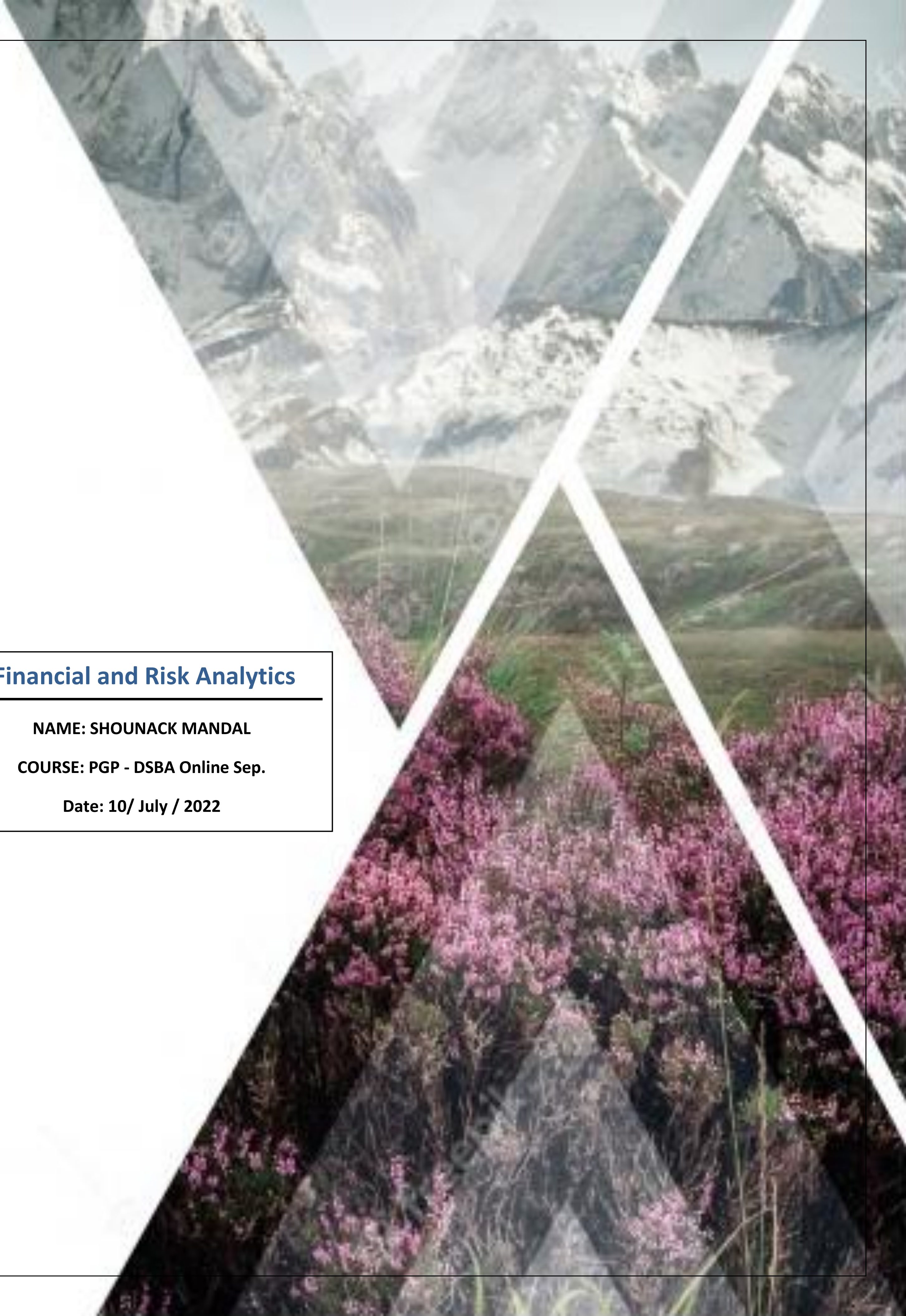


Financial and Risk Analytics

NAME: SHOUNACK MANDAL

COURSE: PGP - DSBA Online Sep.

Date: 10/ July / 2022

Table of contents

LIST OF FIGURE

LIST OF TABLES

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.

Project Objective

- Understanding the structure of dataset.
- Exploratory Data analysis
- Graphical exploration
- Prediction using various machine learning models
- Insights from the dataset

1.1 Outlier Treatment and inferences

1.1.1 Information on dataset

As given in the dataset the column names are modified.

	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]	Debtors Velocity (Days)	Creditors Velocity (Days)	Inventory Velocity (Days)	Value of Output/Total Assets	Value of Output/Gross Block
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	0	0	45.0	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	29	101	2.0	0.31	0.24
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	97	558	0.0	-0.03	-0.26
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	93	63	2.0	0.24	1.90
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	3887	346	0.0	0.01	0.05

Table 1 Information of the dataset head.

	#	Field Name	Description	New Field Name
0	1	Co_Code	Company Code	Co_Code
1	2	Co_Name	Company Name	Co_Name
2	3	Networth Next Year	Value of a company as on 2016 - Next Year(diff...	Networth_Next_Year
3	4	Equity Paid Up	Amount that has been received by the company t...	Equity_Paid_Up
4	5	Networth	Value of a company as on 2015 - Current Year	Networth

Table 2 Data Dictionary

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

Table 3 Column names and data type

There are 118 missing values. For the Credit risk analysis, missing values are not treated using normal methods mostly imputation methods are used before treating missing, it is important to treat outliers for the analysis.

