FRA MILESTONE 1 – CREDIT RISK ANALYSIS

Akshaya Parthasarathy

PGP – DSBA Online June-E

Date: 20/06/2021

Table of Contents

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.

Executive Summary

The purpose of this project is to predict the default/non-default companies from the given balance variables. Logistic regression model has been used in order to identify the most important variables for model building and predicting the probability class of the default variable. Based on the performance metrics, identified which type of error will be risky. Improving recall of the model is chosen as the objective for improving the model.

Data Description

Given a dataset which has balance sheet details of 3586 companies with 67 attributes.

Attributes and their description:

#	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
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
10	Current Assets	All the assets of a company that are expected to be sold or used as a result of standard business
11	Current Liabilities and Provisions	Short-term financial obligations that are due within one year (includes amount that is set aside
12	Total Assets/Liabilities	Ratio of total assets to liabailities 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
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
33	CEPS (annualised) (Unit Curr)	Cash Earnings per Share, profitability ratio that measures the financial performance of a company by
34	Cash Flow From Operating	Use of cash from ongoing regular business activities
35	Cash Flow From Investing	Cash used in the purchase of non-current assets–or long-term assets– that will deliver value in the
36	Cash Flow From Financing	Net flows of cash that are used to fund the company (transactions involving debt, equity, and

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

55	Inventory Ratio[Latest]	Activity ratio, specifies the number of times the stock or inventory has been replaced and sold by
56	Debtors Ratio[Latest]	Measures how quickly cash debtors are paying back to the company
57	Total Asset Turnover	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

Sample of the dataset

	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]	APATM (%) [Latest]	Debt Velo (Da
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	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	-87.18	
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	-7961.51	
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	-51.58	
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	274.79	3

5 rows × 67 columns

Columns have been renamed for ease of use according to the data dictionary given.

Exploratory data analysis

Shape of the dataset

```
The number of rows (observations) is 3586
The number of columns (variables) is 67
```

Information of the 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   Current_Assets                        3586 non-null   float64
 10  Current_Liabilities_and_Provisions    3586 non-null   float64
 11  Total_Assets_to_Liabilities           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  Revenue_earn_in_forex                 3586 non-null   float64
 26  Revenue_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_Operating_Activities   3586 non-null   float64
 33  Cash_Flow_From_Investing_Activities   3586 non-null   float64
 34  Cash_Flow_From_Financing_Activities   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_Revenue_earn_in_forex_perc       3586 non-null   float64
 49  ROG_Revenue_exp_in_forex_perc        3586 non-null   float64
 50  ROG_Market_Capitalisation_perc       3586 non-null   float64
 51  Current_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
```

- All the variables are either float or int. Except company name which will not be useful in the model building.
- There are not a lot of missing values except for the variable Inventory_Vel_Days.

Summary statistics of the variables

	count	mean	std	min	25%	50%	75%	max
Co_Code	3586.0	16065.388734	19776.817379	4.00	3029.2500	6077.500	24269.5000	72493.00
Networth_Next_Year	3586.0	725.045251	4769.681004	-8021.60	3.9850	19.015	123.8025	111729.10
Equity_Paid_Up	3586.0	62.966584	778.761744	0.00	3.7500	8.290	19.5175	42263.46
Networth	3586.0	649.746299	4091.988792	-7027.48	3.8925	18.580	117.2975	81657.35
Capital_Employed	3586.0	2799.611054	26975.135385	-1824.75	7.6025	39.090	226.6050	714001.25
Total_Debt	3586.0	1994.823779	23652.842746	-0.72	0.0300	7.490	72.3500	652823.81
Gross_Block	3586.0	594.178829	4871.547802	-41.19	0.5700	15.870	131.8950	128477.59
Net_Working_Capital	3586.0	410.809665	6301.218546	-13162.42	0.9425	10.145	61.1750	223257.56
Current_Assets	3586.0	1960.349172	22577.570829	-0.91	4.0000	24.540	135.2775	721166.00
Current_Liabilities_and_Provisions	3586.0	391.992078	2675.001631	-0.23	0.7325	9.225	65.6500	83232.98
Total_Assets_to_Liabilities	3586.0	1778.453751	11437.574690	-4.51	10.5550	52.010	310.5400	254737.22
Gross_Sales	3586.0	1123.738985	10603.703837	-62.59	1.4425	31.210	242.2500	474182.94
Net_Sales	3586.0	1079.702579	9996.574173	-62.59	1.4400	30.440	234.4400	443775.16
Other_Income	3586.0	48.729824	426.040665	-448.72	0.0200	0.450	3.6350	14143.40
Value_Of_Output	3586.0	1077.187292	9843.880293	-119.10	1.4125	30.895	235.8375	435559.09
Cost_of_Prod	3586.0	798.544621	9076.702982	-22.65	0.9400	25.990	189.5500	419913.50
Selling_Cost	3586.0	25.554997	194.244466	0.00	0.0000	0.160	3.8825	5283.91
PBIDT	3586.0	248.175282	1949.593350	-4655.14	0.0400	2.045	23.5250	42059.26
PBDT	3586.0	116.268795	956.199566	-5874.53	0.0000	0.795	12.9450	23215.00
PBIT	3586.0	217.659395	1850.972782	-4812.95	0.0000	1.150	16.6675	41402.96
PBT	3586.0	85.752909	799.925768	-6032.34	-0.0600	0.310	7.4225	16798.00
PAT	3586.0	61.218313	620.298432	-6032.34	-0.0600	0.255	5.5400	13383.39
Adjusted_PAT	3586.0	60.058963	580.432912	-4418.72	-0.0900	0.210	5.3425	13384.11
CP	3586.0	91.734200	780.790561	-5874.53	0.0000	0.740	10.9100	20760.20

Revenue_earn_in_forex	3586.0	131.165270	1150.730209	0.00	0.0000	0.000	7.2000	46158.00
Revenue_exp_in_forex	3586.0	256.327002	4132.339619	0.00	0.0000	0.000	6.9875	193979.73
Capital_exp_in_forex	3586.0	7.655689	111.432070	0.00	0.0000	0.000	0.0000	3722.10
Book_Value_Unit_Curr	3586.0	157.237836	1622.664105	-3371.57	7.9625	21.665	71.6675	75790.00
Book_Value_Adj_Unit_Curr	3582.0	2243.152917	128283.728186	-33715.70	7.0600	18.925	60.0100	7677600.29
Market_Capitalisation	3586.0	1664.092387	12805.173084	0.00	0.0000	8.370	111.4575	260865.08
CEPS_annualised_Unit_Curr	3586.0	36.018709	828.420796	-1808.00	0.0000	1.145	8.7725	45438.44
Cash_Flow_From_Operating_Activities	3586.0	65.770750	1455.048376	-25469.23	-0.3075	0.450	12.6475	44529.40
Cash_Flow_From_Investing_Activities	3586.0	-60.870365	701.974713	-23843.45	-5.1175	-0.120	0.1200	3732.98
Cash_Flow_From_Financing_Activities	3586.0	11.436453	1272.257361	-38374.04	-5.8475	0.000	0.4575	28846.00
ROG_Net_Worth_perc	3586.0	1237.624576	41041.930017	-14485.71	-1.4875	1.840	11.3625	2144020.00
ROG_Capital_Employed_perc	3586.0	2988.884612	126472.870285	-8614.63	-3.8350	1.375	12.5875	7412700.00
ROG_Gross_Block_perc	3586.0	37.554306	893.619402	-116.12	0.0000	0.250	6.7200	47400.00
ROG_Gross_Sales_perc	3586.0	242.672962	6103.527897	-5503.70	-8.0775	3.310	21.5250	320200.00
ROG_Net_Sales_perc	3586.0	242.588530	6103.487655	-5503.70	-8.1175	3.205	21.5675	320200.00
ROG_Cost_of_Prod_perc	3586.0	310.488405	5573.215095	-2130.23	-7.2425	4.415	23.1225	267150.00
ROG_Total_Assets_perc	3586.0	2793.282621	125941.653747	-136.13	-3.9725	1.475	12.5000	7422120.00
ROG_PBIDT_perc	3586.0	375.852181	23278.396117	-52200.00	-23.3625	4.570	47.8750	1386200.00
ROG_PBDT_perc	3586.0	336.379947	20353.396660	-52200.00	-30.5975	3.365	52.9150	1208700.00
ROG_PBIT_perc	3586.0	374.699958	22462.789381	-58500.00	-31.3525	2.130	50.1425	1338000.00
ROG_PBT_perc	3586.0	224.070248	19659.232661	-78900.00	-41.2350	0.025	61.9575	1160500.00
ROG_PAT_perc	3586.0	112.231654	13480.515287	-114500.00	-43.7325	0.000	65.3475	774200.00
ROG_CP_perc	3586.0	221.091523	13980.202791	-52200.00	-29.5050	4.615	52.9075	822400.00
ROG_Revenue_earn_in_forex_perc	3586.0	37.227844	658.666041	-100.00	0.0000	0.000	0.0000	29084.77
ROG_Revenue_exp_in_forex_perc	3586.0	364.863221	15233.643027	-100.00	0.0000	0.000	0.0000	894591.69
ROG_Market_Capitalisation_perc	3586.0	63.682220	1047.928144	-98.05	0.0000	0.000	47.5150	61865.26
Current_Ratio_Latest	3585.0	12.056603	108.410131	0.00	0.8800	1.360	2.7700	4813.00
Fixed_Assets_Ratio_Latest	3585.0	51.538840	681.150910	0.00	0.2700	1.560	4.7400	22172.00
Inventory_Ratio_Latest	3585.0	37.798946	458.189394	0.00	0.0000	3.560	8.9400	15472.00
Debtors_Ratio_Latest	3585.0	33.026996	489.563498	0.00	0.4200	3.820	8.5200	22992.67
Total_Asset_Turnover_Ratio_Latest	3585.0	1.237236	2.673228	0.00	0.0700	0.600	1.5500	57.75
Interest_Cover_Ratio_Latest	3585.0	16.387894	351.737840	-5450.00	0.0000	1.080	3.7100	18639.40
PBIDTM_perc_Latest	3585.0	-51.162890	1795.131025	-78870.45	0.0000	8.070	18.9900	19233.33
PBITM_perc_Latest	3585.0	-109.213414	3057.635870	-141600.00	0.0000	5.230	14.2900	19195.70
PBDTM_perc_Latest	3585.0	-311.570357	10921.592639	-590500.00	0.0000	4.690	14.1100	15640.00
CPM_perc_Latest	3585.0	-307.005632	10676.149629	-572000.00	0.0000	3.890	11.3900	15640.00
APATM_perc_Latest	3585.0	-365.056187	12500.051387	-688600.00	0.0000	1.590	7.4100	15266.67
Debtors_Vel_Days	3586.0	603.894032	10636.759580	0.00	8.0000	49.000	106.0000	514721.00
Creditors_Vel_Days	3586.0	2057.854992	54169.479197	0.00	8.0000	39.000	89.0000	2034145.00
Inventory_Vel_Days	3483.0	79.644559	137.847792	-199.00	0.0000	35.000	96.0000	996.00
Value_of_Output_to_Total_Assets	3586.0	0.819757	1.201400	-0.33	0.0700	0.480	1.1600	17.63
Value_of_Output_to_Gross_Block	3586.0	61.884548	976.824352	-61.00	0.2700	1.530	4.9100	43404.00

- Outliers are present in all the variables. Some fields have both positive and negative outliers.
- Range of the variables are different. We have ratios, days, percentage and currency fields. Scaling does not affect the model score. Trend of the predictor and predicted variables would remain the same. Intercept and coefficients of the features will change which would remove any effects that would be present from one variable from having an incorrect magnitude of influence on our predictor variable.

- There are fields which have maximum field values as zero.
- There are no duplicate records in the dataset as well.

