FINANCE AND RISK ANALYTICS

Submitted by:

Pavithra Doraiswamy

Problem Statement

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

Data that is available includes information from the financial statement of the companies for the previous year (2015). Also, information about the Networth of the company in the following year (2016) is provided which can be used to drive the labeled field.

Explanation of data fields available in Data Dictionary, 'Data Dictionary.xlsx'

Hints:

Dependent variable - We need to create a default variable that should take the value of 1 when Net worth next year is negative & 0 when Net worth next year is positive.

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

1.1 Outlier Treatment

The given dataset shape is (**3586,67).**
There was 3586 number of rows and 67 columns.
The data type of the 67 columns are:

a. float64-63 columns
b. int64 -3 columns
c. object -1 column

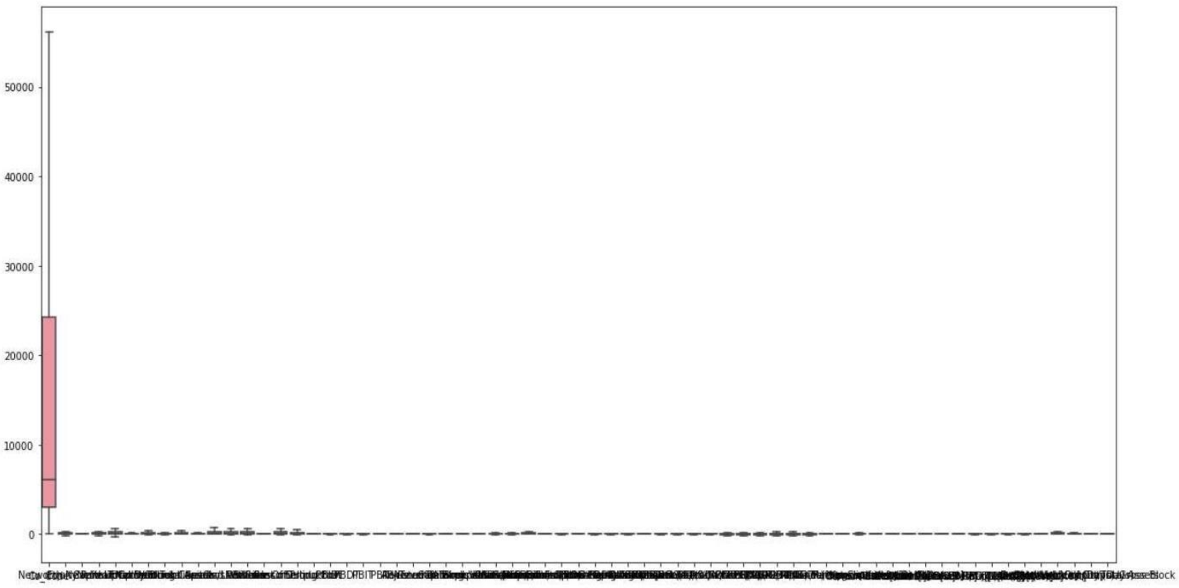
There were no duplicate rows in the given dataset.
There were outliers present in the given dataset.

Outlier treatment was carried out on the dataset by considering the following:

1. Arranging data in ascending order.
2. Calculate the 25^{th} percentile value Q1.
3. Calculate the 75^{th} percentile value Q3.
4. Find **IQR** which is (Q3 - Q1)
5. Find the lower Range using Q1 -(1.5 * **IQR**)

6. Find the upper Range using Q3 + (1.5 * **IQR**)
Using IQR capping of outliers was done.

After outlier treatment the boxplot is given below:



From the above boxplot there is no outliers.

1.2 Missing Value Treatment

From the below figures there were total of 118 null values in the given dataset which has a total of 3586 no of rows which is 3.2 percent.

```
pd.set_option('max_rows', None)
data.isnull().sum()
```

```
Co_Code                                  0
Co_Name                                  0
Networth Next Year                       0
Equity Paid Up                           0
Networth                                 0
Capital Employed                         0
Total Debt                               0
Gross Block                              0
Net Working Capital                      0
Current Assets                           0
Current Liabilities and Provisions       0
Total Assets/Liabilities                 0
Gross Sales                              0
Net Sales                                0
Other Income                             0
Value Of Output                          0
Cost of Production                       0
Selling Cost                             0
PBIDT                                    0
PBDT                                     0
PBIT                                     0
PBT                                      0
PAT                                      0
Adjusted PAT                             0
CP                                       0
Revenue earnings in forex                0
Revenue expenses in forex                0
Capital expenses in forex                0
Book Value (Unit Curr)                   0
Book Value (Adj.) (Unit Curr)            4
```

```
Cash Flow From Operating Activities        0
Cash Flow From Investing Activities        0
Cash Flow From Financing Activities        0
ROG-Net Worth (%)                          0
ROG-Capital Employed (%)                   0
ROG-Gross Block (%)                        0
ROG-Gross Sales (%)                        0
ROG-Net Sales (%)                          0
ROG-Cost of Production (%)                 0
ROG-Total Assets (%)                       0
ROG-PBIDT (%)                              0
ROG-PBDT (%)                               0
ROG-PBIT (%)                               0
ROG-PBT (%)                                0
ROG-PAT (%)                                0
ROG-CP (%)                                 0
ROG-Revenue earnings in forex (%)          0
ROG-Revenue expenses in forex (%)          0
ROG-Market Capitalisation (%)              0
Current Ratio[Latest]                      1
Fixed Assets Ratio[Latest]                 1
Inventory Ratio[Latest]                    1
Debtors Ratio[Latest]                      1
Total Asset Turnover Ratio[Latest]         1
Interest Cover Ratio[Latest]               1
PBIDTM (%)[Latest]                         1
PBITM (%)[Latest]                          1
PBDTM (%)[Latest]                          1
CPM (%)[Latest]                            1
APATM (%)[Latest]                          1
Debtors Velocity (Days)                    0
Creditors Velocity (Days)                  0
Inventory Velocity (Days)                103
Value of Output/Total Assets               0
Value of Output/Gross Block                0
dtype: int64
```

```
data.isnull().sum().sum()
```

```
118
```

The column Inventory Velocity (Days) which had more no of null values is dropped.
After dropping null values, the shape of the dataset is as follows:

(3581,66).

1.3 Transform Target variable into 0 and 1

As already mentioned, we need to create a default variable that should take the value of 1 when Net worth next year is negative & 0 when Net worth next year is positive. So, a target variable with the mentioned condition is created.

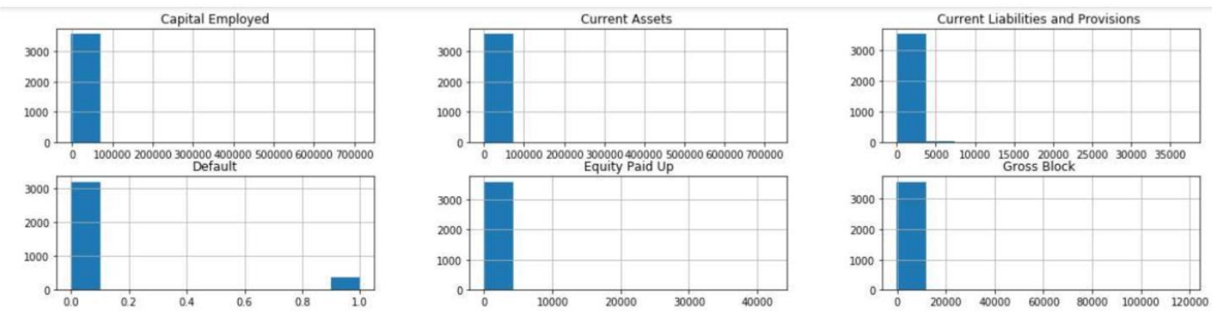
```
0.0    3194
1.0     387
```

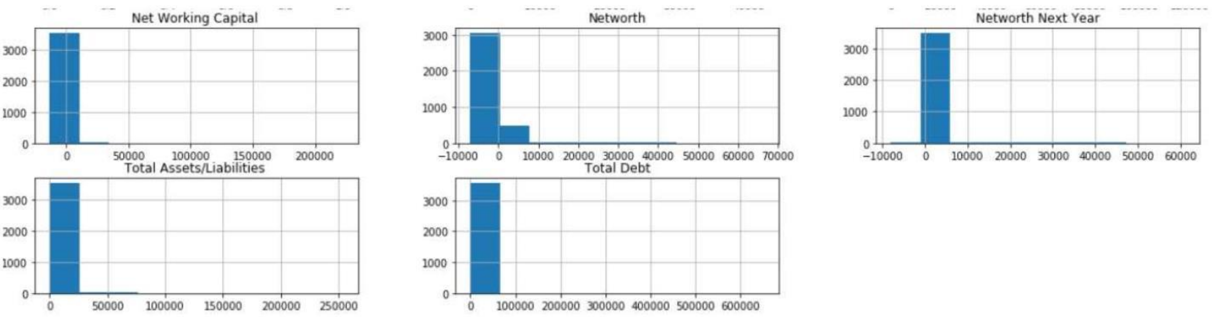
The value counts of 0s and 1s in the target column default is given above which is 3194 and 387 respectively.