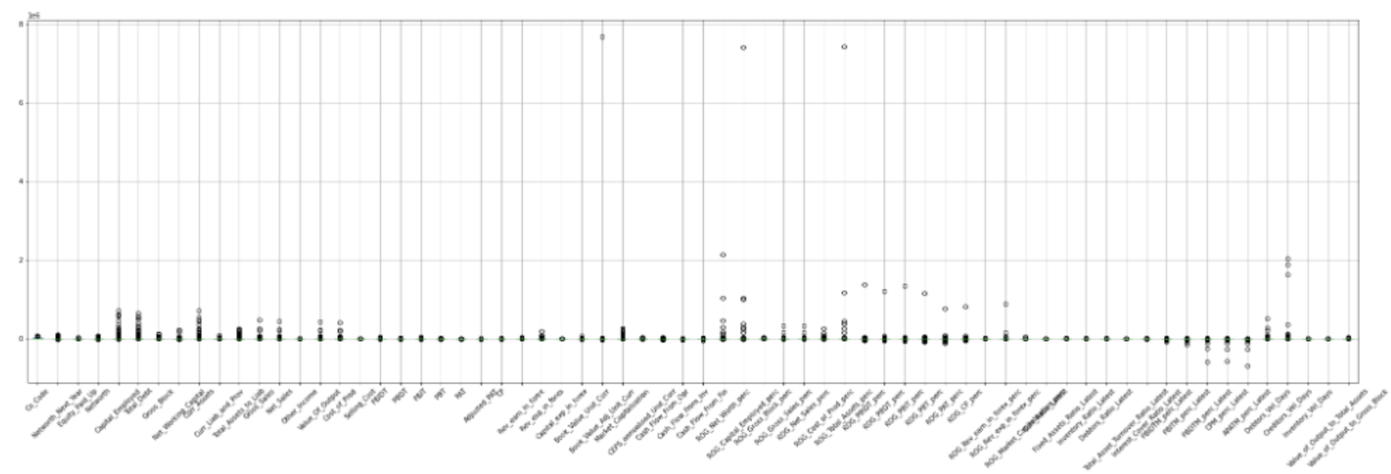


Figure 1 Boxplot for outliers

In many variables there are negative and positive outliers.

greatlearning
Power Ahead

1.1.2 Treatment for outliers

As we have various companies financial data and also the outliers are too large in numbers. But these are not actually outliers so in other scenario we have this financial data and the outliers might very well reflect the information which is genuine in the nature. Since data capture for various parameter of companies so we do not trim or Winsorize data.

1.1.3 Summary of the dataset

	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
Curr_Assets	3586.0	1960.349172	22577.570829	-0.91	4.0000	24.540	135.2775	721166.00
Curr_Liab_and_Prov	3586.0	391.992078	2675.001631	-0.23	0.7325	9.225	65.6500	83232.98
Total_Assets_to_Liab	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

Table 4 Summary f the datasetet(First few rows)

1.1.4 Inference

- The number of rows (observations) is 3586
- The number of columns (variables) is 67
- Minimum Networth_Next_Year - (-8021)
- Maximum Networth_Next_Year - (111729.10)
- Maximum Total Debt - 652823.81
- There are no duplicates in the dataset

1.2 Missing Value Treatment

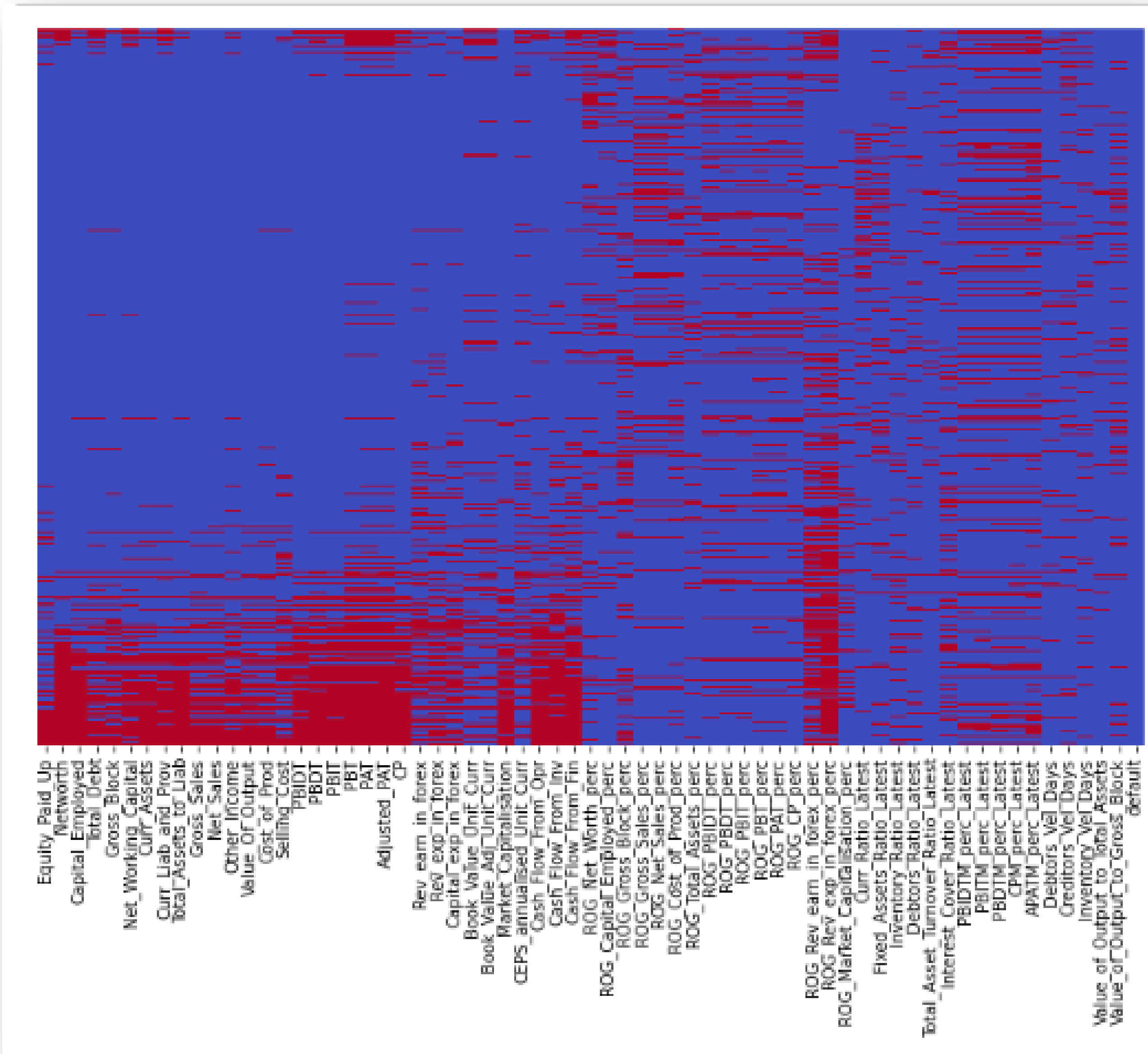


Figure 2 Heat map to visually inspect the missing values

Inspecting Data is scaled and preprocessed before imputating missing data using KNN Imputer. Imputation for completing missing values using k-Nearest Neighbors. Instead the missing values can`t be ignored so we have interpolated with the other financial side-by given data.