```
There are  0  duplicate records in the dataset
```

Creating a binary target variable for net worth next year – Target variable

Default is the newly created field based on the existing field networth_next_year. The value of default will be either 0 or 1. 0 denotes the positive values network_next_year as **not default companies** and 1 denotes the negative values of networth_next_year field as **default companies**.

	default	Networth_Next_Year
0	1	-8021.60
1	1	-3986.19
2	1	-3192.58
3	1	-3054.51
4	1	-2967.36

	default	Networth_Next_Year
3581	0	72677.77
3582	0	79162.19
3583	0	88134.31
3584	0	91293.70
3585	0	111729.10

Dropping the fields networth_next_year, co_code and co_name. Since these fields is not useful in the model building.

Value counts and proportion of the default variable:

```
Value counts:

0    3198
1     388
Name: default, dtype: int64

Proportion of the default:

0    89.18
1    10.82
Name: default, dtype: float64
```

This cannot be considered as an imbalanced dataset since we have at least 10% of records having 1 in the default variable.

There are a total of 67 variables in the dataset. Out of which 13 were identified as significant in the logistic regression model building.

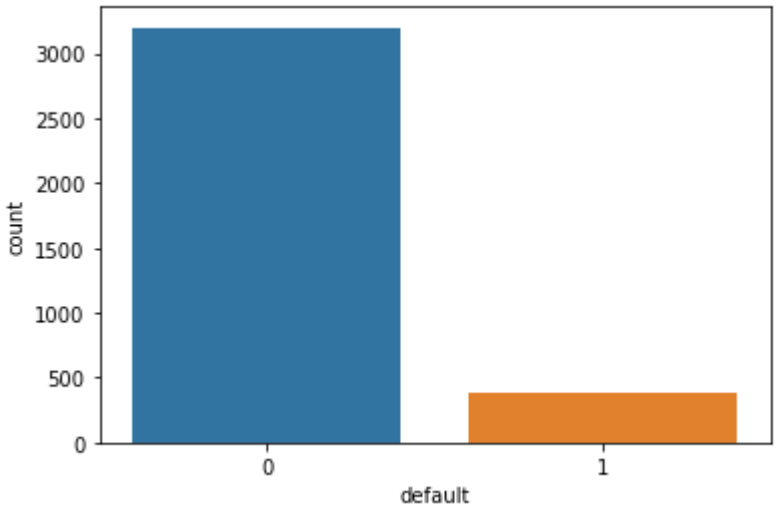
Field Name	Description
Networth	Value of a company as on 2015 - Current Year
Capital Employed	Total amount of capital used for the acquisition of profits by a company
Gross Block	Total value of all of the assets that a company owns
Total Assets/Liabilities	Ratio of total assets to liabailities of the company
Value Of Output	Product of physical output of goods and services produced by company and its market price
Cost of Production	Costs incurred by a business from manufacturing a product or providing a service
Book Value (Unit Curr)	Net asset value
Book Value (Adj.) (Unit Curr)	Book value adjusted to reflect asset's true fair market value
ROG-Net Worth (%)	Rate of Growth - Networth
ROG-Cost of Production (%)	Rate of Growth - Cost of Production
Current Ratio[Latest]	Liquidity ratio, company's ability to pay short-term obligations or those due within one year
Fixed Assets Ratio[Latest]	Solvency ratio, the capacity of a company to discharge its obligations towards long-term lenders indicating
Interest Cover Ratio[Latest]	Determines how easily a company can pay interest on its outstanding debt

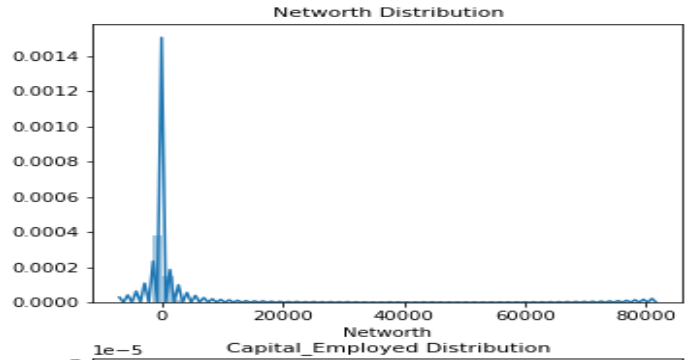
Hence including only these variables for the univariate and bivariate analysis.

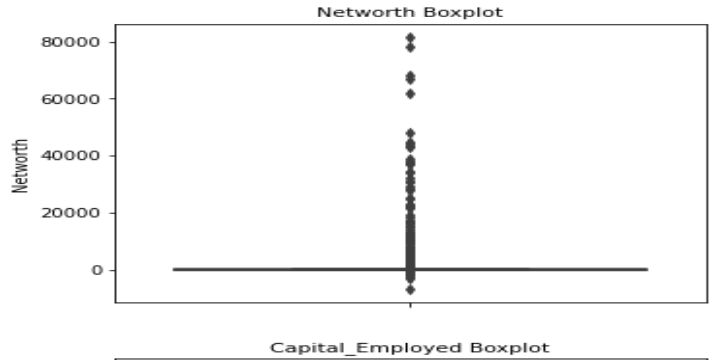
Univariate Analysis of the significant variables

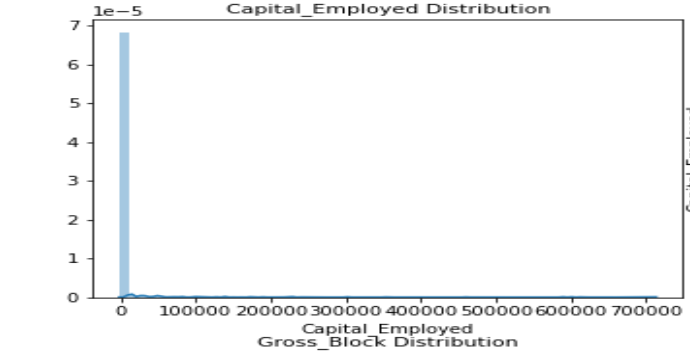
- The below plots confirm the presence of outliers in the significant variables field.
- All the variables have right skewed distribution for which the mean is greater than the median and the tail of the distribution is longer on the right-hand side than the left.

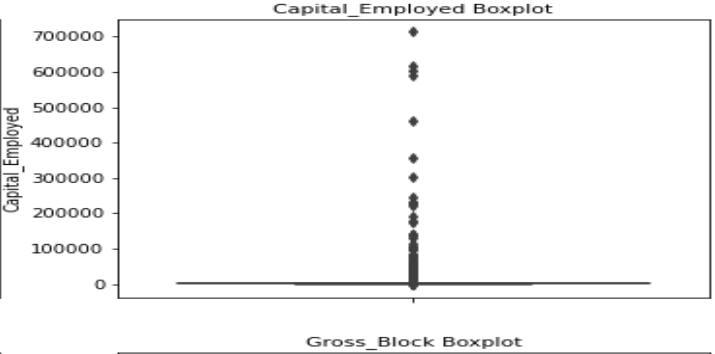
Count plot of default variable:

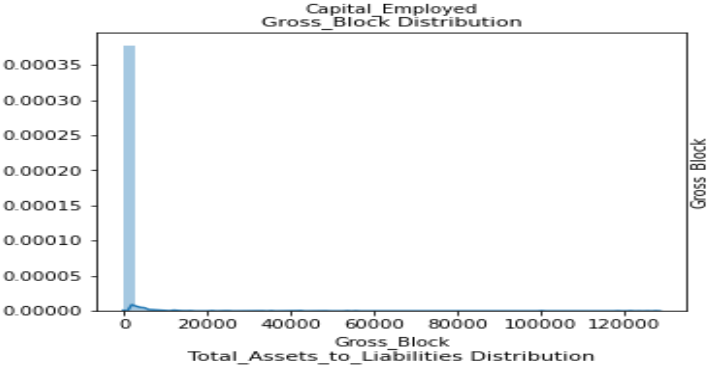


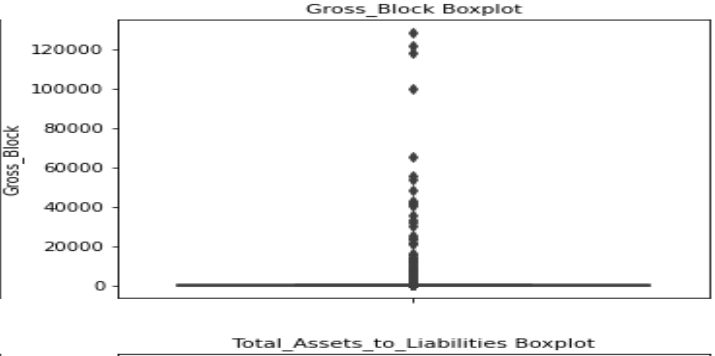

Networth Distribution
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
20000
40000
60000
80000
Networth

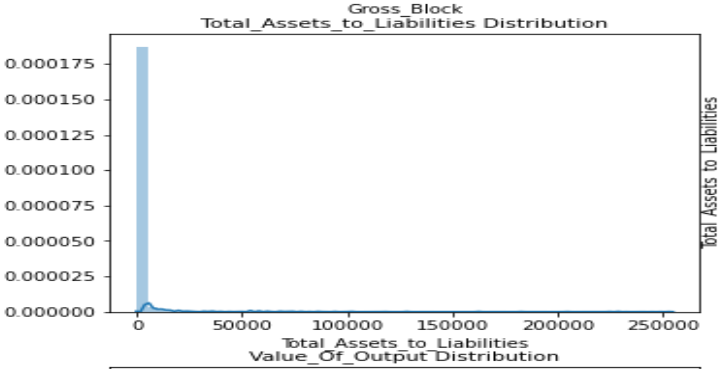

Networth Boxplot
80000
60000
40000
20000
0
Networth

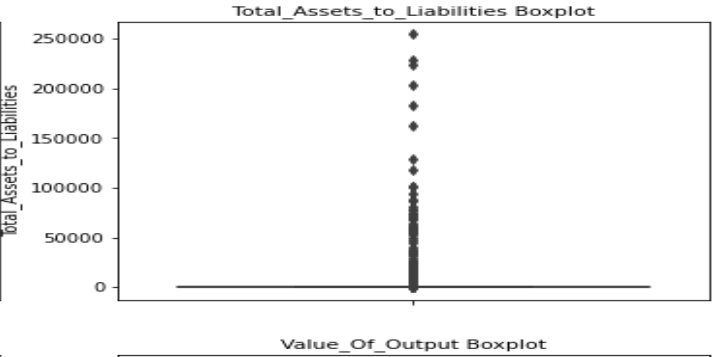

1e−5
Capital_Employed Distribution
7
6
5
4
3
2
1
0
0
100000
200000
300000
400000
500000
600000
700000
Capital_Employed

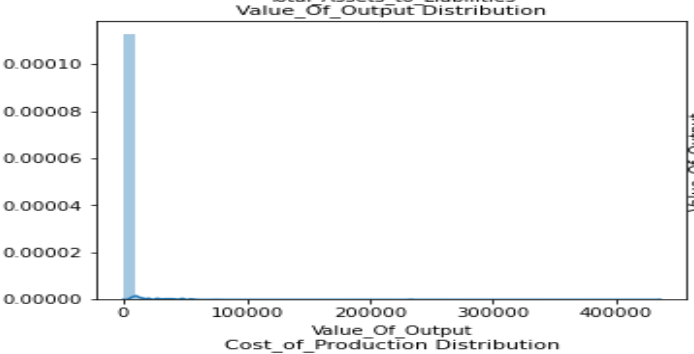

Capital_Employed Boxplot
700000
600000
500000
400000
300000
200000
100000
0
Capital_Employed

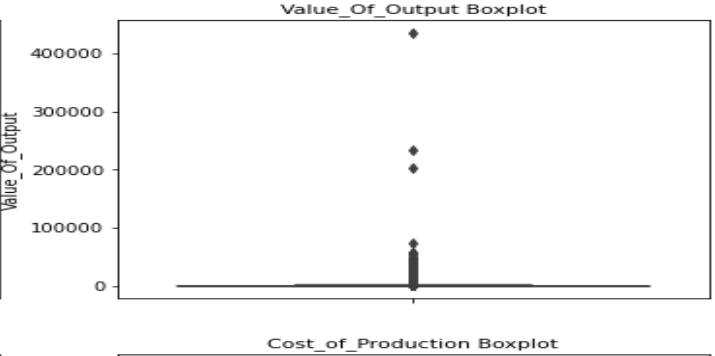

Gross_Block Distribution
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
20000
40000
60000
80000
100000
120000
Gross_Block

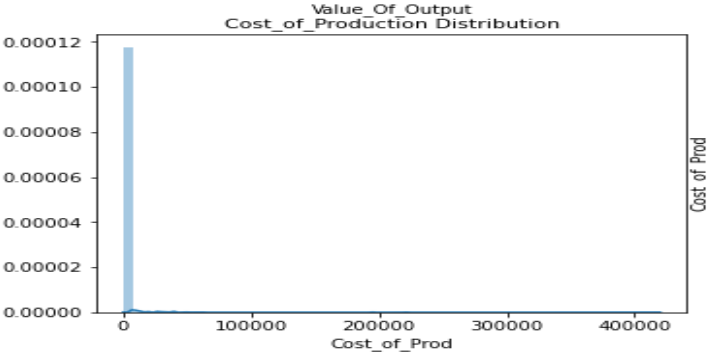

Gross_Block Boxplot
120000
100000
80000
60000
40000
20000
0
Gross_Block

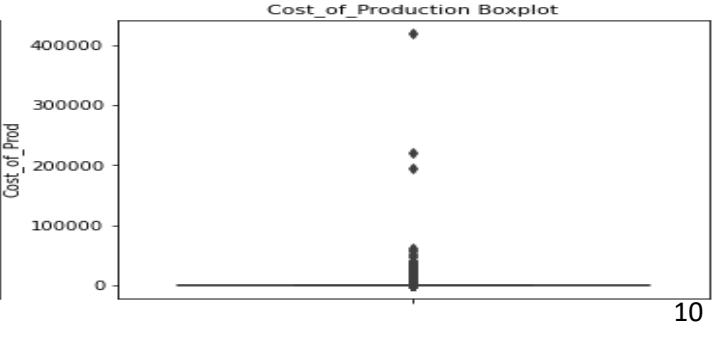

Total_Assets_to_Liabilities Distribution
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
0
50000
100000
150000
200000
250000
Total_Assets_to_Liabilities

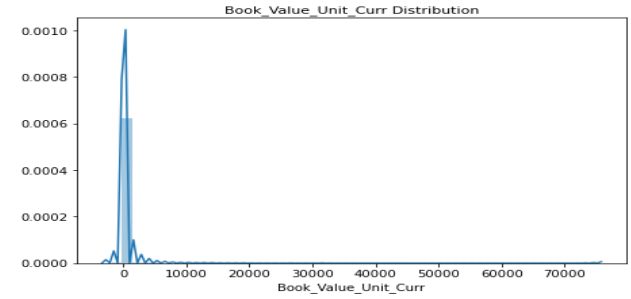

Total_Assets_to_Liabilities Boxplot
250000
200000
150000
100000
50000
0
Total_Assets_to_Liabilities

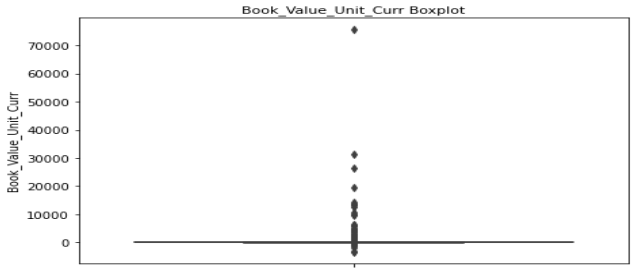