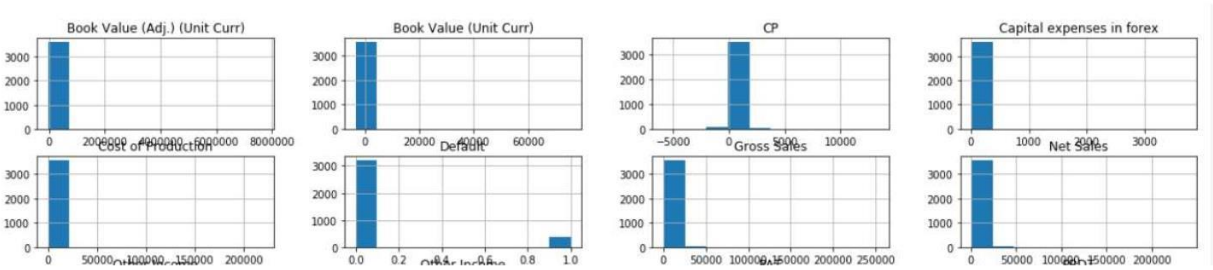
1.4 Univariate & Bivariate analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building)

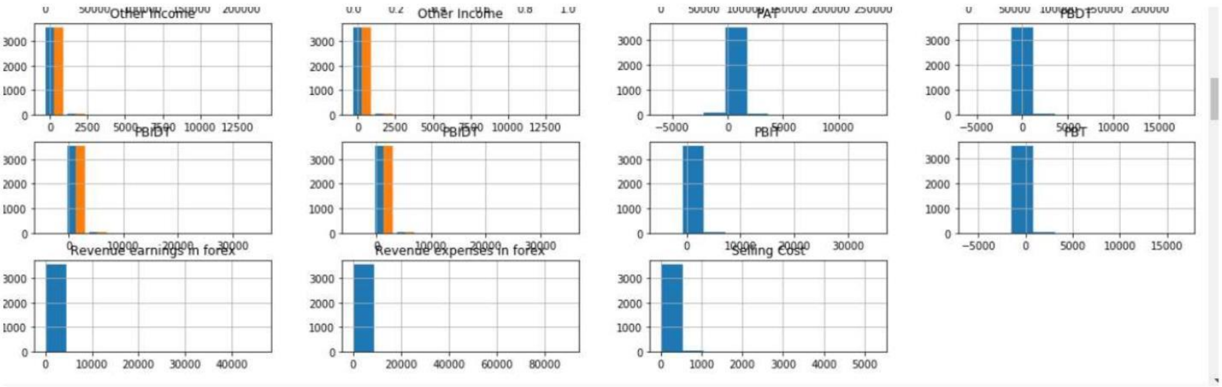
Histogram:

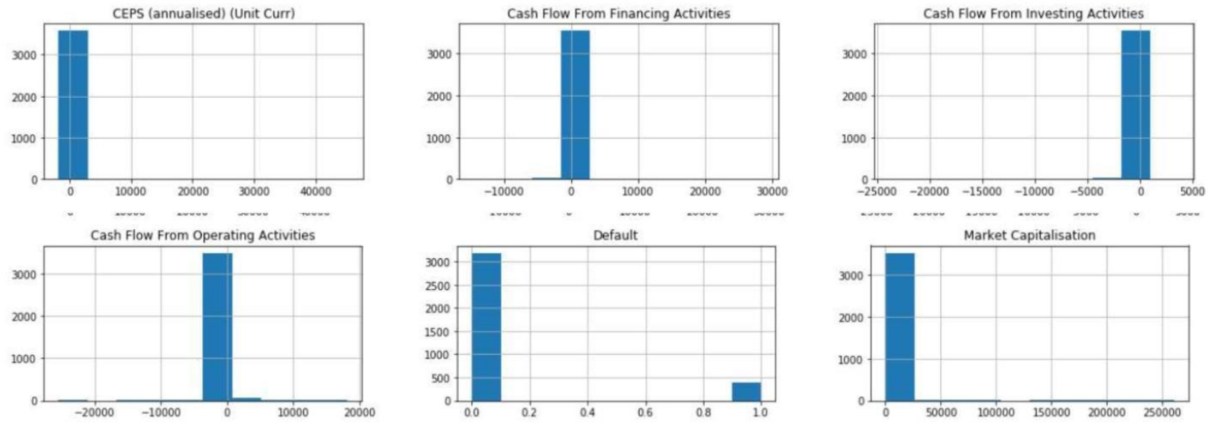
Below Histogram on all Variables is done. From the plot we can say that the data is not normally distributed.