Since the outliers are too large in the number.it will affect the model.but Also given the fact that this is a financial data and the outliers might very well reflect the information which is genuine in nature. Since data captured from different size of companies

Equity_Paid_Up	448	ROG_Net_Worth_perc	747
Networth	650	ROG_Capital_Employed_perc	572
Capital_Employed	596	ROG_Gross_Block_perc	830
Total_Debt	583	ROG_Gross_Sales_perc	671
Gross_Block	540	ROG_Net_Sales_perc	667
Net_Working_Capital	625	ROG_Cost_of_Prod_perc	675
Curr_Assets	577	ROG_Total_Assets_perc	483
Curr_Liab_and_Prov	581	ROG_PBIDT_perc	611
Total_Assets_to_Liab	574	ROG_PBDT_perc	628
Gross_Sales	554	ROG_PBIT_perc	616
Net_Sales	556	ROG_PBT_perc	611
Other_Income	603	ROG_PAT_perc	598
Value_Of_Output	559	ROG_CP_perc	637
Cost_of_Prod	560	ROG_Rev_earn_in_forex_perc	1317
Selling_Cost	605	ROG_Rev_exp_in_forex_perc	1615
PBIDT	671	ROG_Market_Capitalisation_perc	497
PBDT	815	Curr_Ratio_Latest	566
PBIT	720	Fixed_Assets_Ratio_Latest	496
PBT	941	Inventory_Ratio_Latest	376
PAT	959	Debtors_Ratio_Latest	372
Adjusted_PAT	954	Total_Asset_Turnover_Ratio_Latest	202
CP	816	Interest_Cover_Ratio_Latest	726
Rev_earn_in_forex	738	PBIDTM_perc_Latest	596
Rev_exp_in_forex	693	PBITM_perc_Latest	718
Capital_exp_in_forex	694	PBDTM_perc_Latest	696
Book_Value_Unit_Curr	485	CPM_perc_Latest	721
Book_Value_Adj_Unit_Curr	490	APATM_perc_Latest	934
Market_Capitalisation	639	Debtors_Vel_Days	398
CEPS_annualised_Unit_Curr	602	Creditors_Vel_Days	391
Cash_Flow_From_Opr	801	Inventory_Vel_Days	365
Cash_Flow_From_Inv	876	Value_of_Output_to_Total_Assets	150
Cash_Flow_From_Fin	1005	Value_of_Output_to_Gross_Block	481
ROG_Net_Worth_perc	747	dtype: int64	

Table 6 Table showing outliers with null value counts

Although most outliers have nan values which is a missing data which should be treated with missing data imputation method so here KNN imputation method is used.

Dropping columns with more than 30% missing values.

```
Equity_Paid_Up                 0
Networth                       0
Capital_Employed               0
Total_Debt                     0
Gross_Block                    0
Net_Working_Capital            0
Curr_Assets                    0
Curr_Liab_and_Prov             0
Total_Assets_to_Liab           0
Gross_Sales                    0
Net_Sales                      0
Other_Income                   0
Value_Of_Output                0
Cost_of_Prod                   0
Selling_Cost                   0
PBIDT                          0
PBDT                           0
PBIT                           0
PBT                            0
PAT                            0
Adjusted_PAT                   0
CP                             0
Rev_earn_in_forex              0
Rev_exp_in_forex               0
Capital_exp_in_forex           0
Book_Value_Unit_Curr           0
Book_Value_Adj_Unit_Curr       0
Market_Capitalisation          0
CEPS_annualised_Unit_Curr      0
Cash_Flow_From_Opr             0
Cash_Flow_From_Inv             0
Cash_Flow_From_Fin             0
ROG_Net_Worth_perc             0
ROG_Capital_Employed_perc      0
ROG_Capital_Employed_perc      0
ROG_Gross_Block_perc           0
ROG_Gross_Sales_perc           0
ROG_Net_Sales_perc             0
ROG_Cost_of_Prod_perc          0
ROG_Total_Assets_perc          0
ROG_PBIDT_perc                 0
ROG_PBDT_perc                  0
ROG_PBIT_perc                  0
ROG_PBT_perc                   0
ROG_PAT_perc                   0
ROG_CP_perc                    0
ROG_Market_Capitalisation_perc 0
Curr_Ratio_Latest              0
Fixed_Assets_Ratio_Latest      0
Inventory_Ratio_Latest         0
Debtors_Ratio_Latest           0
Total_Asset_Turnover_Ratio_Latest 0
Interest_Cover_Ratio_Latest    0
PBIDTM_perc_Latest             0
PBITM_perc_Latest              0
PBDTM_perc_Latest              0
CPM_perc_Latest                0
APATM_perc_Latest              0
Debtors_Vel_Days               0
Creditors_Vel_Days             0
Inventory_Vel_Days             0
Value_of_Output_to_Total_Assets 0
Value_of_Output_to_Gross_Block 0
default                        0
dtype: int64
```

Table 7 Values check after inmputation

Data is scaled and preprocessed before imputating missing data using KNN Imputer.Imputation for completing missing values using k-Nearest Neighbors.

1.3 Transform Target variable into 0 and 1

We have converted next year net-worth into target variable which has binary variable 0 and 1 also considered as responsible variable. All independent variables except default(dependent) variable also known as predictors or independent.

