**Company Data**

**Analysis**

**FRA 2**

Name: Saurabh Dharmadhikari

PGP-DSBA Online January’ 21

Table of Contents:

Problem Statement……………………...................................................................................................................................3

Data Description ............................................................................................................................................................4

Sample Data. …………………………………………………………………………………………………………………………………………….............5

Data Summary: ……………………….…………………………………………………………………………………………………………………………….7

Outliers ………………………..………………………………………………………………......…………………………………………………………........10

Missing value treatment. …………………………………….....................................................................................................13

Univariant and bivariant analysis. ..............................................................................................................................14

Scatter Plot:. …………………………………...........................................................................................................................17

Heat Map of Correlation. ............................................................................................................................................21

Train and Test Split …………………………………………………………………………........................................................................22

Linear Regression Model. …………….............................................................................................................................22

Model 1 ………………………………………………………………………………………………………………………………………………………………..24

Model 2 Using REF …………………………….………………………………………………………………………………………………………………….24

Model 3 Using SMOTE (75:25) ……………..……………………………………………………………………………………………………………….25

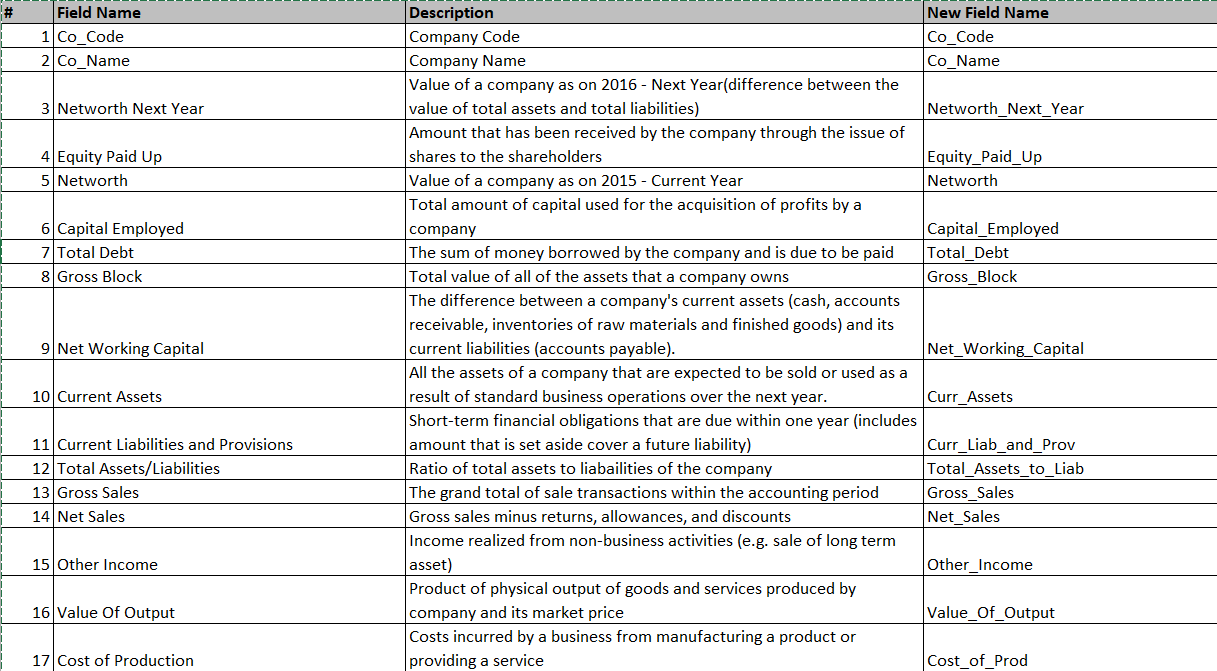
Model 4 SMOTE (65:35) ………………………………………………………………………………………………………………………………………..26