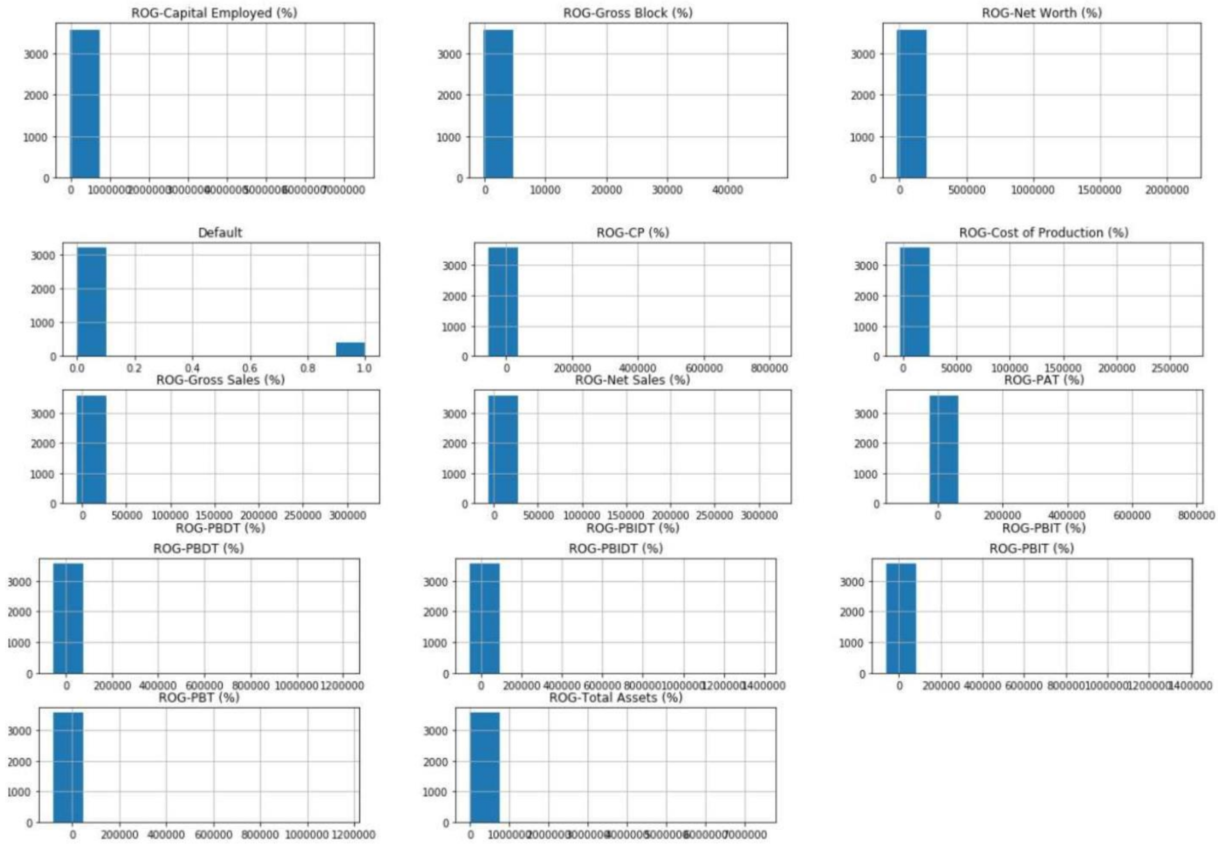


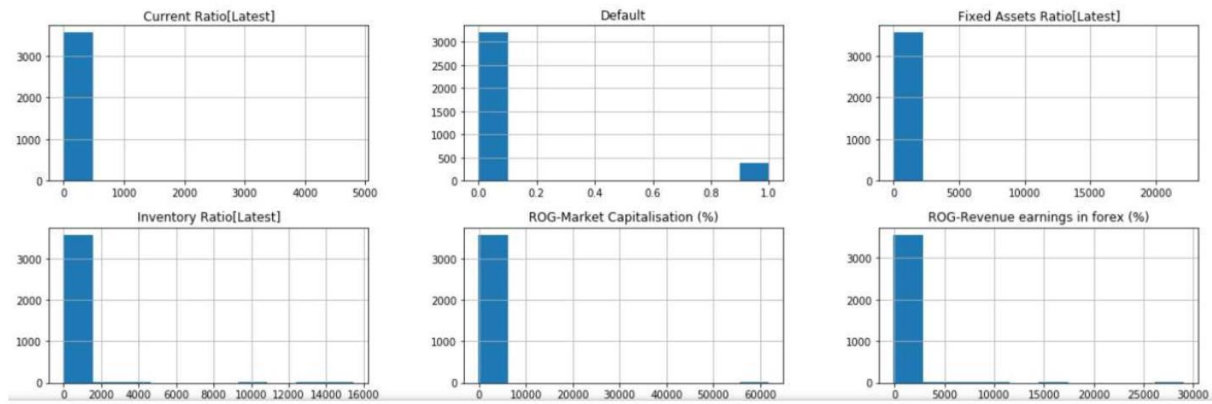

Net Working Capital
Networth
Networth Next Year
Total Assets/Liabilities
Total Debt

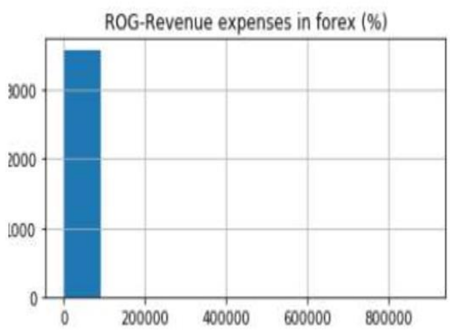

Book Value (Adj.) (Unit Curr)
Book Value (Unit Curr)
CP
Capital expenses in forex
Cost of Production
Default
Gross Sales
Net Sales

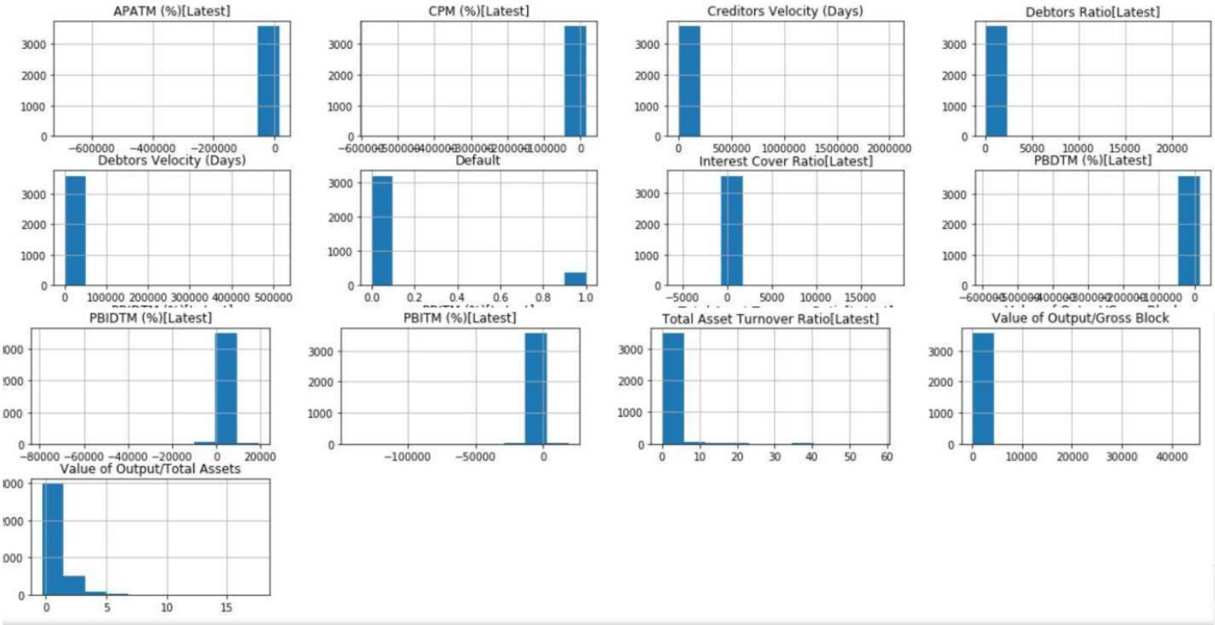

Other Income
Other Income
PAT
PBDT
PBIDT
PBIDT
PBIT
PBT
Revenue earnings in forex
Revenue expenses in forex
Selling Cost


CEPS (annualised) (Unit Curr)
Cash Flow From Financing Activities
Cash Flow From Investing Activities
Cash Flow From Operating Activities
Default
Market Capitalisation


ROG-Capital Employed (%)
ROG-Gross Block (%)
ROG-Net Worth (%)
Default
ROG-CP (%)
ROG-Cost of Production (%)
ROG-Gross Sales (%)
ROG-Net Sales (%)
ROG-PAT (%)
ROG-PBDT (%)
ROG-PBIDT (%)
ROG-PBIT (%)
ROG-PBDT (%)
ROG-PBIDT (%)
ROG-PBIT (%)
ROG-PBT (%)
ROG-Total Assets (%)

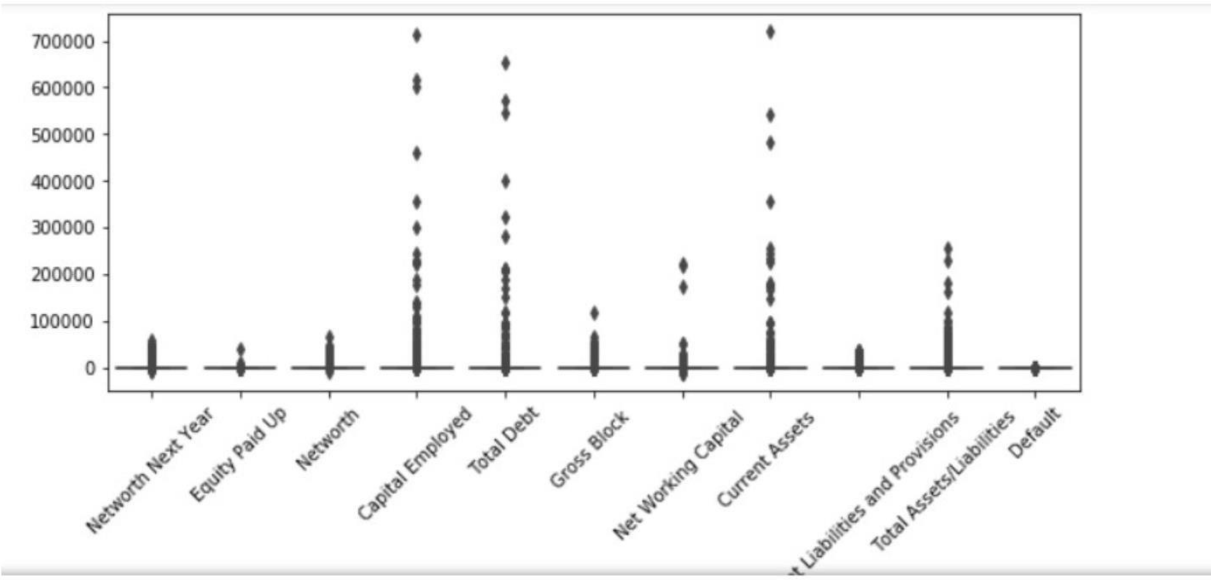

Current Ratio[Latest]
Default
Fixed Assets Ratio[Latest]
Inventory Ratio[Latest]
ROG-Market Capitalisation (%)
ROG-Revenue earnings in forex (%)