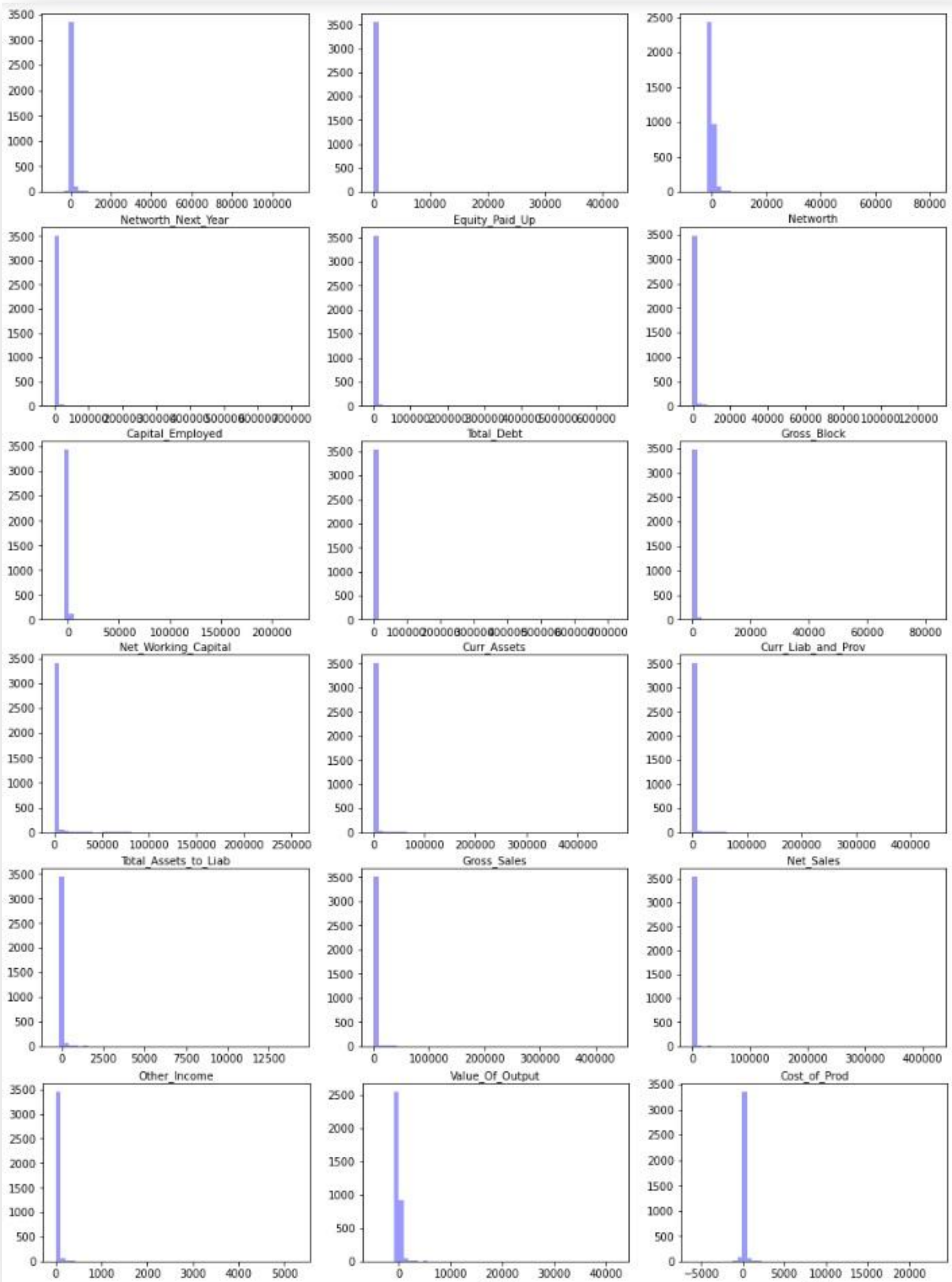
	default	Networth_Next_Year
0	1	-8021.60
1	1	-3986.19
2	1	-3192.58
3	1	-3054.51
4	1	-2967.36
5	1	-2519.40
6	1	-2125.05
7	1	-2100.56
8	1	-1695.75
9	1	-1677.18

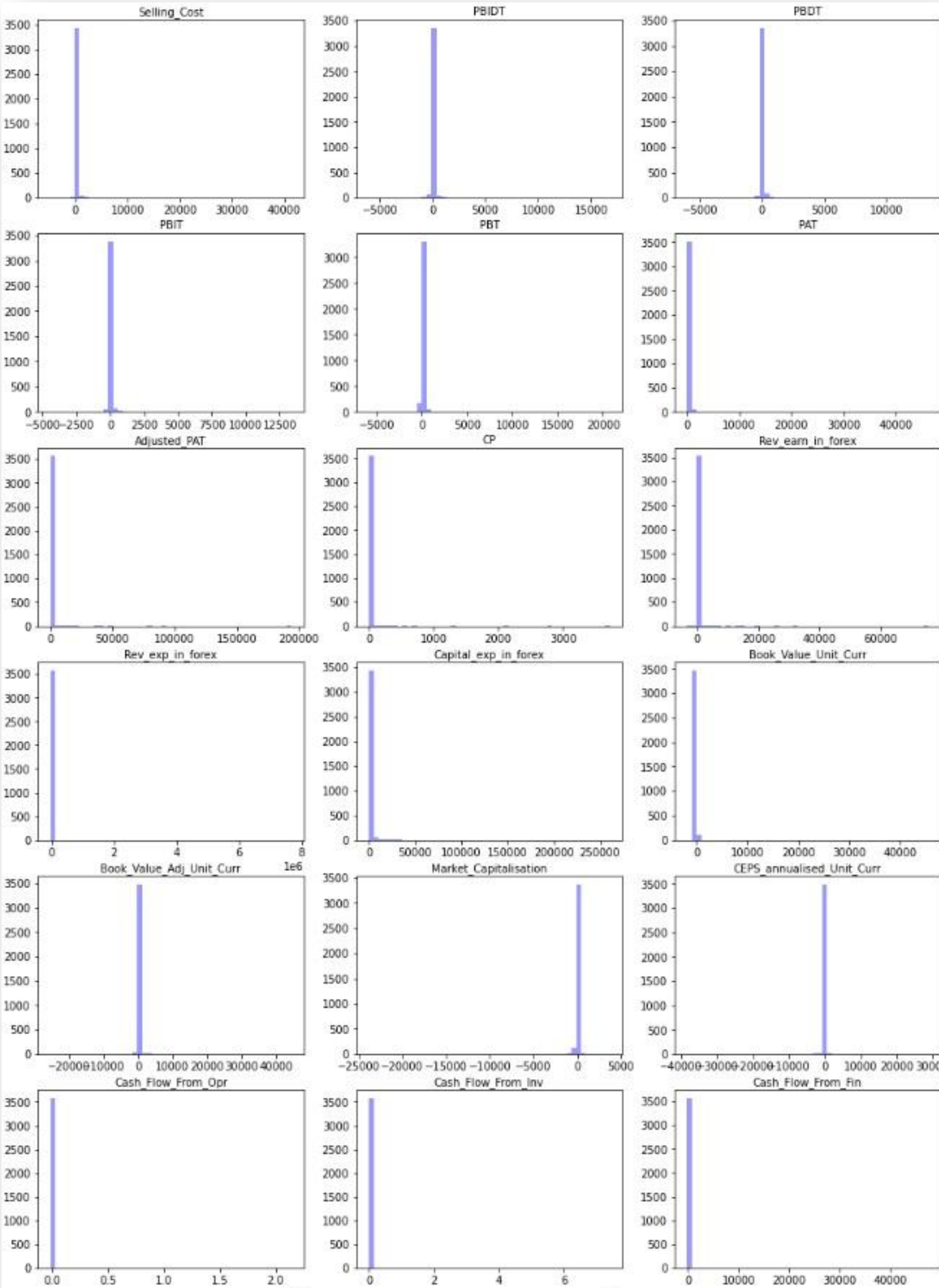
	default	Networth_Next_Year
3576	0	43811.23
3577	0	46637.38
3578	0	47261.30
3579	0	53164.91
3580	0	61082.00
3581	0	72677.77
3582	0	79162.19
3583	0	88134.31
3584	0	91293.70
3585	0	111729.10