Value_Of_Output Distribution
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
100000
200000
300000
400000
Value_Of_Output

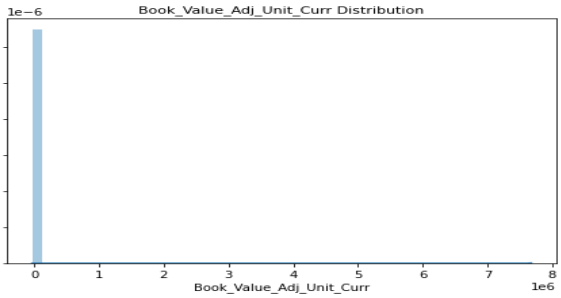

Value_Of_Output Boxplot
400000
300000
200000
100000
0
Value_Of_Output

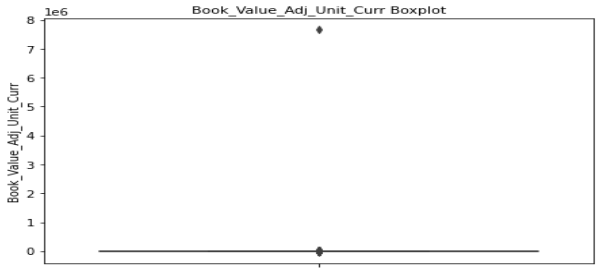

Cost_of_Production Distribution
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
100000
200000
300000
400000
Cost_of_Prod

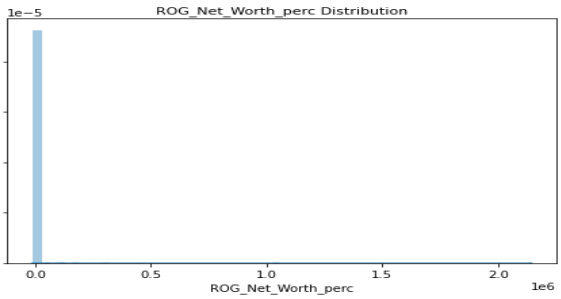

Cost_of_Production Boxplot
400000
300000
200000
100000
0
Cost_of_Prod

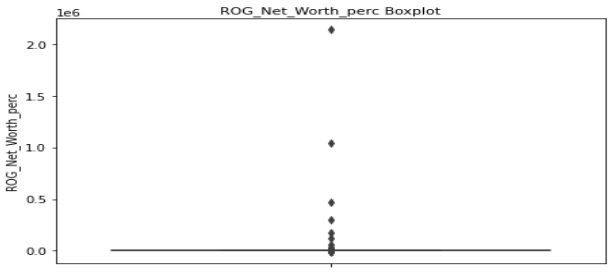

Book_Value_Unit_Curr Distribution
Book_Value_Unit_Curr

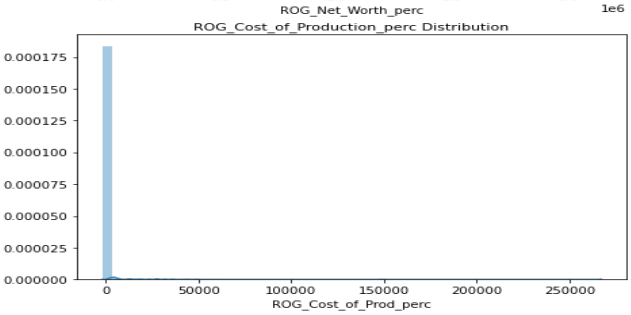

Book_Value_Unit_Curr Boxplot
Book_Value_Unit_Curr

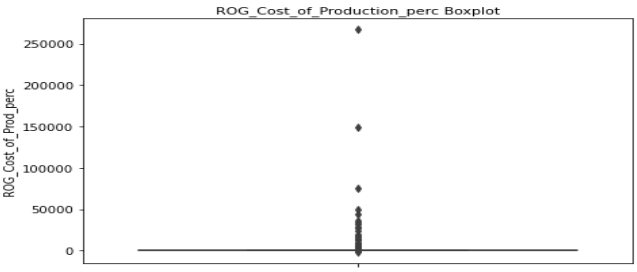

Book_Value_Adj_Unit_Curr Distribution
1e−6
Book_Value_Adj_Unit_Curr
1e6

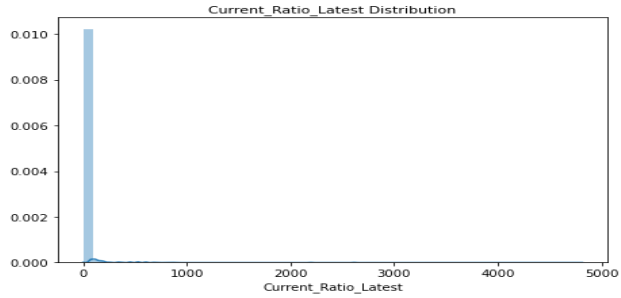

Book_Value_Adj_Unit_Curr Boxplot
1e6
Book_Value_Adj_Unit_Curr

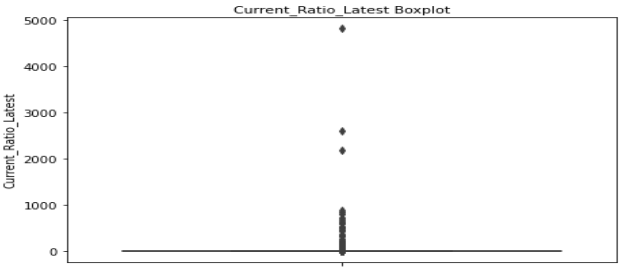

ROG_Net_Worth_perc Distribution
1e−5
ROG_Net_Worth_perc
1e6

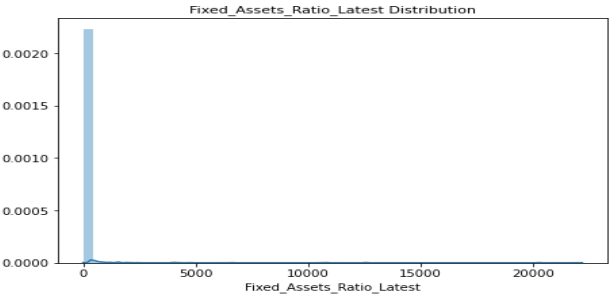

ROG_Net_Worth_perc Boxplot
1e6
ROG_Net_Worth_perc

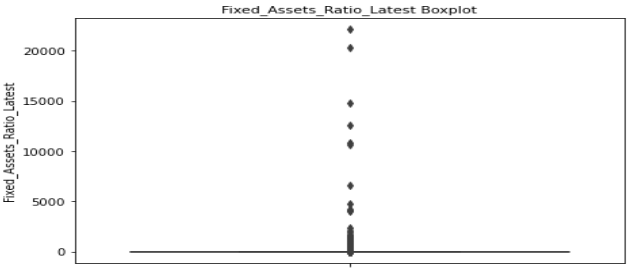

ROG_Cost_of_Production_perc Distribution
ROG_Cost_of_Prod_perc

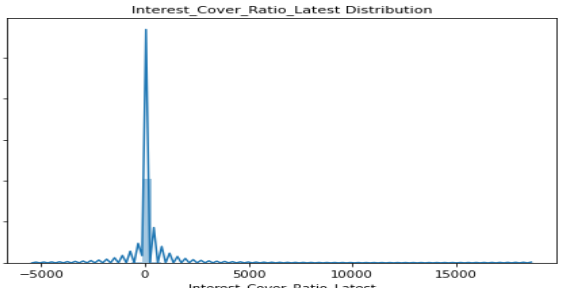

ROG_Cost_of_Production_perc Boxplot
ROG_Cost_of_Prod_perc

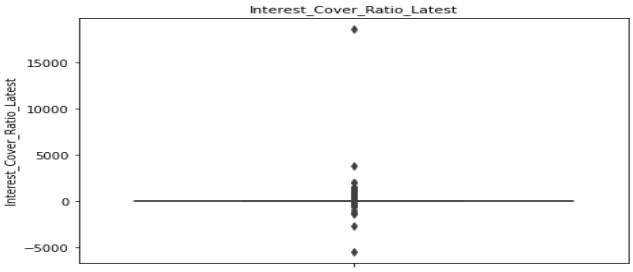

Current_Ratio_Latest Distribution
Current_Ratio_Latest

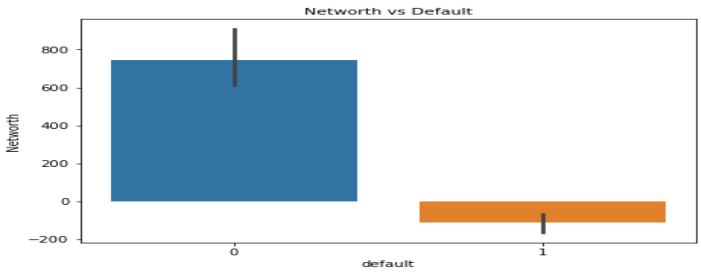

Current_Ratio_Latest Boxplot
Current_Ratio_Latest

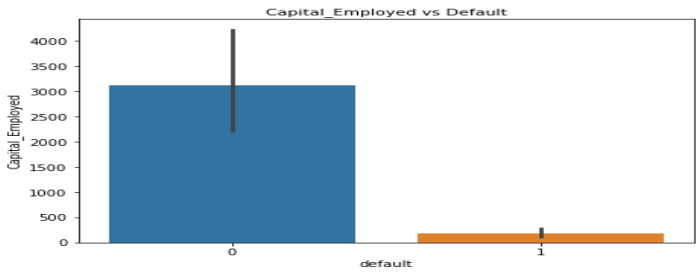

Fixed_Assets_Ratio_Latest Distribution
Fixed_Assets_Ratio_Latest

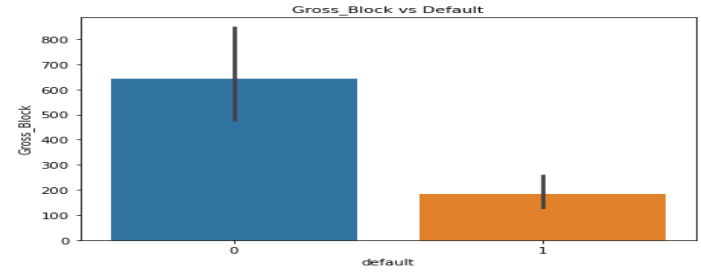

Fixed_Assets_Ratio_Latest Boxplot
Fixed_Assets_Ratio_Latest

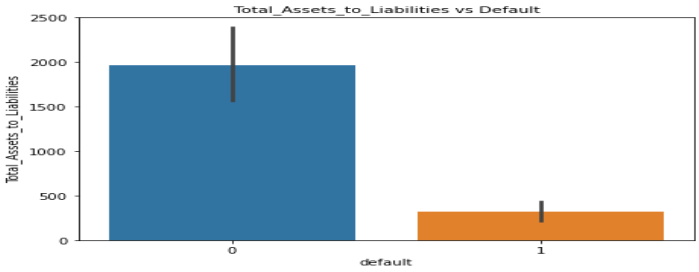

Interest_Cover_Ratio_Latest Distribution

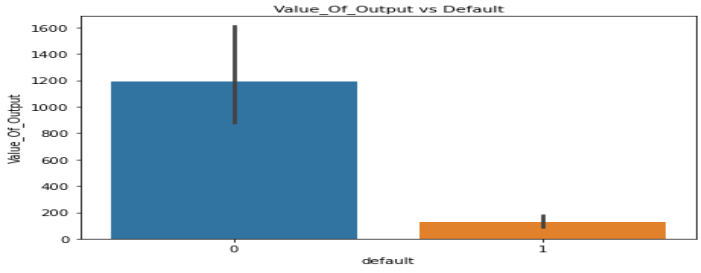

Interest_Cover_Ratio_Latest
Interest_Cover_Ratio_Latest

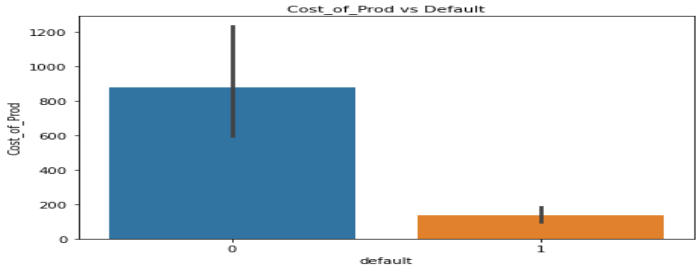

Networth vs Default
Networth
default

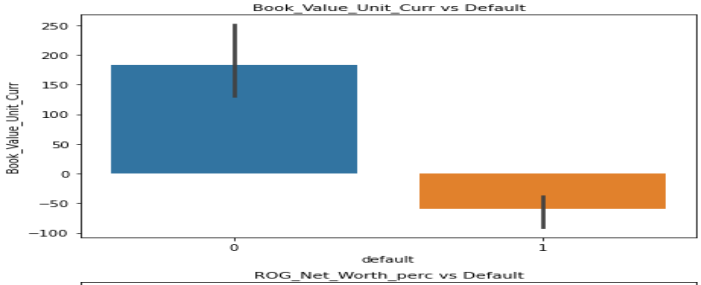

Capital_Employed vs Default
Capital_Employed
default

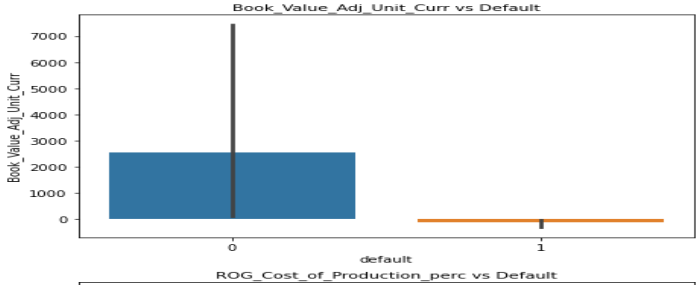

Gross_Block vs Default
Gross_Block
default

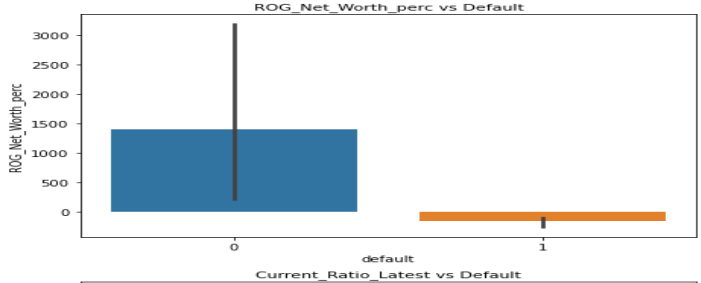

Total_Assets_to_Liabilities vs Default
Total_Assets_to_Liabilities
default

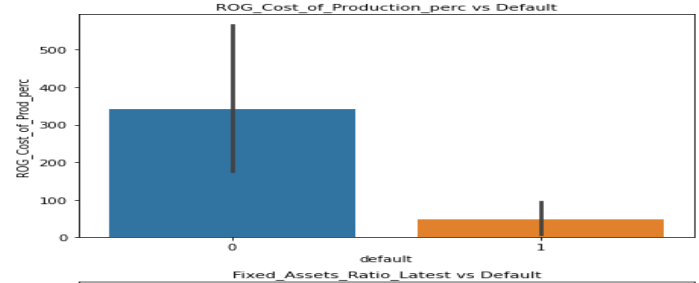

Value_Of_Output vs Default
Value_Of_Output
default

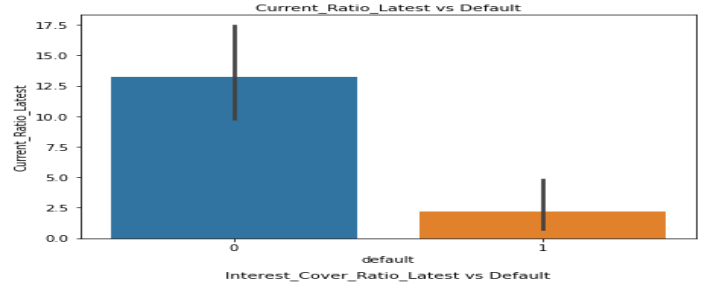

Cost_of_Prod vs Default
Cost_of_Prod
default

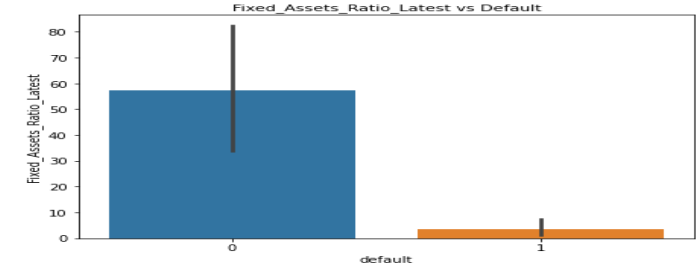

Book_Value_Unit_Curr vs Default
Book_Value_Unit_Curr
default

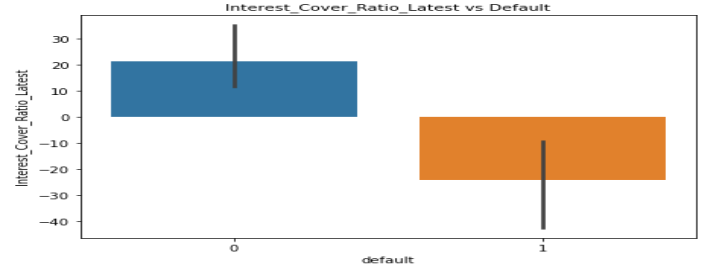

Book_Value_Adj_Unit_Curr vs Default
Book_Value_Adj_Unit_Curr
default

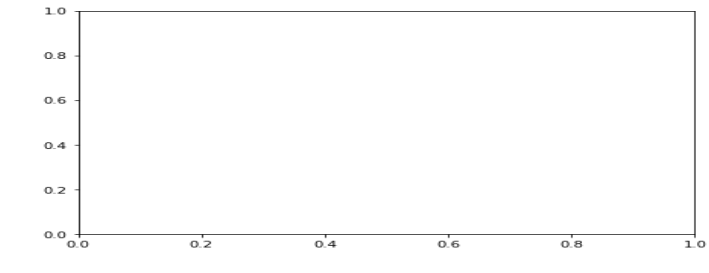

ROG_Net_Worth_perc vs Default
ROG_Net_Worth_perc
default


ROG_Cost_of_Production_perc vs Default
ROG_Cost_of_Prod_perc
default


Current_Ratio_Latest vs Default
Current_Ratio_Latest
default


Fixed_Assets_Ratio_Latest vs Default
Fixed_Assets_Ratio_Latest
default

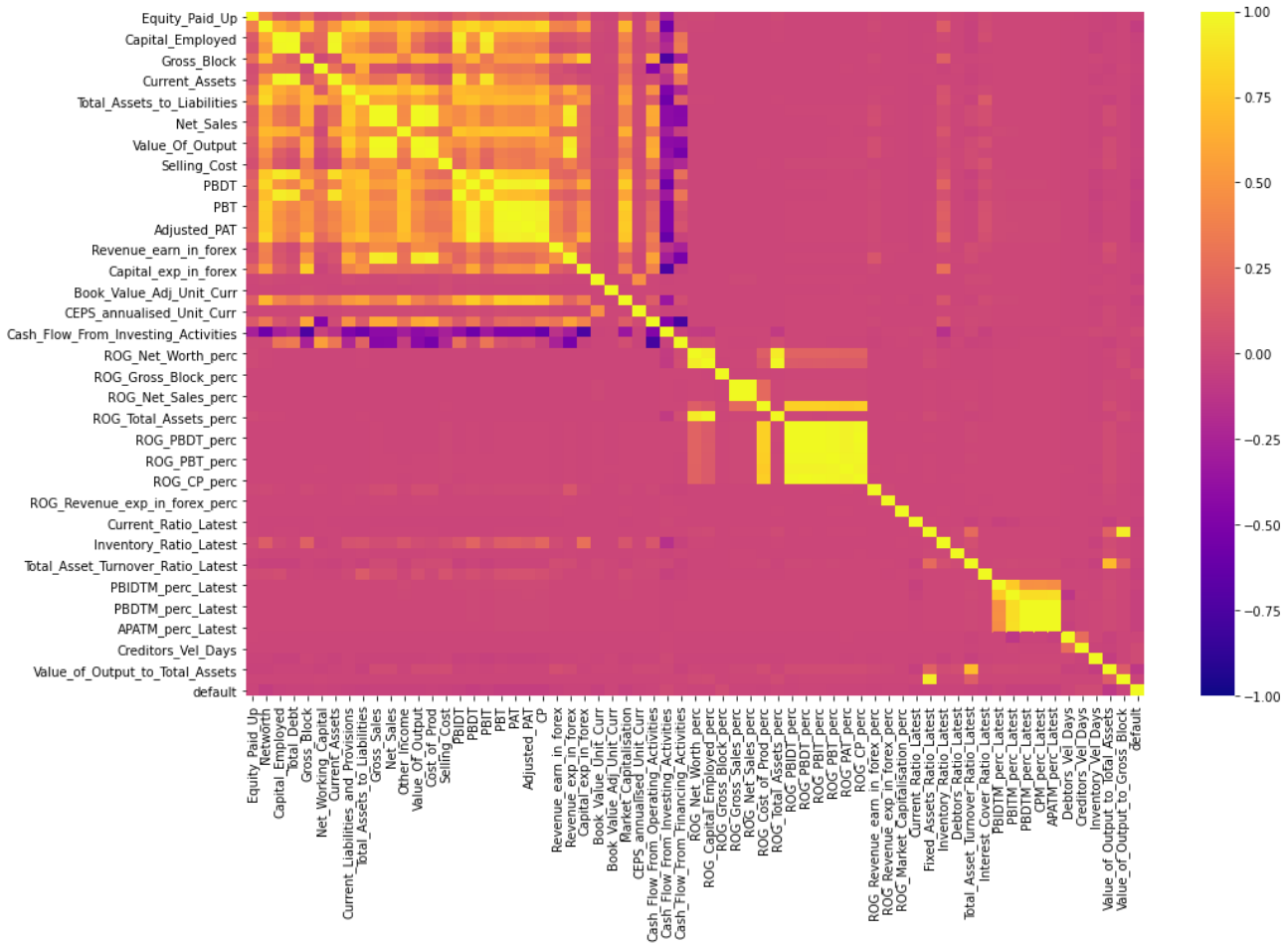

Interest_Cover_Ratio_Latest vs Default
Interest_Cover_Ratio_Latest
default



Bivariate analysis of significant variables vs default

- Clear distinction for the default variable can be identified from the Networth, Book_value_unit_curr, Book_value_Adj_unit_curr, ROG_Networth_perc and Interest_cover_ratio_latest. In these fields the value of default is in the negative side whereas the value for non-default is in the positive side.

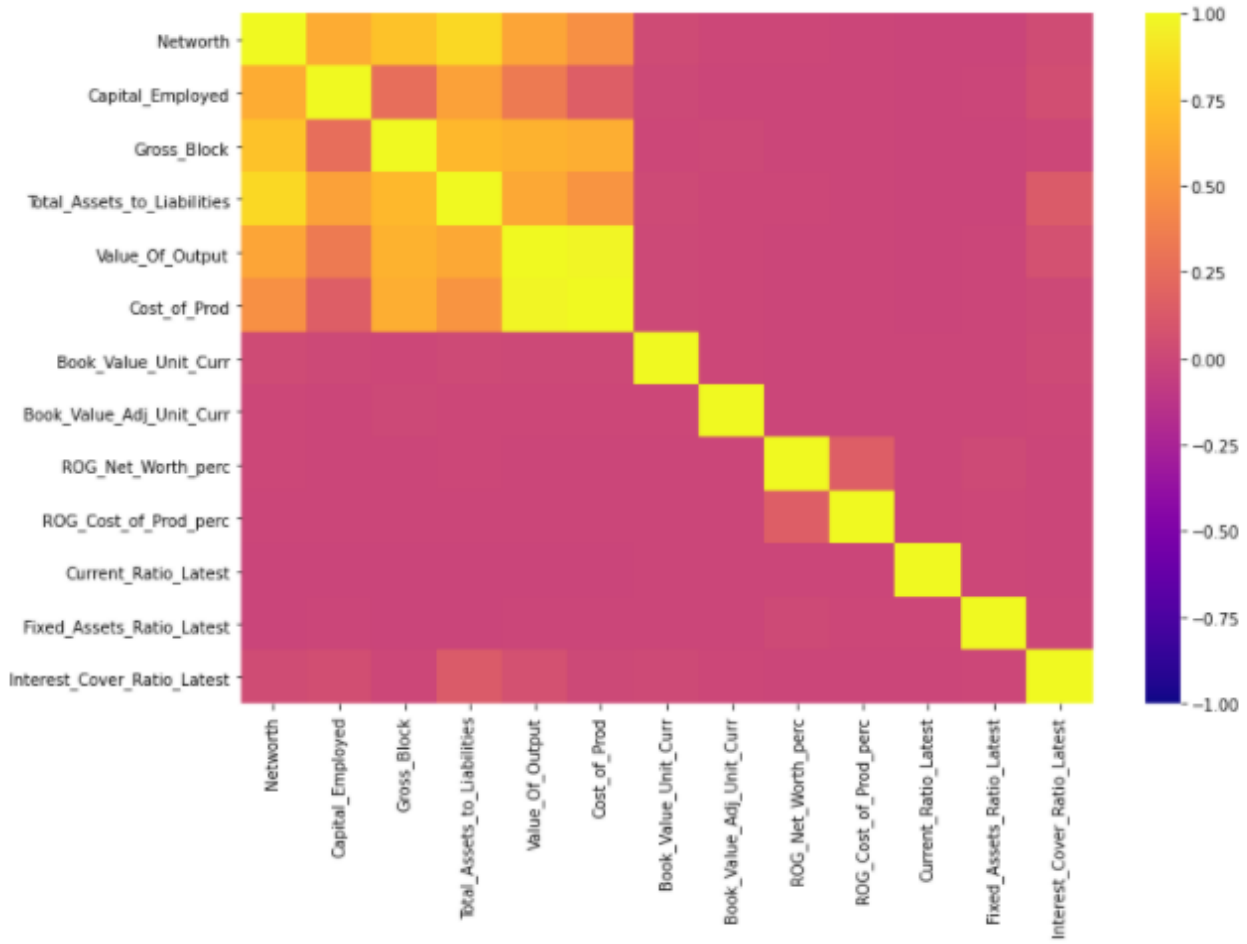
Correlation plot of all variables before imputing



There are variables which are highly correlated either positively or negatively. These variables might not be useful in the model building and prediction of default variable.

Default variable as such does not have any correlation with other fields.

Now, when we look at the correlation plot of the significant variables got from the model building, we can see that most of the variables are not correlated with each other, except a few. Variables cost_of_prod and value_of_output are the most highly correlated independent variables. These correlations have been depicted by the pair plot in the python file.



Missing values