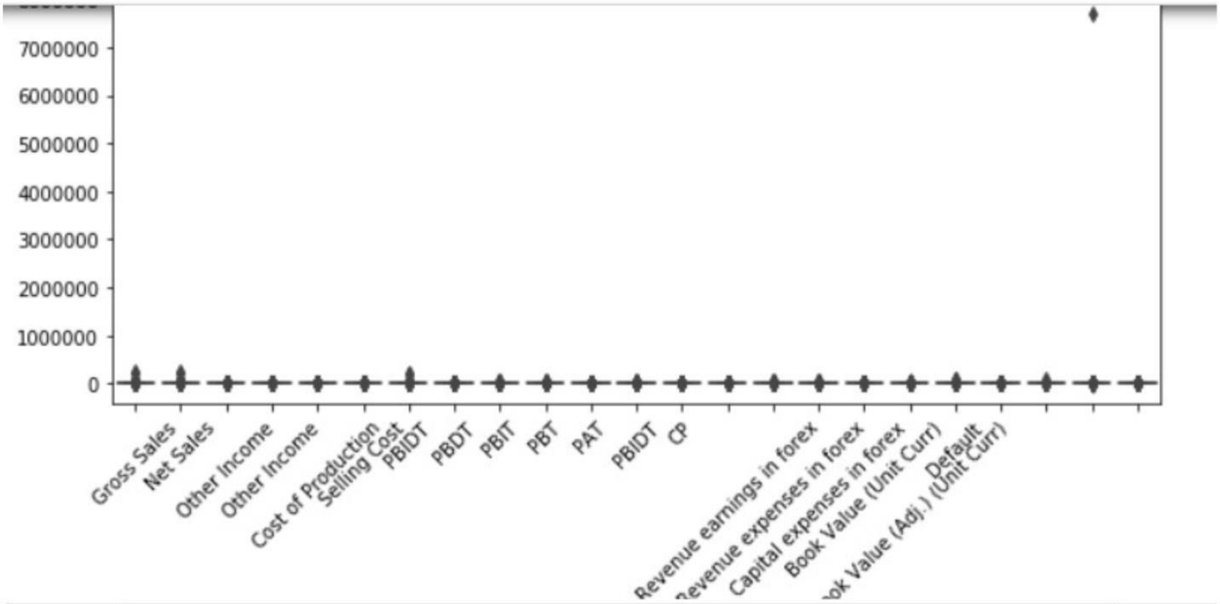

ROG-Revenue expenses in forex (%)



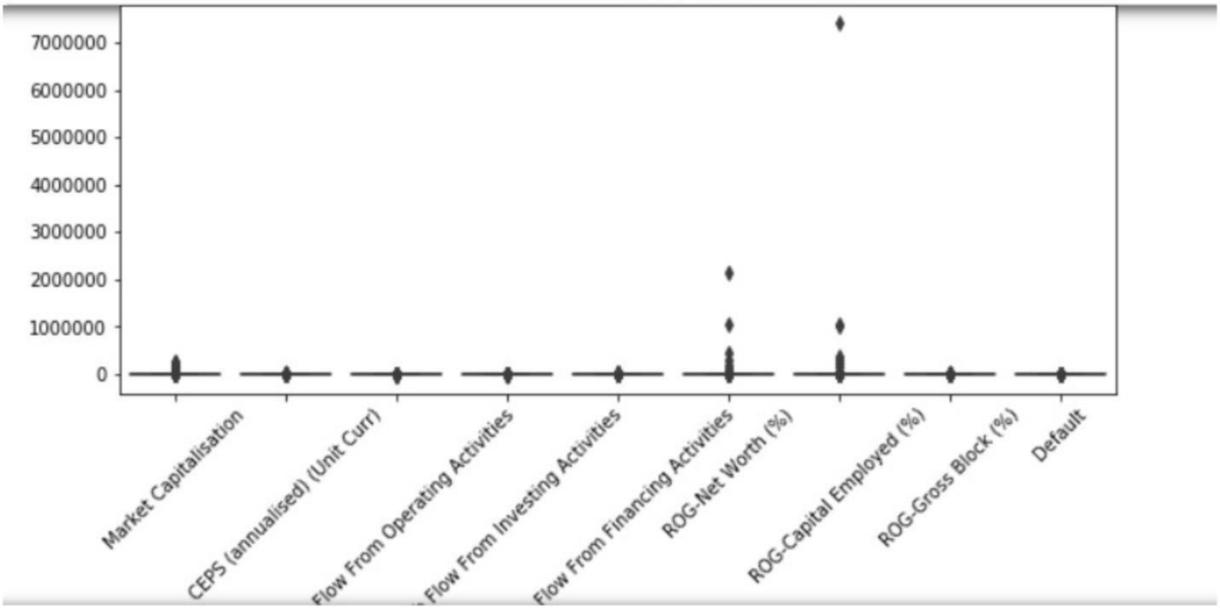
Boxplots:

Box plots to understand the distribution of data. We can see from below plots that outlier is present in all the variables. We can also assess the distribution of data using kurtosis and skew.