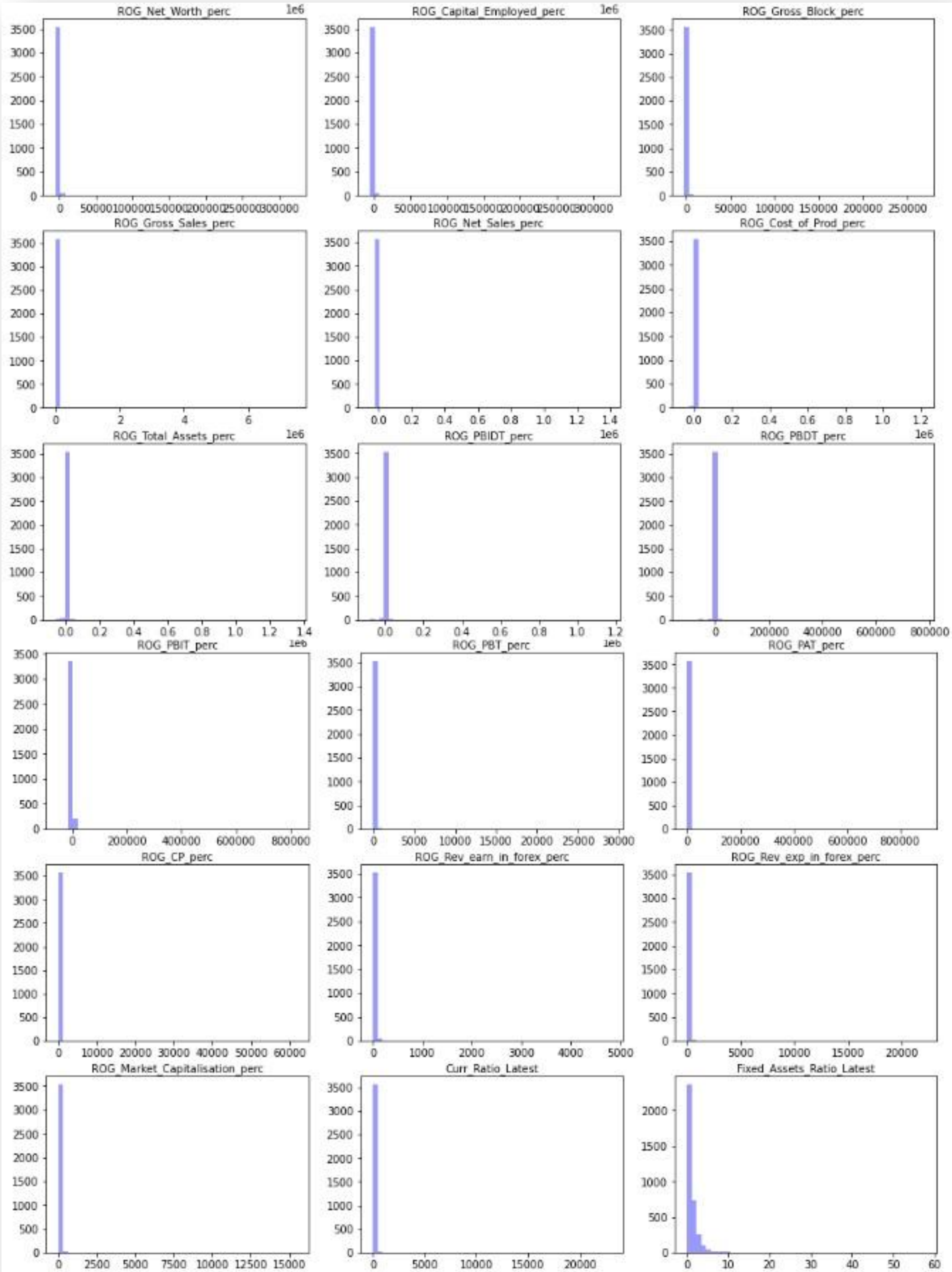
Table 8 Transformed targer variable 0 and 1

