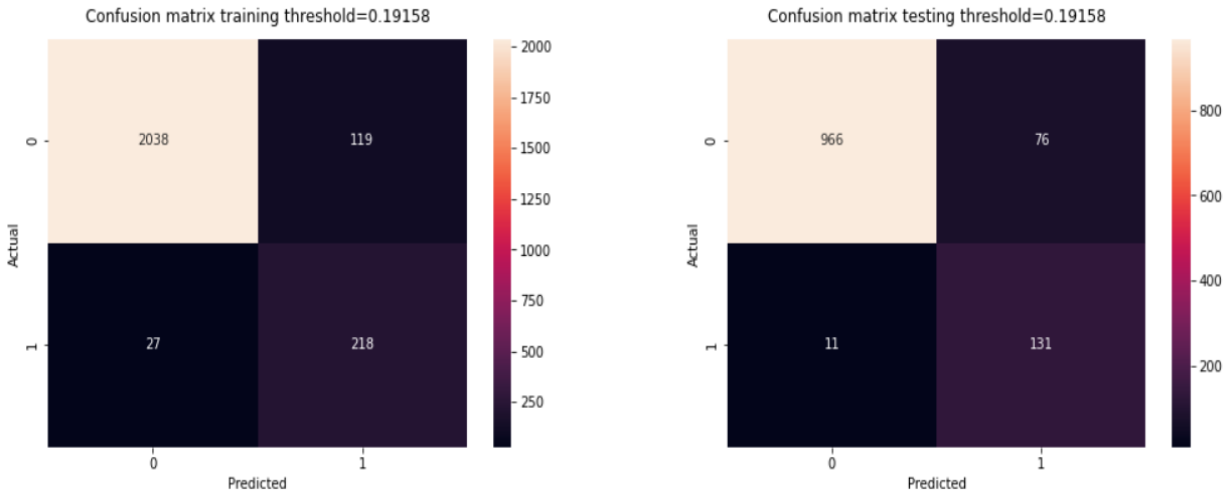
```
AUC Training:  0.9644971757826916
AUC Testing:  0.9791030250601498
```



This was pretty good result on its own, however to further improve the on the results. We decided to look for the optimum threshold.

Optimal Threshold: 0.19083113625604242

After evaluating using the optimal threshold. Below was the new classification matrix.



Classification report for train:

	precision	recall	f1-score	support
0	0.99	0.94	0.97	2157
1	0.65	0.89	0.75	245
accuracy			0.94	2402
macro avg	0.82	0.92	0.86	2402
weighted avg	0.95	0.94	0.94	2402

Accuracy of over 94% was achieved while recall, precision and f1 score were also very high at 99,94% and 97% respectively.

Classification report for test:

	precision	recall	f1-score	support
0	0.99	0.93	0.96	1042
1	0.63	0.92	0.75	142
accuracy			0.93	1184
macro avg	0.81	0.92	0.85	1184
weighted avg	0.95	0.93	0.93	1184

We also evaluated the test data set for the same model which was built after the above process.

Accuracy of 93% and very high recall, precision and f1 score of 99% ,93% and 96% respectively were also observed on the test set. This clearly indicates that the model which has been built is highly efficient and has been able to capture the correct variable for prediction.

Hence, it has been proven to work on train as well as test data.