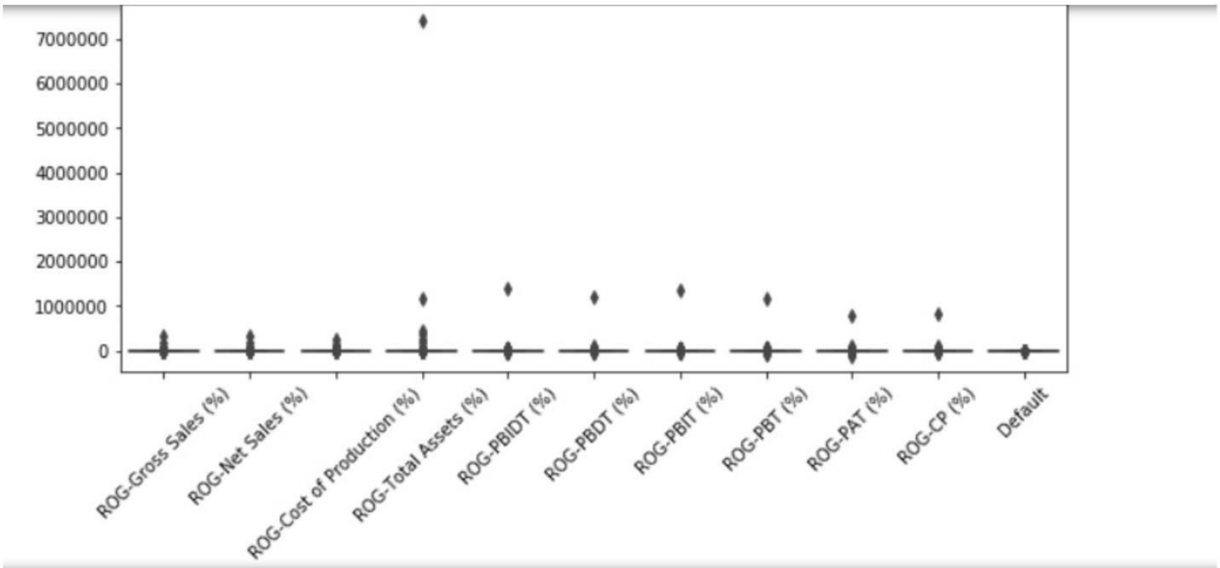


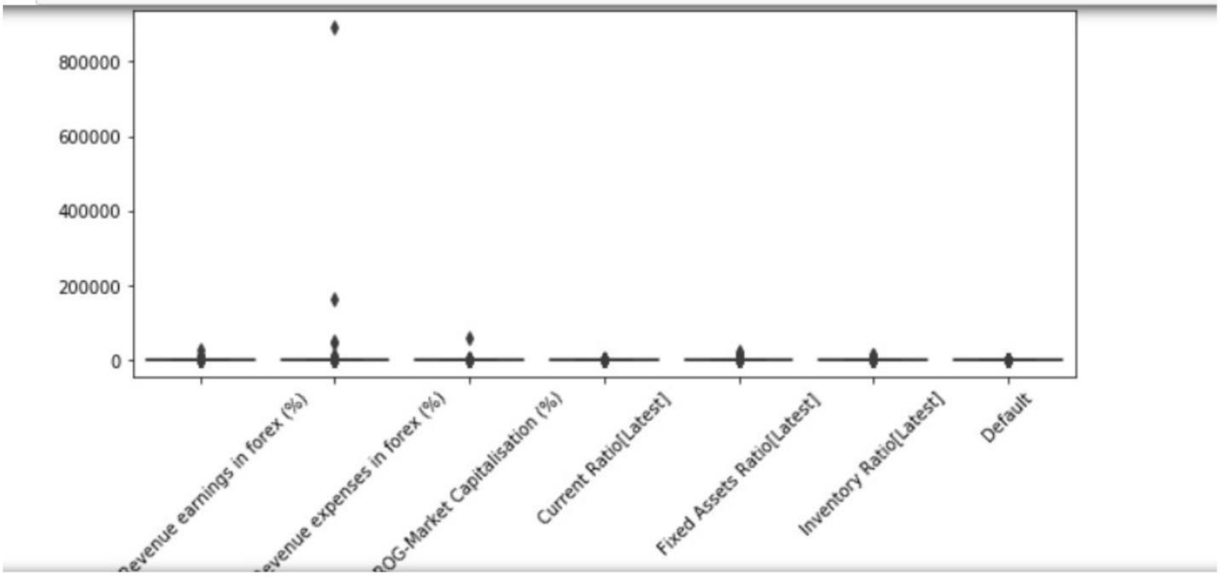


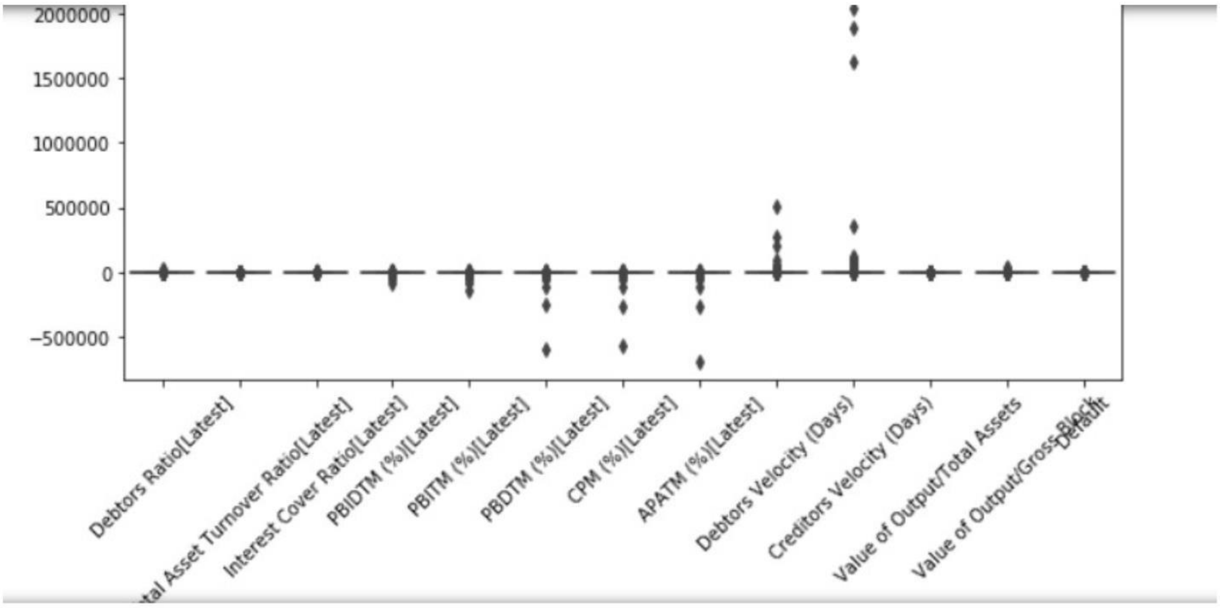
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
Gross Sales
Net Sales
Other Income
Other Income
Cost of Production
Selling Cost
PBIDT
PBDT
PBIT
PBT
PAT
PBIDT
CP
Revenue earnings in forex
Revenue expenses in forex
Capital expenses in forex
Book Value (Unit Curr)
Default
Book Value (Adj.) (Unit Curr)



7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
Market Capitalisation
CEPS (annualised) (Unit Curr)
Flow From Operating Activities
Flow From Investing Activities
Flow From Financing Activities
ROG-Net Worth (%)
ROG-Capital Employed (%)
ROG-Gross Block (%)
Default