1.4 Univariate & Bivariate analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building)

1.4.1 Univariate Analysis




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

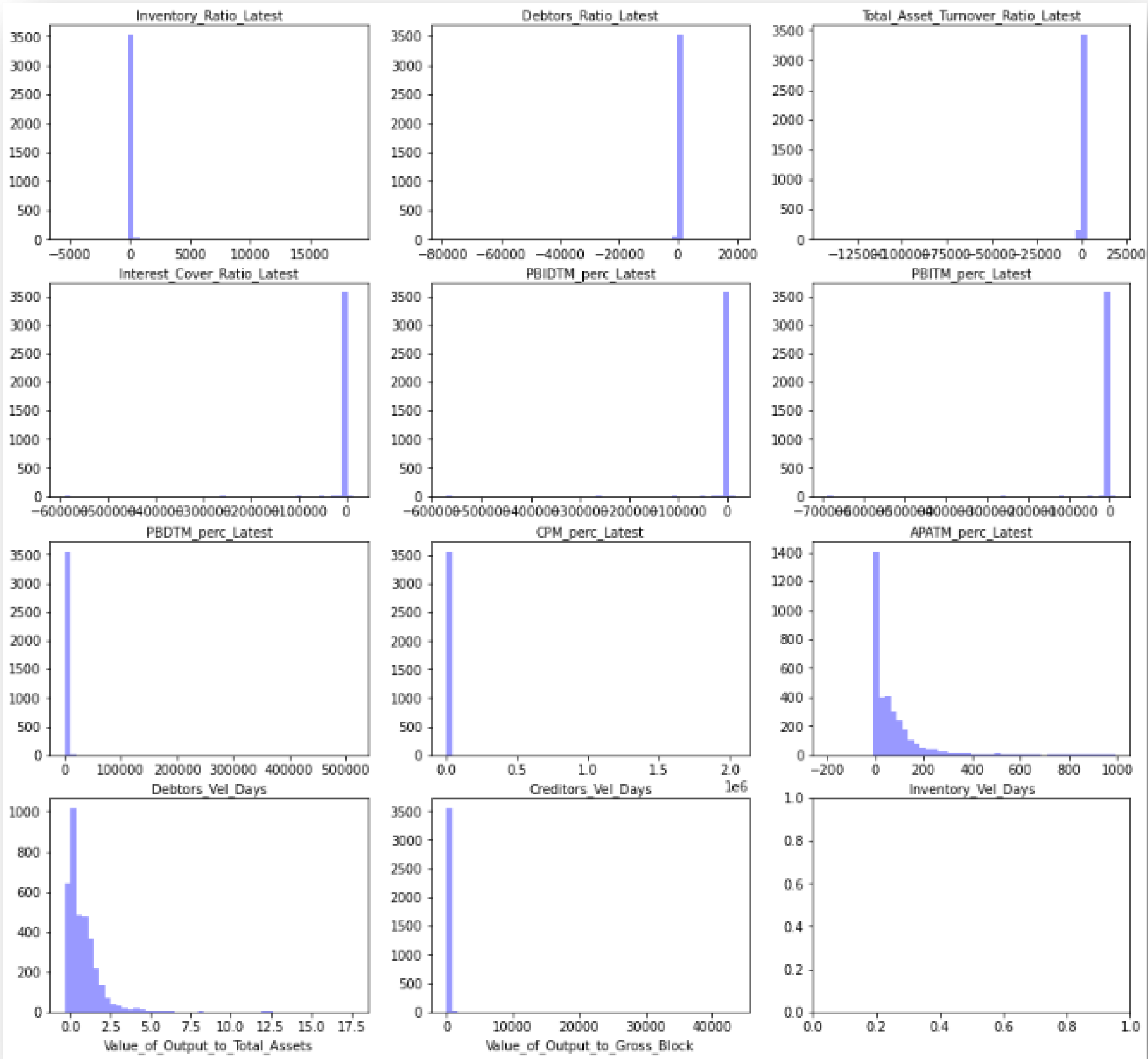


Figure 3 Univariate analysis with histplot

None of the variables show perfect normal distribution. Most of the variables have left positive skewness only six variable right negative skewness.