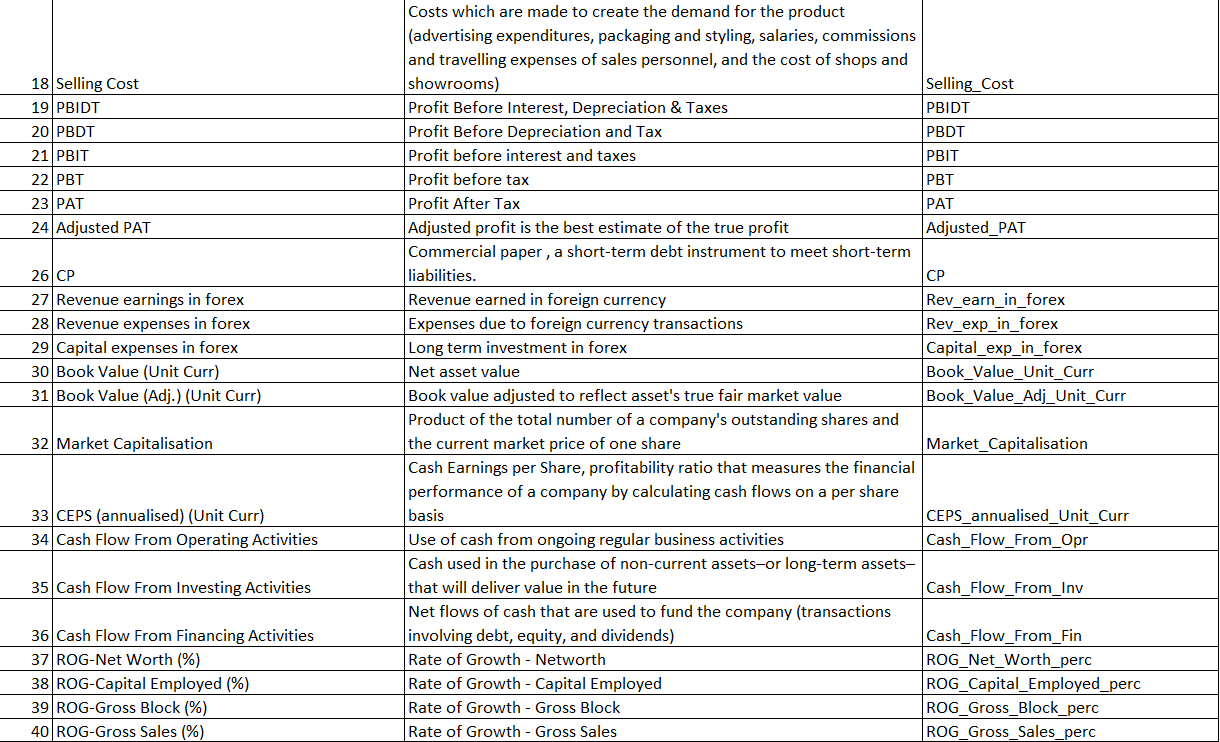
Model Validation …..………………………………………………………………………………………………………………………………………………26