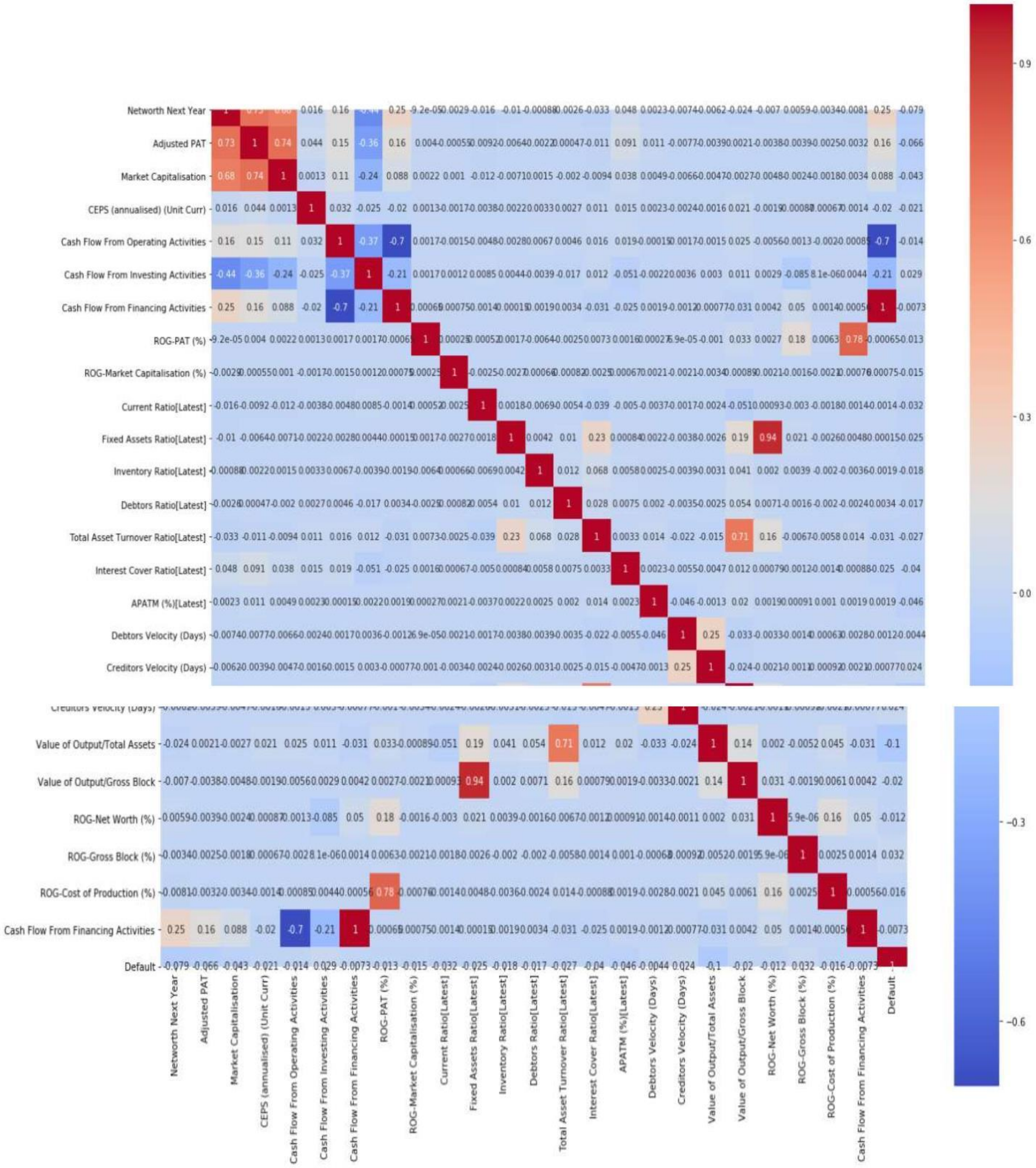

7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
ROG-Gross Sales (%)
ROG-Net Sales (%)
ROG-Cost of Production (%)
ROG-Total Assets (%)
ROG-PBIDT (%)
ROG-PBDT (%)
ROG-PBIT (%)
ROG-PBT (%)
ROG-PAT (%)
ROG-CP (%)
Default


800000
600000
400000
200000
0
ROG-Market Capitalisation (%)
Current Ratio[Latest]
Fixed Assets Ratio[Latest]
Inventory Ratio[Latest]
Default



Heatmap:

From the heatmap below, we can see many variables are highly correlated with each other. We can use VIF (variable inflation factors) to assess if there is multicollinearity between independent variables.



1.5 Train Test Split

The train_test_split is for splitting a single dataset for two different purposes: training and testing. The testing subset is for building your model. The testing subset is for using the model on unknown data to evaluate the performance of the model.

So, let us divide the data into training and test dataset in the ratio 67:33. There are a total of 2399 records in Train and 1182 records in Test dataset.

```
Train.shape
```

```
(2399, 14)
```

```
Test.shape
```

```
(1182, 14)
```

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

Logistic Regression:

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables.

The Below is the VIF value for different independent variables.

:

	VIF	variable
0	1.289656e+00	Co_Code
1	2.574517e+01	Networth Next Year
2	1.489876e+00	Equity Paid Up
3	6.287015e+01	Networth
4	2.608702e+03	Capital Employed
5	1.273709e+03	Total Debt
6	1.916345e+01	Gross Block
7	4.969258e+01	Net Working Capital
8	1.835601e+02	Current Assets
9	2.388149e+01	Current Liabilities and Provisions
10	1.256741e+02	Total Assets/Liabilities
11	2.299312e+03	Gross Sales
12	1.251686e+04	Net Sales
13	8.377212e+00	Other Income
14	1.049763e+04	Value Of Output
15	1.066718e+03	Cost of Production
16	2.856506e+00	Selling Cost
17	1.117136e+01	Adjusted PAT
18	2.658947e+00	Revenue earnings in forex
19	2.551364e+01	Revenue expenses in forex
20	7.971651e+00	Capital expenses in forex

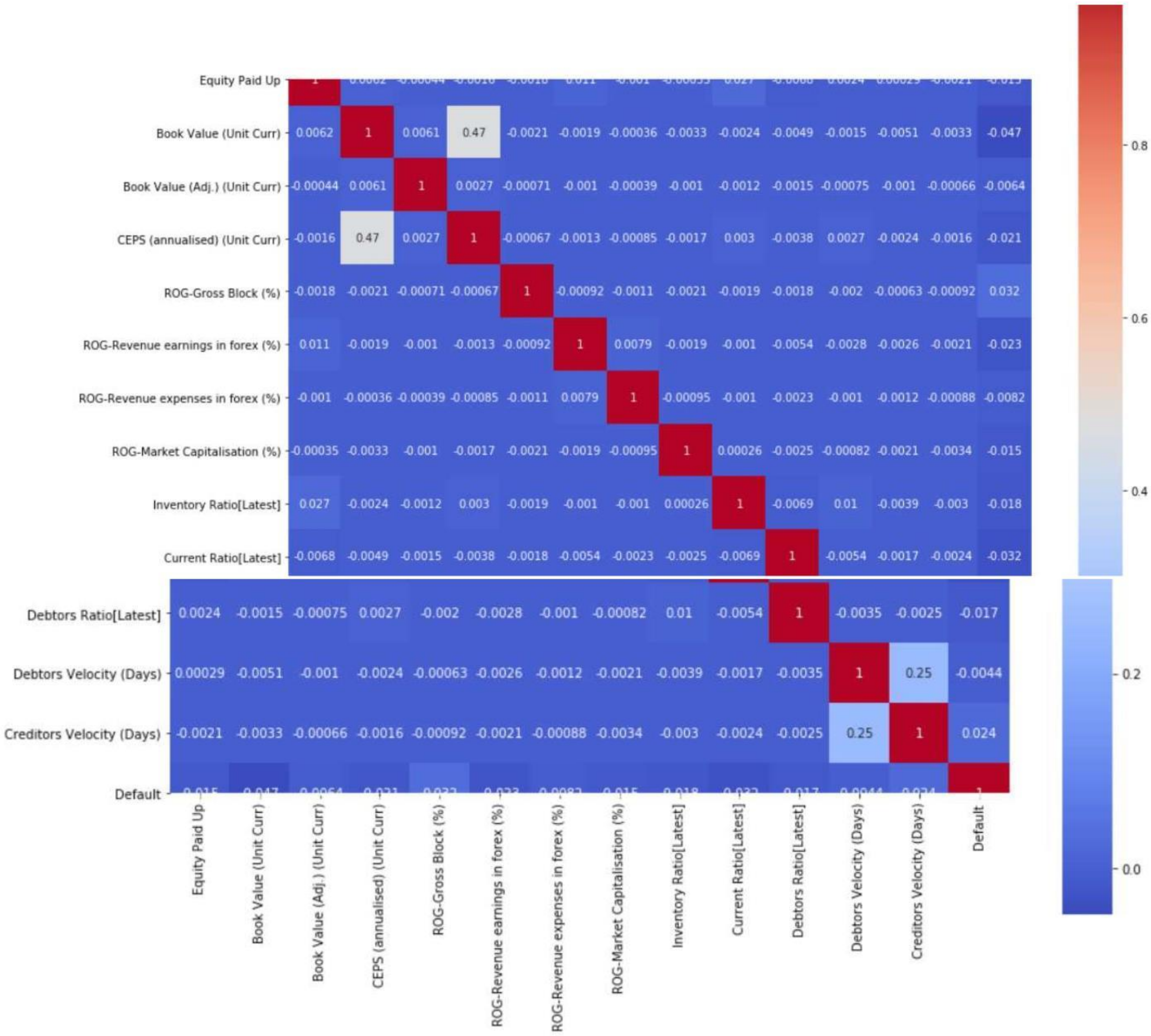
21	1.324715e+00	Book Value (Unit Curr)
22	1.005903e+00	Book Value (Adj.) (Unit Curr)
23	3.945892e+00	Market Capitalisation
24	1.341744e+00	CEPS (annualised) (Unit Curr)
25	1.729468e+01	Cash Flow From Operating Activities
26	8.846726e+00	Cash Flow From Investing Activities
27	1.484555e+01	Cash Flow From Financing Activities
28	2.192063e+01	ROG-Net Worth (%)
29	3.130189e+03	ROG-Capital Employed (%)
30	1.003600e+00	ROG-Gross Block (%)
31	2.060871e+06	ROG-Gross Sales (%)
32	2.060850e+06	ROG-Net Sales (%)
33	3.389346e+00	ROG-Cost of Production (%)
34	3.429078e+03	ROG-Total Assets (%)
35	3.961328e+02	ROG-PBIDT (%)
36	4.276766e+02	ROG-PBDT (%)
37	2.486160e+02	ROG-PBIT (%)
38	8.028446e+01	ROG-PBT (%)
39	2.197622e+01	ROG-PAT (%)
40	9.660137e+01	ROG-CP (%)
41	1.149727e+00	ROG-Revenue earnings in forex (%)
42	1.001445e+00	ROG-Revenue expenses in forex (%)
43	1.002520e+00	ROG-Market Capitalisation (%)