1.4.1 Bivariate Analysis()

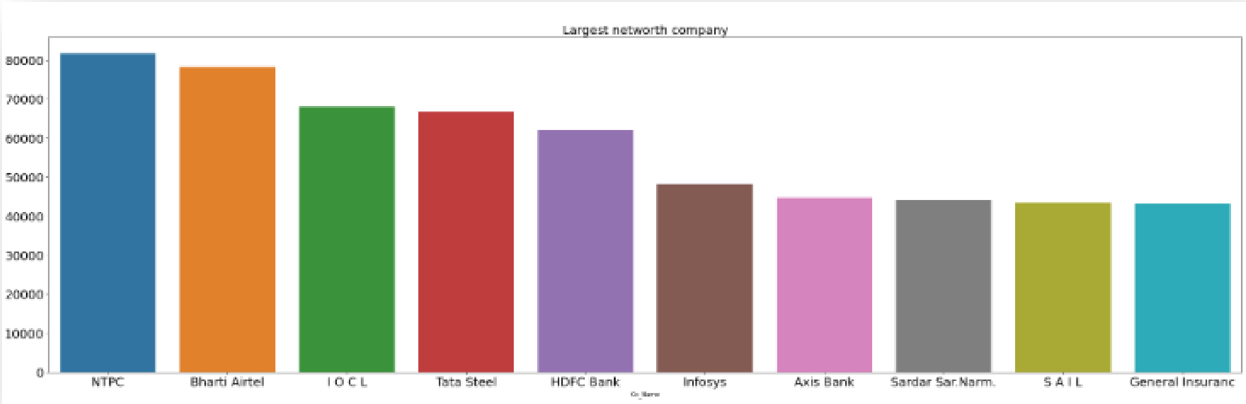


Figure 4 NTPC has highest networth followed by the Bharti Airtel

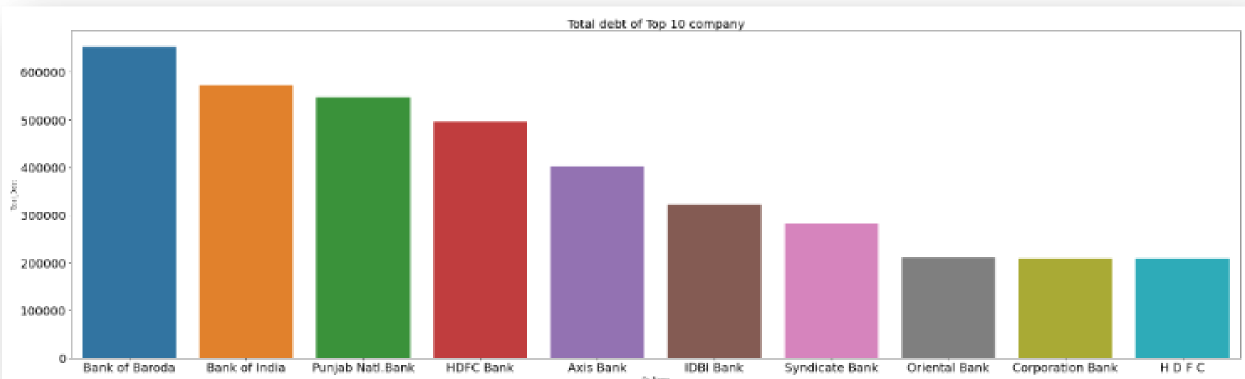


Figure 6 Highest debt Company is Bank of Baroda Second Highest is Bank of india

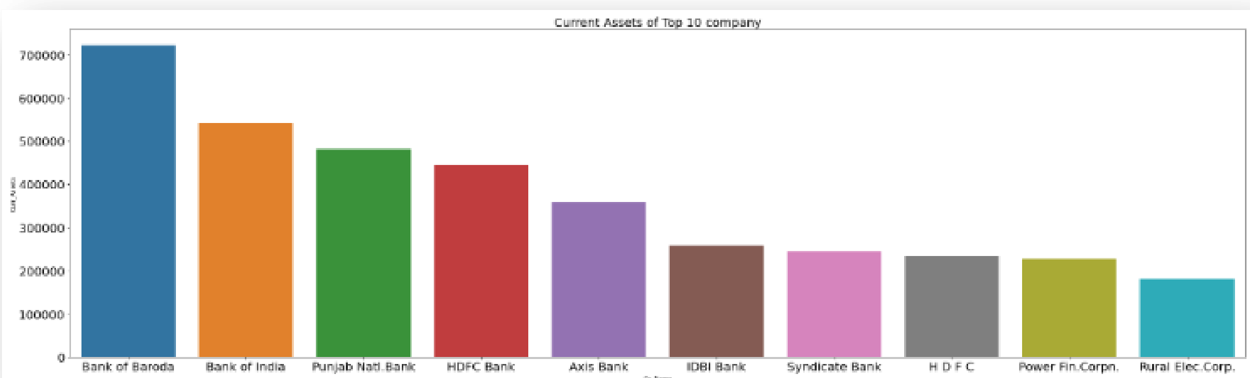


Figure 5 Highest debt Company is Bank of Baroda Second Highest is Bank of india

1.4.2 Univariate Analysis

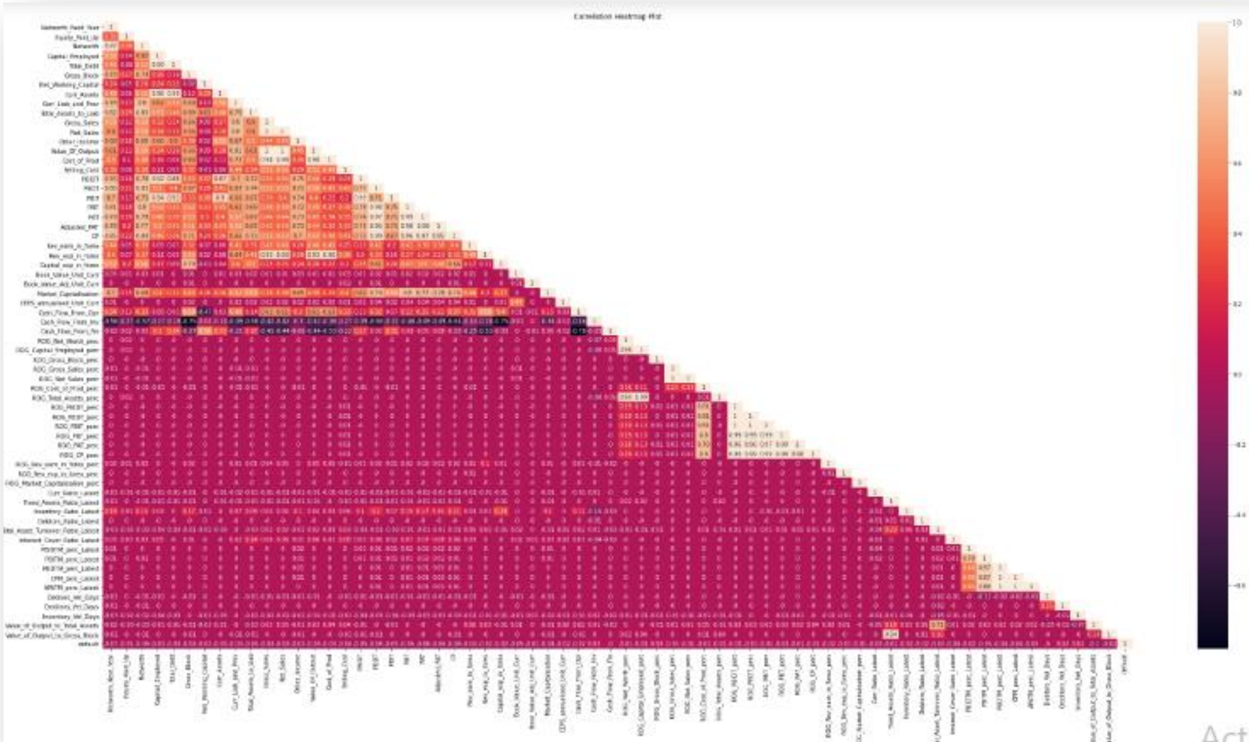


Figure 7 Univariate analysis - heatmap

- We can observe that some of the variables are highly positive correlated and some are slightly negative correlated.
- Given variables are highly correlated. Lighter the color higher the relationship is.

1.4.3 Inspect possible correlations between independent variables

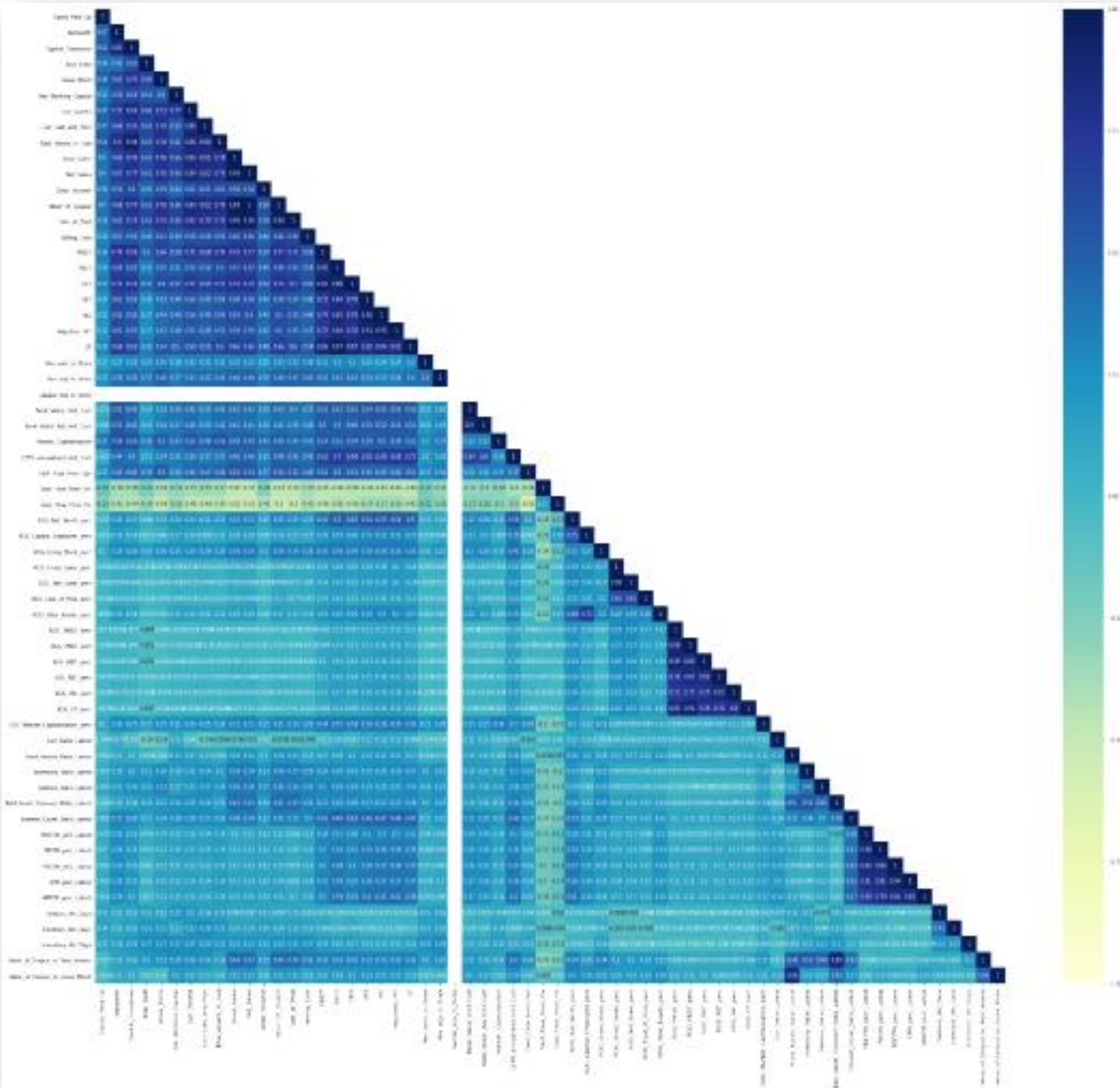


Figure 8 Correlation heat map

Some of the variable is high positively correlated and some of the variables are slightly negative correlated.