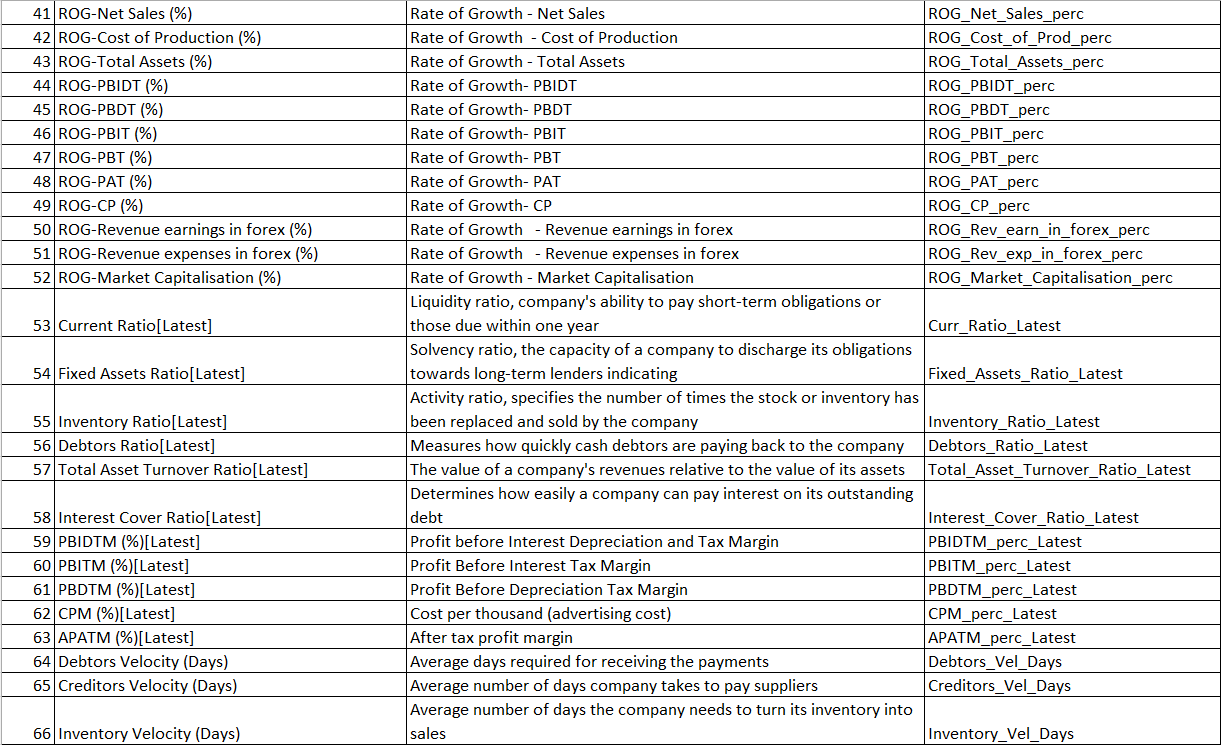
Problem Statement

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.

Data Description:

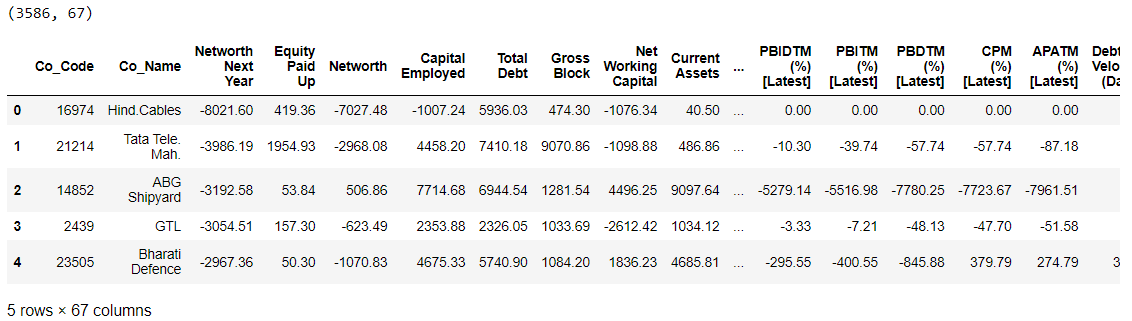








Sample Data



We have already cleaned the data.

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\_BY\_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\_PRODUCTION 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\_EARNINGS\_IN\_FOREX 3586 non-null float64

26 REVENUE\_EXPENSES\_IN\_FOREX 3586 non-null float64

27 CAPITAL\_EXPENSES\_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\_PRODUCTION\_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\_EARNINGS\_IN\_FOREX\_PERC 3586 non-null float64