44	1.012902e+00	Current Ratio[Latest]
45	9.250245e+00	Fixed Assets Ratio[Latest]
46	1.241912e+00	Inventory Ratio[Latest]
47	1.010024e+00	Debtors Ratio[Latest]
48	2.577469e+00	Total Asset Turnover Ratio[Latest]
49	2.083430e+00	Interest Cover Ratio[Latest]
50	1.387110e+11	PBIDTM (%)[Latest]
51	4.026104e+11	PBITM (%)[Latest]
52	5.635599e+03	PBDTM (%)[Latest]
53	4.905882e+12	CPM (%)[Latest]
54	6.726810e+12	APATM (%)[Latest]
55	1.518385e+00	Debtors Velocity (Days)
56	1.099604e+00	Creditors Velocity (Days)
57	2.781496e+00	Value of Output/Total Assets
58	8.971307e+00	Value of Output/Gross Block

For model building we can drop columns having VIF>2 to reduce multicollinearity.
So, considering only below mentioned columns for model building.

	VIF	variable
0	1.001824	Equity Paid Up
1	1.287746	Book Value (Unit Curr)
2	1.000062	Book Value (Adj.) (Unit Curr)
3	1.287241	CEPS (annualised) (Unit Curr)
4	1.001952	ROG-Gross Block (%)
5	1.000376	ROG-Revenue earnings in forex (%)
6	1.000092	ROG-Revenue expenses in forex (%)
7	1.000122	ROG-Market Capitalisation (%)
8	1.001524	Inventory Ratio[Latest]
9	1.000138	Current Ratio[Latest]
10	1.000395	Debtors Ratio[Latest]
11	1.068841	Debtors Velocity (Days)
12	1.069915	Creditors Velocity (Days)
13	1.003743	Default

Below is the heatmap for to see the correlation of the parameters with less VIF factor.



With this predictor the logistic regression model is built.

```
import statsmodels.formula.api as SM
logitmodel = SM.logit(formula=f1,data=Train).fit()
logitmodel.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.180894
         Iterations 13
```

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

Model Performance

1. Create the formula variable and load for all the independent variables as:

   f1='Default ~ EquityPaidUp + BookValurUnitCurr + BookValueAdjUnitCurr + CEPSannualisedUnitCurr + ROGGrossBlockPerc + ROGRevenueearningsinforexper + ROGRevenueexpensesinforexper + ROGMarketCapitalisationper + CurrentRatioLatest + DebtorsRatioLatest+ DebtorsVelocityDays + CreditorsVelocityDays + InventoryRatioLatest '

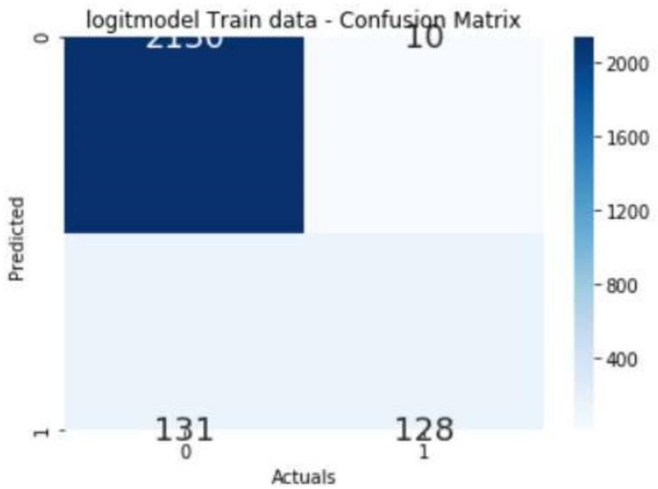
2. From the results we can see the BookValueAdjUnitCurr, ROGRevenueexpensesinforexper variables are not statistically significant as p value is higher than 0.05. Let us use the variable selection method. Drop the variable with the highest p-value (least significant variable) in the first iteration of the model and run the Logit model once again.
3. The accuracy scores, Precision and Recall values, F Values for two models are almost the same.

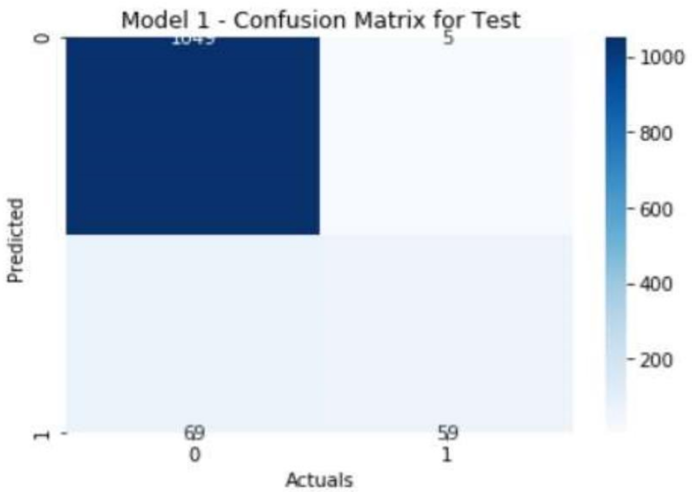
Model 1:

Dep. Variable:	Default	**No. Observations:**	2399
Model:	Logit	**Df Residuals:**	2385
Method:	MLE	**Df Model:**	13
Date:	Sun, 26 Dec 2021	**Pseudo R-squ.:**	0.4714
Time:	17:18:45	**Log-Likelihood:**	-433.96
converged:	True	**LL-Null:**	-821.02
Covariance Type:	nonrobust	**LLR p-value:**	4.848e-157

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.2273	0.137	-1.656	0.098	-0.496	0.042
EquityPaidUp	-0.0008	0.001	-0.871	0.384	-0.002	0.001
BookValurUnitCurr	-0.0670	0.006	-11.465	0.000	-0.078	-0.056
BookValueAdjUnitCurr	1.068e-06	4.57e-05	0.023	0.981	-8.85e-05	9.07e-05
CEPSannualisedUnitCurr	0.0559	0.010	5.614	0.000	0.036	0.075
ROGGrossBlockPerc	-0.0065	0.003	-2.112	0.035	-0.012	-0.000
ROGRevenueearningsinforexper	-0.0040	0.002	-2.038	0.042	-0.008	-0.000
ROGRevenueexpensesinforexper	-0.0002	0.000	-0.420	0.674	-0.001	0.001
ROGMarketCapitalisationper	-0.0019	0.001	-1.759	0.079	-0.004	0.000
CurrentRatioLatest	-0.6951	0.102	-6.842	0.000	-0.894	-0.496
DebtorsRatioLatest	-0.0003	0.001	-0.326	0.744	-0.002	0.001
DebtorsVelocityDays	-1.512e-05	2.96e-05	-0.510	0.610	-7.32e-05	4.3e-05
CreditorsVelocityDays	5.895e-06	4.46e-06	1.322	0.186	-2.84e-06	1.46e-05
InventoryRatioLatest	-0.0002	0.001	-0.317	0.751	-0.001	0.001

Heat Map:


logitmodel Train data - Confusion Matrix
10
131
128
0
1
Predicted
0
1
Actuals
2000
1600
1200
800
400


Model 1 - Confusion Matrix for Test
5
69
59
0
1
Predicted
0
1
Actuals
1000
800
600
400
200