```
Total null values in the dataset:  118
```

	Co_Code	Co_Name	Networth_Next_Year	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block	Net_Working_Capital	Current_Assets
24	5135	Alps Inds.	-647.90	39.11	-668.12	914.09	1080.44	816.58	425.64	544.97
97	3058	Ashima	-77.29	33.37	-363.21	204.64	539.72	205.81	130.00	171.47
153	24971	Runeecha Textile	-26.87	23.57	-5.27	46.09	48.79	92.84	35.22	43.39
168	2891	Apple Credit	-20.57	19.42	-20.10	-2.85	10.34	0.45	-2.92	34.29
170	2681	Navcom Inds.	-19.26	8.23	-19.22	-19.22	0.00	0.01	-19.23	4.12

5 rows × 68 columns

Check for missing values in each row:

Rows that have more than 5 missing values

	Co_Code	Co_Name	Networth_Next_Year	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block	Net_Working_Capital	Current_Assets
2825	3240	G M Breweries	180.99	11.71	126.2	158.3	32.1	130.8	-7.65	39.86

1 rows × 68 columns

Since there is only one, we can impute the missing values of this row, instead of dropping it.

Outliers

These are the number of outliers in each columns whose values are greater than or less than the upper and lower limit respectively.

Equity_Paid_Up	448	Cash_Flow_From_Financing_Activities	1005
Networth	650	ROG_Net_Worth_perc	747
Capital_Employed	596	ROG_Capital_Employed_perc	572
Total_Debt	583	ROG_Gross_Block_perc	830
Gross_Block	540	ROG_Gross_Sales_perc	671
Net_Working_Capital	625	ROG_Net_Sales_perc	667
Current_Assets	577	ROG_Cost_of_Prod_perc	675
Current_Liabilities_and_Provisions	581	ROG_Total_Assets_perc	483
Total_Assets_to_Liabilities	574	ROG_PBIDT_perc	611
Gross_Sales	554	ROG_PBDT_perc	628
Net_Sales	556	ROG_PBIT_perc	616
Other_Income	603	ROG_PBT_perc	611
Value_Of_Output	559	ROG_PAT_perc	598
Cost_of_Prod	560	ROG_CP_perc	637
Selling_Cost	605	ROG_Revenue_earn_in_forex_perc	1317
PBIDT	671	ROG_Revenue_exp_in_forex_perc	1615
PBDT	815	ROG_Market_Capitalisation_perc	497
PBIT	720	Current_Ratio_Latest	565
PBT	941	Fixed_Assets_Ratio_Latest	495
PAT	959	Inventory_Ratio_Latest	375
Adjusted_PAT	954	Debtors_Ratio_Latest	371
CP	816	Total_Asset_Turnover_Ratio_Latest	201
Revenue_earn_in_forex	738	Interest_Cover_Ratio_Latest	725
Revenue_exp_in_forex	693	PBIDTM_perc_Latest	595
Capital_exp_in_forex	694	PBITM_perc_Latest	717
Book_Value_Unit_Curr	485	PBDTM_perc_Latest	695
Book_Value_Adj_Unit_Curr	486	CPM_perc_Latest	720
Market_Capitalisation	639	APATM_perc_Latest	933
CEPS_annualised_Unit_Curr	602	Debtors_Vel_Days	398
Cash_Flow_From_Operating_Activities	801	Creditors_Vel_Days	391
Cash_Flow_From_Investing_Activities	876	Inventory_Vel_Days	262
		Value_of_Output_to_Total_Assets	150
		Value_of_Output_to_Gross_Block	481
		dtype: int64	

As we saw from the summary statistics of these variables, these variables have a lot of values as 0. Hence, converting these outliers to null values and imputing them using KNN (K Nearest Neighbours) imputer in order to get a comparable outcome.

After converting the outliers to null values:

```
Total null values in the dataset:  41473
```

Creating a dataset with records having more than 5 null values after converting the outliers to null:

Value count of default of the newly created dataset:

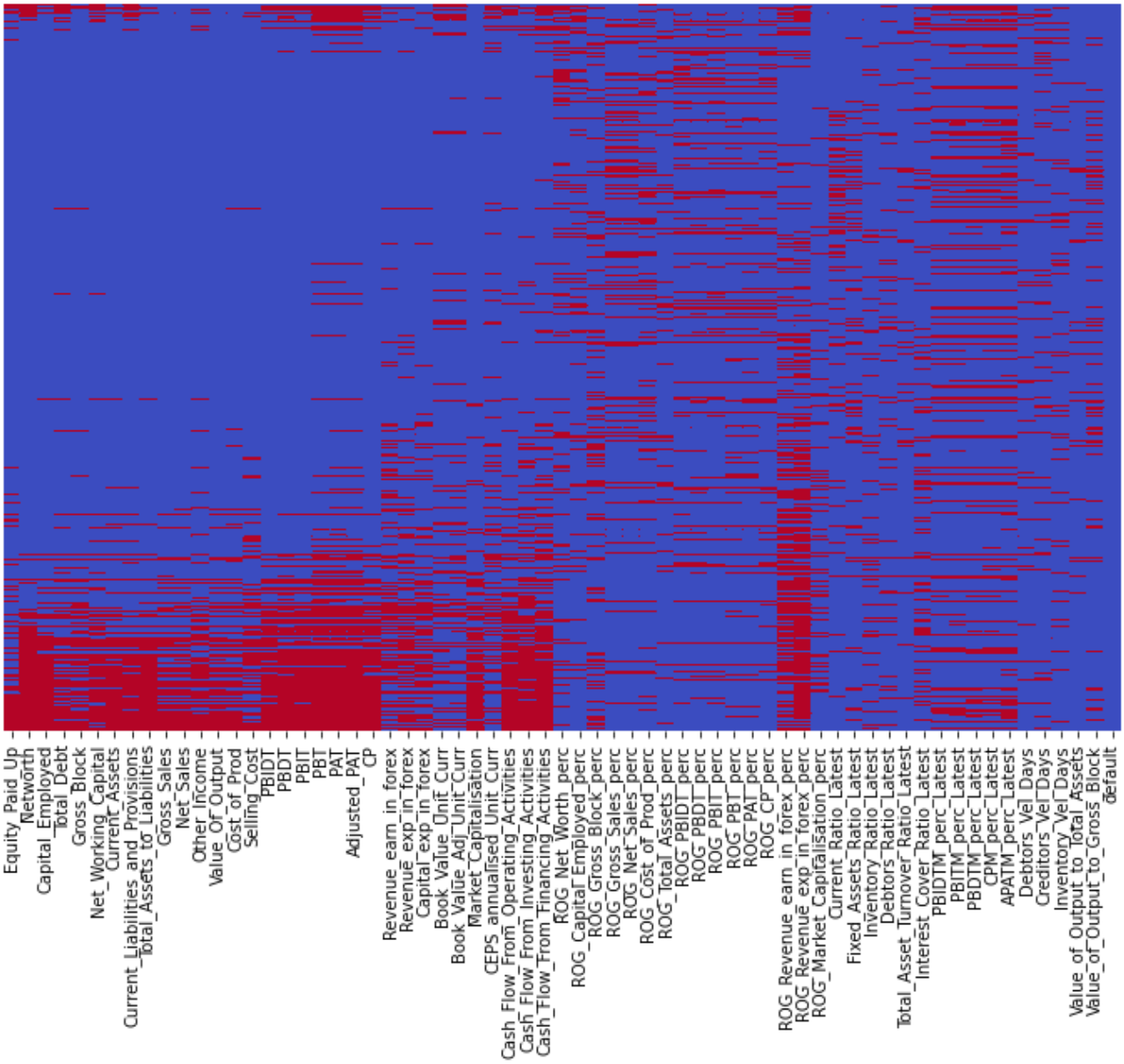
```
0    2326
1     292
Name: default, dtype: int64
```

Value count of the original dataset:

```
0    3198
1     388
Name: default, dtype: int64
```

If we are considering availability of features for deciding the observations to be considered then we would lose up to 75% of the actual default. Hence, imputing these null values using KNN imputer from sklearn.

Visual representation of missing values:



Inspecting the columns for the percentage of missing values:

Top columns with missing values percentage greater than 25% are

ROG_Revenue_exp_in_forex_perc	0.450363
ROG_Revenue_earn_in_forex_perc	0.367262
Cash_Flow_From_Financing_Activities	0.280257
PAT	0.267429
Adjusted_PAT	0.266035
PBT	0.262409
APATM_perc_Latest	0.260457

Removing these columns before model building.

Missing value and outlier treatment

For this dataset, converted the outlier to null values so that both can be treat together using KNN imputer. KNN method is to identify 'k' samples in the dataset that are similar to other data points. Then using these 'k' samples to estimate the value of the missing data. Each missing values are imputed using the mean value of the k-neighbours found in the dataset.

After imputing, head of the dataset:

	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block	Net_Working_Capital	Current_Assets	Current_Liabilities_and_Provisions
0	23.746	68.469	71.144	49.995	172.258	0.000	40.500	40.388
1	10.871	-3.802	0.732	2.913	8.490	-1.722	3.732	5.449
2	19.431	181.316	377.636	51.842	217.240	72.327	235.171	106.390
3	18.326	64.958	179.874	48.934	149.370	49.616	119.568	47.866
4	24.055	166.635	352.921	56.499	199.574	37.377	233.286	86.089

5 rows × 58 columns

Train and test split

For the given business problem, 'default' is the target variable since the problem is to come up with a model to predict whether a particular company will default or not.

X – Independent variable (Removing 'default' variable)

Y – Dependent/ Target variable (Having only 'default' variable)

Next step is to Split the data into training and testing test. Splitting the data as 67% training and 33% testing with a random state = 42 and stratify = y. Output of this step will be: Training independent variable (X_train), Testing independent variable (X-test), Training dependent variable (y_train) and testing dependent variable (y_test).

Proportion of default in train dataset:

```
0    0.891757
1    0.108243
Name: default, dtype: float64
```

Proportion of default in test dataset:

```
0    0.891892
1    0.108108
Name: default, dtype: float64
```

Logistic Regression model (using statsmodel library)

Logistic Regression is a supervised learning method for classification. It establishes relationship between dependent class variables and independent class variables using regression. Logistic regression assign probabilities to different classes to which a data point is likely to belong. In order to do this, the classifier takes the weighted sum of the features and bias to represent the class of interest of a particular data point, this linear output is passed through a sigmoid function in order to get the values between the range (0,1).

Using **logit function from statsmodels** in order to determine the p-value of variables and to determine if it's a good predictor. This approach requires the labelled data, therefore concatenating the train and test labels to the respective datasets.

Model building approach

Started building the model with ten variables in the formula.

Logit Regression Results

Dep. Variable:	default	**No. Observations:**	2402
Model:	Logit	**Df Residuals:**	2392
Method:	MLE	**Df Model:**	9
Date:	Sun, 20 Jun 2021	**Pseudo R-squ.:**	0.3249
Time:	19:11:58	**Log-Likelihood:**	-555.96
converged:	True	**LL-Null:**	-823.47
Covariance Type:	nonrobust	**LLR p-value:**	1.808e-109

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.0877	0.119	-17.494	0.000	-2.322	-1.854
Equity_Paid_Up	0.0432	0.011	3.772	0.000	0.021	0.066
Networth	-0.0486	0.004	-11.054	0.000	-0.057	-0.040
Capital_Employed	-0.0187	0.003	-6.298	0.000	-0.025	-0.013
Total_Debt	0.0108	0.004	2.482	0.013	0.002	0.019
Gross_Block	0.0131	0.002	5.324	0.000	0.008	0.018
Total_Assets_to_Liabilities	0.0148	0.002	6.287	0.000	0.010	0.019
Value_Of_Output	-0.0180	0.005	-3.999	0.000	-0.027	-0.009
Cost_of_Prod	0.0163	0.005	3.148	0.002	0.006	0.026
Selling_Cost	-0.1626	0.083	-1.964	0.050	-0.325	-0.000

Eliminating the variables which has p-value > 0.05 and running the model again.

Logit Regression Results

Dep. Variable:	default	**No. Observations:**	2402
Model:	Logit	**Df Residuals:**	2392
Method:	MLE	**Df Model:**	9
Date:	Sun, 20 Jun 2021	**Pseudo R-squ.:**	0.3249
Time:	19:14:42	**Log-Likelihood:**	-555.96
converged:	True	**LL-Null:**	-823.47
Covariance Type:	nonrobust	**LLR p-value:**	1.808e-109

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.0877	0.119	-17.494	0.000	-2.322	-1.854
Equity_Paid_Up	0.0432	0.011	3.772	0.000	0.021	0.066
Networth	-0.0486	0.004	-11.054	0.000	-0.057	-0.040
Capital_Employed	-0.0187	0.003	-6.298	0.000	-0.025	-0.013
Total_Debt	0.0108	0.004	2.482	0.013	0.002	0.019
Gross_Block	0.0131	0.002	5.324	0.000	0.008	0.018
Total_Assets_to_Liabilities	0.0148	0.002	6.287	0.000	0.010	0.019
Value_Of_Output	-0.0180	0.005	-3.999	0.000	-0.027	-0.009
Cost_of_Prod	0.0163	0.005	3.148	0.002	0.006	0.026
Selling_Cost	-0.1626	0.083	-1.964	0.050	-0.325	-0.000

Adding next variables and find the p-values:

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.4154	0.142	-9.946	0.000	-1.694	-1.136
Equity_Paid_Up	0.0019	0.014	0.137	0.891	-0.025	0.029
Networth	-0.0238	0.004	-6.047	0.000	-0.031	-0.016
Capital_Employed	-0.0228	0.003	-6.925	0.000	-0.029	-0.016
Total_Debt	0.0104	0.005	2.087	0.037	0.001	0.020
Gross_Block	0.0122	0.003	4.152	0.000	0.006	0.018
Total_Assets_to_Liabilities	0.0175	0.002	7.169	0.000	0.013	0.022
Value_Of_Output	-0.0176	0.005	-3.279	0.001	-0.028	-0.007
Cost_of_Prod	0.0185	0.006	2.983	0.003	0.006	0.031
Selling_Cost	-0.0466	0.091	-0.512	0.609	-0.225	0.132
PBIDT	-0.0303	0.017	-1.790	0.073	-0.064	0.003
PBDT	0.0647	0.046	1.412	0.158	-0.025	0.154
PBIT	0.0363	0.025	1.451	0.147	-0.013	0.085
CP	-0.1304	0.052	-2.523	0.012	-0.232	-0.029
Book_Value_Unit_Curr	-0.0370	0.015	-2.500	0.012	-0.066	-0.008
Book_Value_Adj_Unit_Curr	-0.0473	0.016	-2.883	0.004	-0.080	-0.015

Now after adding few variables, previously useful variables now have a p-value > 0.05. Therefore, they are removed from further model building. Likewise, added variables consecutively in an iterative process and found out the top 13 variables which are significant in logistic regression model building.

Final Logistic Regression model

Important variables

Dep. Variable:	default	No. Observations:	2402
Model:	Logit	Df Residuals:	2388
Method:	MLE	Df Model:	13
Date:	Sun, 20 Jun 2021	Pseudo R-squ.:	0.5588
Time:	19:51:36	Log-Likelihood:	-363.32
converged:	True	LL-Null:	-823.47
Covariance Type:	nonrobust	LLR p-value:	2.252e-188

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.0568	0.167	-0.340	0.734	-0.384	0.270
Networth	-0.0212	0.004	-5.897	0.000	-0.028	-0.014
Capital_Employed	-0.0165	0.003	-5.463	0.000	-0.022	-0.011
Gross_Block	0.0091	0.003	3.522	0.000	0.004	0.014
Total_Assets_to_Liabilities	0.0140	0.002	5.809	0.000	0.009	0.019
Value_Of_Output	-0.0198	0.004	-4.775	0.000	-0.028	-0.012
Cost_of_Prod	0.0256	0.005	4.950	0.000	0.015	0.036
Book_Value_Unit_Curr	-0.0357	0.013	-2.809	0.005	-0.061	-0.011
Book_Value_Adj_Unit_Curr	-0.0420	0.014	-2.941	0.003	-0.070	-0.014
ROG_Net_Worth_perc	-0.0306	0.013	-2.363	0.018	-0.056	-0.005
ROG_Cost_of_Prod_perc	-0.0148	0.005	-2.929	0.003	-0.025	-0.005
Current_Ratio_Latest	-1.0723	0.146	-7.330	0.000	-1.359	-0.786
Fixed_Assets_Ratio_Latest	-0.2250	0.082	-2.729	0.006	-0.387	-0.063
Interest_Cover_Ratio_Latest	-0.2713	0.069	-3.944	0.000	-0.406	-0.137

Selected variables have p-value < 0.05 and are therefore statistically significant.

Fit the model and predict the probabilities on train dataset:

There are two values in default variable – 0 for non-defaulters and 1 for defaulters.

In this problem, our primary focus is to predict the default class (1).

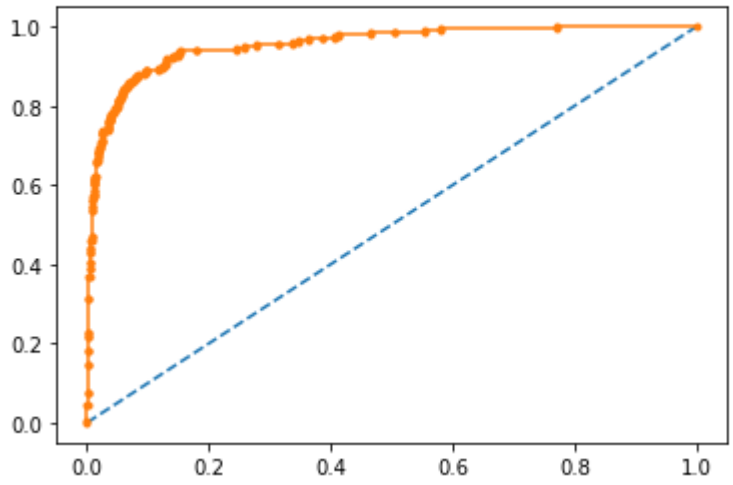
Optimum cut-off

Optimum cut-off Point for making the prediction would be the point where true positive rate is high and false positive rate is low. For calculating the optimal threshold, we would be looking at the ROC-AUC curve of the train dataset:

```
Optimum threshold:  0.16308163831301595
```

Keeping this as a cut-off when predicting the probability class of the train and test dataset.

AUC for the Training Data: 0.956



Performance metrics

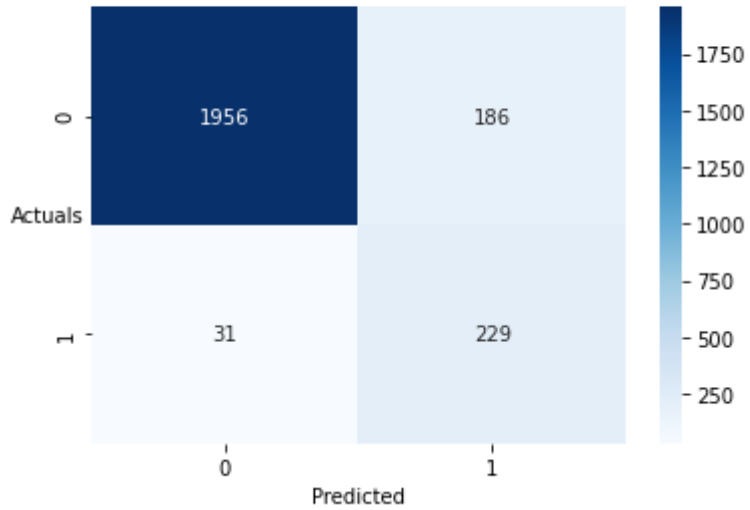
Train dataset

Accuracy of the train dataset: 90.97

Classification Report:

	precision	recall	f1-score	support
0	0.984	0.913	0.947	2142
1	0.552	0.881	0.679	260
accuracy			0.910	2402
macro avg	0.768	0.897	0.813	2402
weighted avg	0.938	0.910	0.918	2402

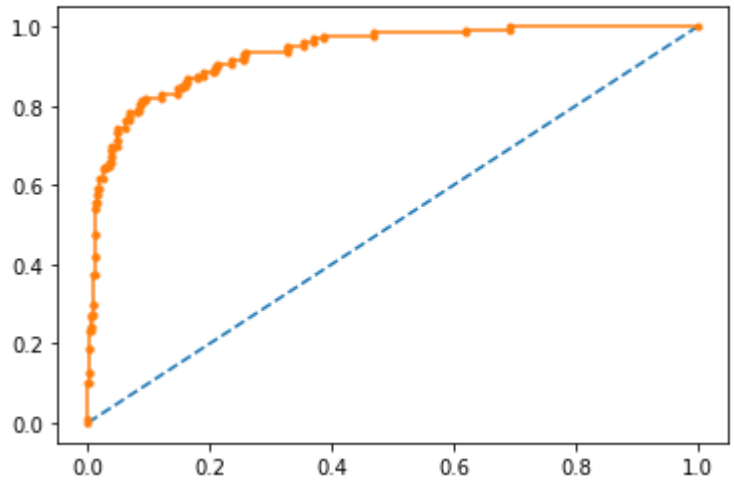
Confusion matrix:



Validating the model on the test dataset and Performance metrics:

Probability class was predicted with the cut-off of 0.13 which was found before.

```
AUC for the Training Data: 0.935
```

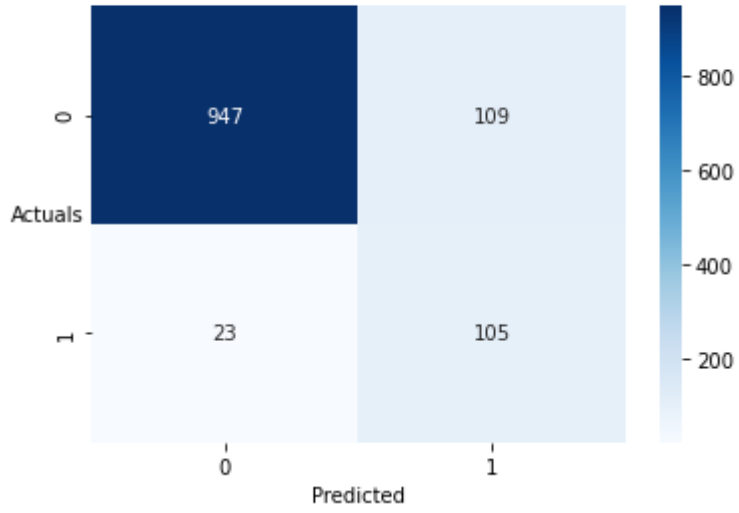



```
Accuracy of the test dataset:  88.85
Classification Report:
              precision    recall  f1-score   support

           0      0.976     0.897     0.935      1056
           1      0.491     0.820     0.614       128

    accuracy                          0.889      1184
   macro avg      0.733     0.859     0.774      1184
weighted avg      0.924     0.889     0.900      1184
```

```
Confusion matrix:
```



Business Insights:

With increasing amount of data, companies and industries try to remain competitive by keeping themselves ahead of the curve. By analysing huge amounts of financial data, companies are able to obtain valuable information to determine their strategic plans such as risk control, crisis management or growth management. Logistic regression model has been employed in predicting the defaulters of companies.

13 significant variables have been identified and model is based on these variables.

Accuracy of the training dataset was 90.97% and testing dataset was 88.85%. This indicates that the model does not overfit and it's a valid model for predicting the default variable.

Consolidated performance metrics (for the default class):

Metrics	**Training set**	**Testing set**
Accuracy	90.97	88.85
Precision	0.55	0.49
Recall	0.88	0.82
F1 score	0.67	0.61

- **False positive (FP)** - Datapoints that are actually non-default but predicted as default. This is also known as type 1 error. In order to reduce the type 1 error, we have to increase the precision of the model (among the points identified as positives by the model how many are actually positive).

  Type 1 error in this case study means model has classified the data point as 1 instead of 0. The company which are identified wrongly as default might try to reassess the net worth, change in working capital or change in strategy as required. They might as well choose to sale their non-profitable investments. This might not be of priority for our case study, since predicting the actual non-defaulters as default might have minimal impact than false negatives tend to pose.

- **False negative (FN)** – Datapoints that are actually default but predicted as non-defaulters. This is known as type 2 error. In order to reduce to type 2 error. We have to increase recall (how many actual true data points are identified as true by the model)

  Type 2 error is that the model has classified the data point as 0 instead of 1. In which case, the company will continue their strategy which makes it risky investment for shareholders. Concentrating on this type of error is a priority.

There is trade-off between precision and recall, this model has chosen a cut off of 0.16 for predicting the probability class which identifies most of the false negative cases.