1.5 Train Test Split

Splitting the data into train and test sets for the models, test train split ration is as given 67:33 and also have to use random state = 42.

```
The training set for the independent variables: (2402, 62)
The training set for the dependent variable: (2402,)
The test set for the independent variables: (1184, 62)
The test set for the dependent variable: (1184,)
```

Table 9 Table showing train and test split

1.6 Build Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff. Also showcase your model building approach.

1.6.1 Model using logistic regression

Logistic regression using StatsModel:- Statsmodel is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration.

First we have to define set of dependent(y) and independent(x) variables. In case the dependent variable as non numeric form, it is first converted to numeric using dummies. Statsmodel provide a Logit() function for performing logistic regression. The Logit function accepts y and x as parameter and returns the Logit object which then fitted into the data.

Logit Regression Results

Dep. Variable:	default	No. Observations:	2402
Model:	Logit	Df Residuals:	2386
Method:	MLE	Df Model:	15
Date:	Sun, 10 Jul 2022	Pseudo R-squ.:	0.5863
Time:	15:54:22	Log-Likelihood:	-327.37
converged:	True	LL-Null:	-791.34
Covariance Type:	nonrobust	LLR p-value:	3.686e-188

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-5.2239	0.292	-17.872	0.000	-5.797	-4.651
Networth	-1.5555	0.334	-4.664	0.000	-2.209	-0.902
Capital_Employed	-0.7493	0.309	-2.424	0.015	-1.355	-0.143
Gross_Block	0.8500	0.228	3.733	0.000	0.404	1.296
Curr_Liab_and_Prov	0.7379	0.236	3.125	0.002	0.275	1.201
Total_Assets_to_Liab	0.7680	0.306	2.509	0.012	0.168	1.368
Value_Of_Output	-1.8154	0.552	-3.290	0.001	-2.897	-0.734
Cost_of_Prod	1.6849	0.489	3.447	0.001	0.727	2.643
PBIDT	-1.2197	0.257	-4.745	0.000	-1.724	-0.716
PBIT	0.9219	0.251	3.670	0.000	0.430	1.414
Book_Value_Unit_Curr	-2.0100	0.544	-3.693	0.000	-3.077	-0.943
Book_Value_Adj_Unit_Curr	-1.5899	0.539	-2.950	0.003	-2.646	-0.533
ROG_Net_Worth_perc	-0.5607	0.149	-3.768	0.000	-0.852	-0.269
ROG_Capital_Employed_perc	0.4830	0.132	3.672	0.000	0.225	0.741
Curr_Ratio_Latest	-1.0811	0.163	-6.639	0.000	-1.400	-0.762
Interest_Cover_Ratio_Latest	-0.7117	0.167	-4.265	0.000	-1.039	-0.385

Table 10Logit Regression results

1.6.2 Inference:

➢ The sign of a regression coefficient tells you whether there is a positive or negative correlation between each independent variable the dependent variable. A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease.

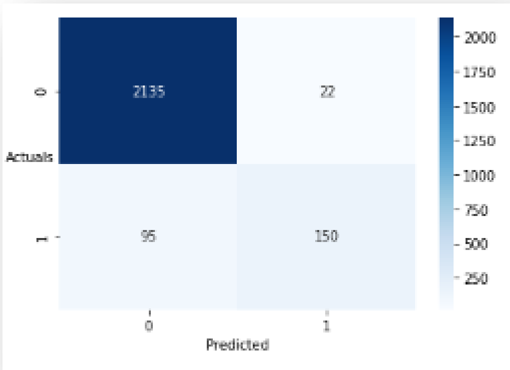
➢ Gross_Block, Curr_Liab_and_Prov, Total_Assets_to_Liab, Cost_of_Prod, ROG_Capital_Employed_perc has positive coefficients. When these features increase Credit Score also increases.

➢ Other features have negative coefficients. When these features increases then Credit score is decreases.

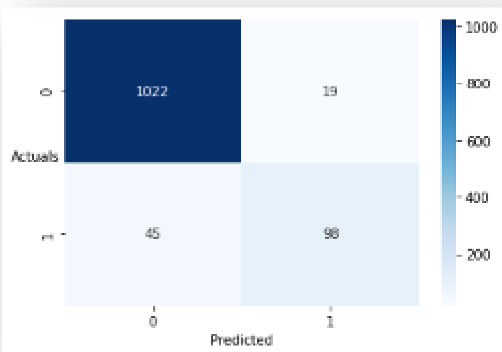
➢ The parameter estimates table summarizes the effect of each predictor. The ratio of the coefficient to its standard error, squared, equals the Wald statistic. If the significance level of the Wald statistic is small (less than 0.05) then the parameter is useful to the model. The predictors and coefficient values shown in the last step are used by the procedure to make predictions.

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

Confusion matrix on the training and test data:-



	precision	recall	f1-score	support
0.0	0.96	0.99	0.97	2157
1.0	0.87	0.61	0.72	245
accuracy			0.95	2402
macro avg	0.91	0.80	0.85	2402
weighted avg	0.95	0.95	0.95	2402



	precision	recall	f1-score	support
0.0	0.96	0.98	0.97	1041
1.0	0.84	0.69	0.75	143
accuracy			0.95	1184
macro avg	0.90	0.83	0.86	1184
weighted avg	0.94	0.95	0.94	1184

Figure 9 Confusion matrix on the training and test data

Table 11 Classification matrix report for training and test data

Inference

Training data:

True Negative : 2135 False Positive : 22

False Negative : 95 True Positive : 150

Test data:

True Negative : 1022 False Positive : 19

False Negative : 45 True Positive : 98

Train Data:

➢ Accuracy: 95%

➢ precision : 87%

➢ recall : 61%

➢ f1 :72%

Test Data:

➢ Accuracy: 95%

➢ precision: 84%

➢ recall : 69%

➢ f1 : 75%

Interpretation:-

Credit report analysis provides information on the credit worthiness of a potential customer The model with selected features will predict a relatively high probability of default. Next step is to integrate with classification model where defaulters further classified into “very high risk”, “high risk”, “medium risk”, “low risk”, etc. Later embed these models in Web and Database Integration.