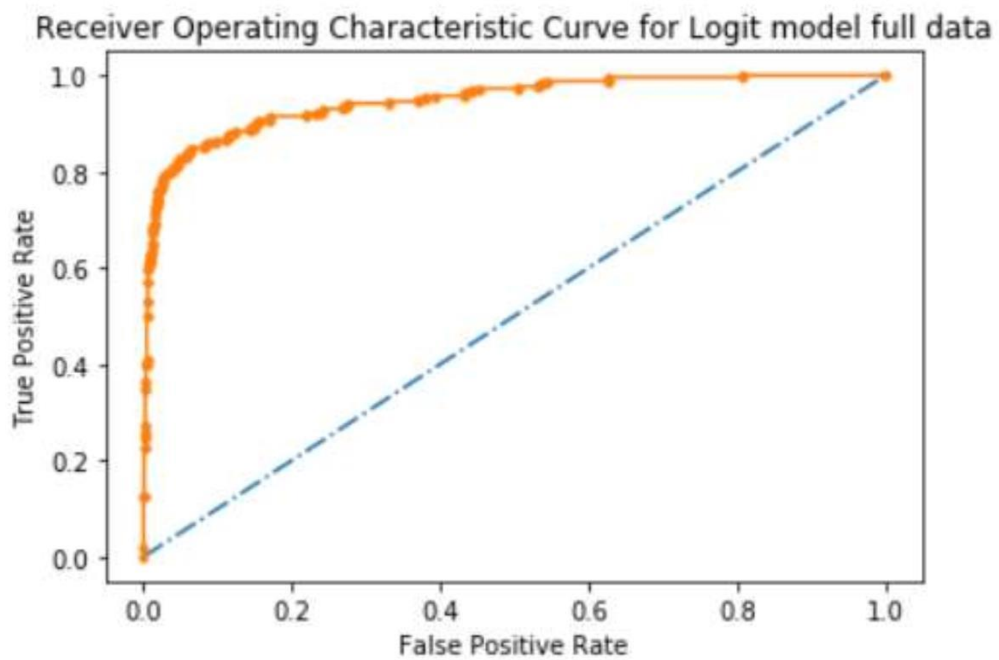
Report for Model 1:

Classification Report for Logit model _1 Train data

	precision	recall	f1-score	support
0.0	0.94	1.00	0.97	2140
1.0	0.93	0.49	0.64	259
accuracy			0.94	2399
macro avg	0.93	0.74	0.81	2399
weighted avg	0.94	0.94	0.93	2399

ROC Model 1 full Data:

AUC: 0.9489

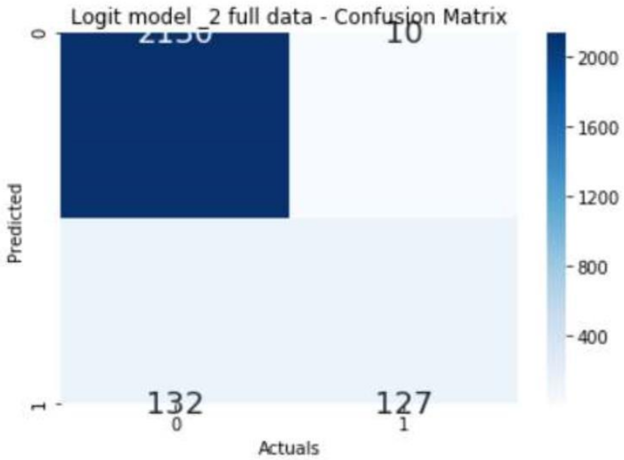


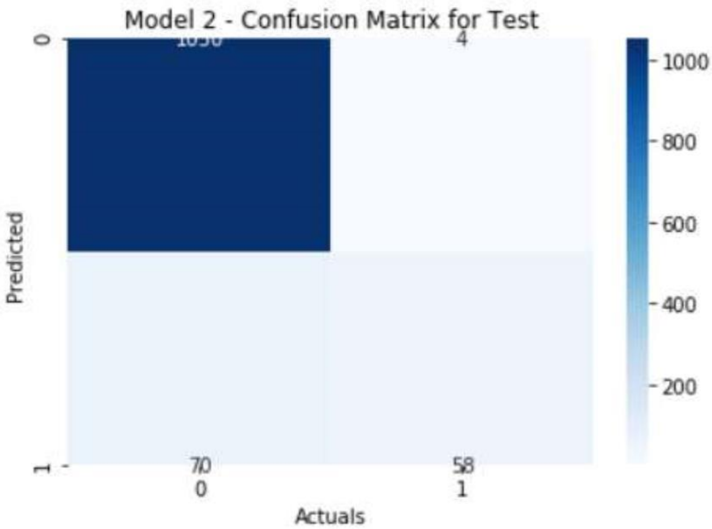
Model 2:

Dep. Variable:	Default	**No. Observations:**	2399
Model:	Logit	**Df Residuals:**	2388
Method:	MLE	**Df Model:**	10
Date:	Sun, 26 Dec 2021	**Pseudo R-squ.:**	0.4704
Time:	18:48:54	**Log-Likelihood:**	-434.80
converged:	True	**LL-Null:**	-821.02
Covariance Type:	nonrobust	**LLR p-value:**	1.742e-159

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.2263	0.138	-1.641	0.101	-0.497	0.044
EquityPaidUp	-0.0007	0.001	-0.856	0.392	-0.002	0.001
BookValurUnitCurr	-0.0670	0.006	-11.475	0.000	-0.078	-0.056
CEPSannualisedUnitCurr	0.0561	0.010	5.651	0.000	0.037	0.076
ROGGrossBlockPerc	-0.0065	0.003	-2.137	0.033	-0.013	-0.001
ROGRevenueearningsinforexper	-0.0040	0.002	-2.032	0.042	-0.008	-0.000
ROGRevenueexpensesinforexper	-0.0002	0.000	-0.421	0.674	-0.001	0.001
ROGMarketCapitalisationper	-0.0019	0.001	-1.802	0.072	-0.004	0.000
CurrentRatioLatest	-0.6936	0.102	-6.829	0.000	-0.893	-0.495
DebtorsRatioLatest	-0.0003	0.001	-0.326	0.744	-0.002	0.001
InventoryRatioLatest	-0.0002	0.001	-0.319	0.750	-0.001	0.001

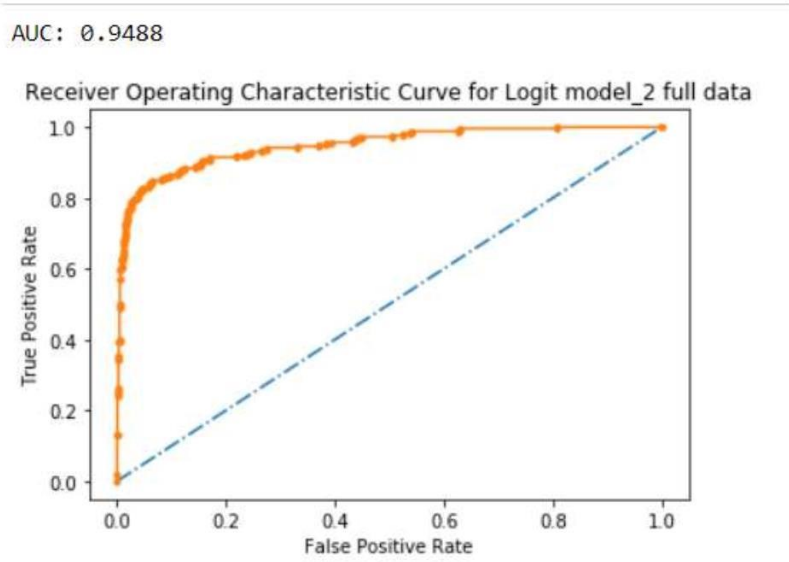
Heat Map:





Classification Report for Logit model _2 Train data

	precision	recall	f1-score	support
0.0	0.94	1.00	0.97	2140
1.0	0.93	0.49	0.64	259
accuracy			0.94	2399
macro avg	0.93	0.74	0.80	2399
weighted avg	0.94	0.94	0.93	2399



Comparing values:

Model 1 Pseudo R2 =0.4714
Model 1 Logit Accuracy= 0.94
Model 1 Logit Recall= 0.74
Model 1 Logit Precision= 0.93

Comparing values:

Model 1 Pseudo R2 =0.4704
Model 1 Logit Accuracy= 0.94
Model 1 Logit Recall= 0.74
Model 1 Logit Precision= 0.93

We could see that both the models perform equally same when considering different performance metrics.

Business Interpretation

Thus, we were able to predict default value for a company to assess the credit risk based on our logistic model.