49 ROG\_REVENUE\_EXPENSES\_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\_VELOCITY\_DAYS 3586 non-null int64

63 CREDITORS\_VELOCITY\_DAYS 3586 non-null int64

64 INVENTORY\_VELOCITY\_DAYS 3483 non-null float64

65 VALUE\_OF\_OUTPUT\_BY\_TOTAL\_ASSETS 3586 non-null float64

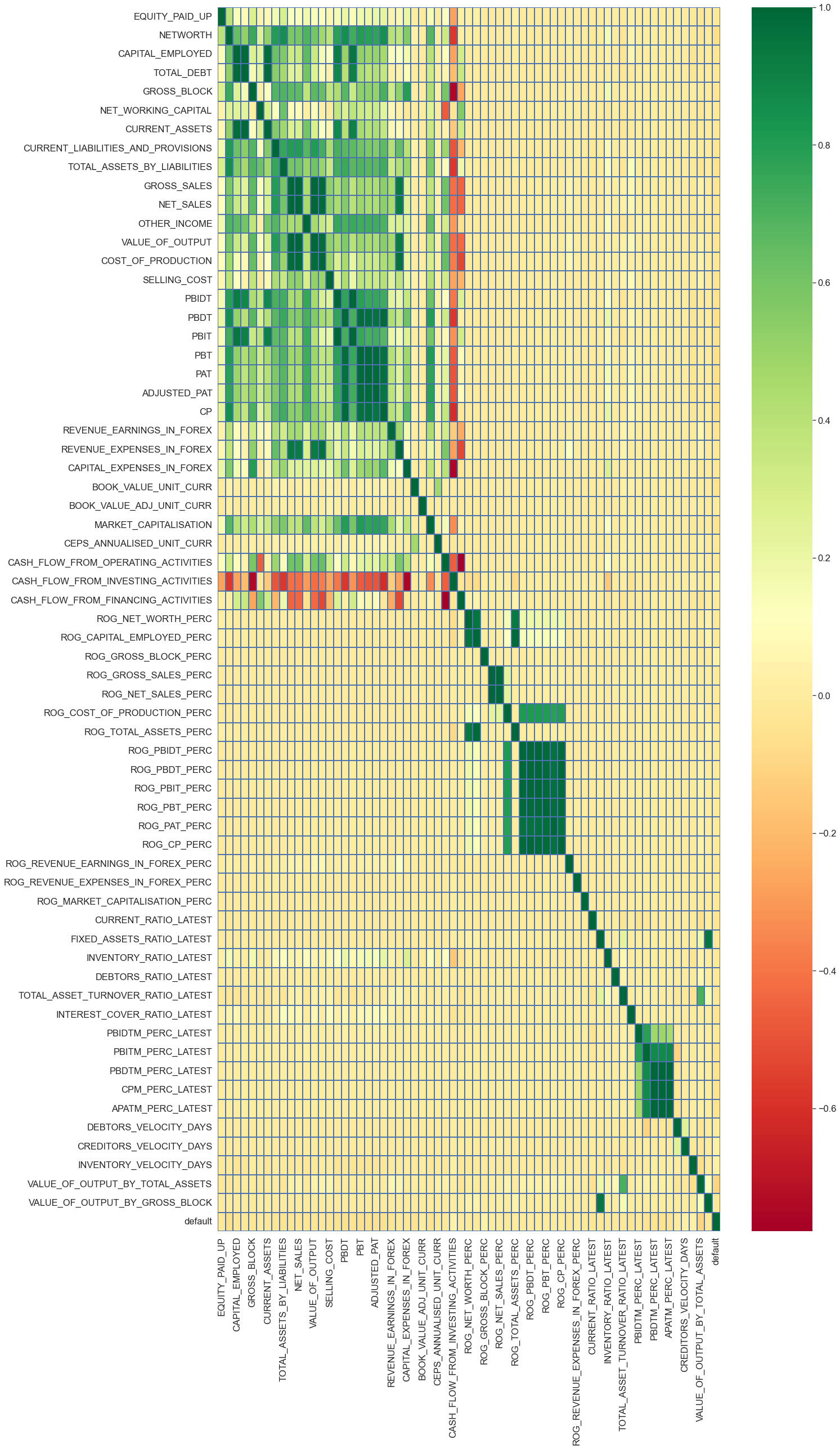
66 VALUE\_OF\_OUTPUT\_BY\_GROSS\_BLOCK 3586 non-null float64

dtypes: float64(63), int64(3), object(1)

There is only 1 categorical column all other columns are numeric in nature.

**Data Summary:**

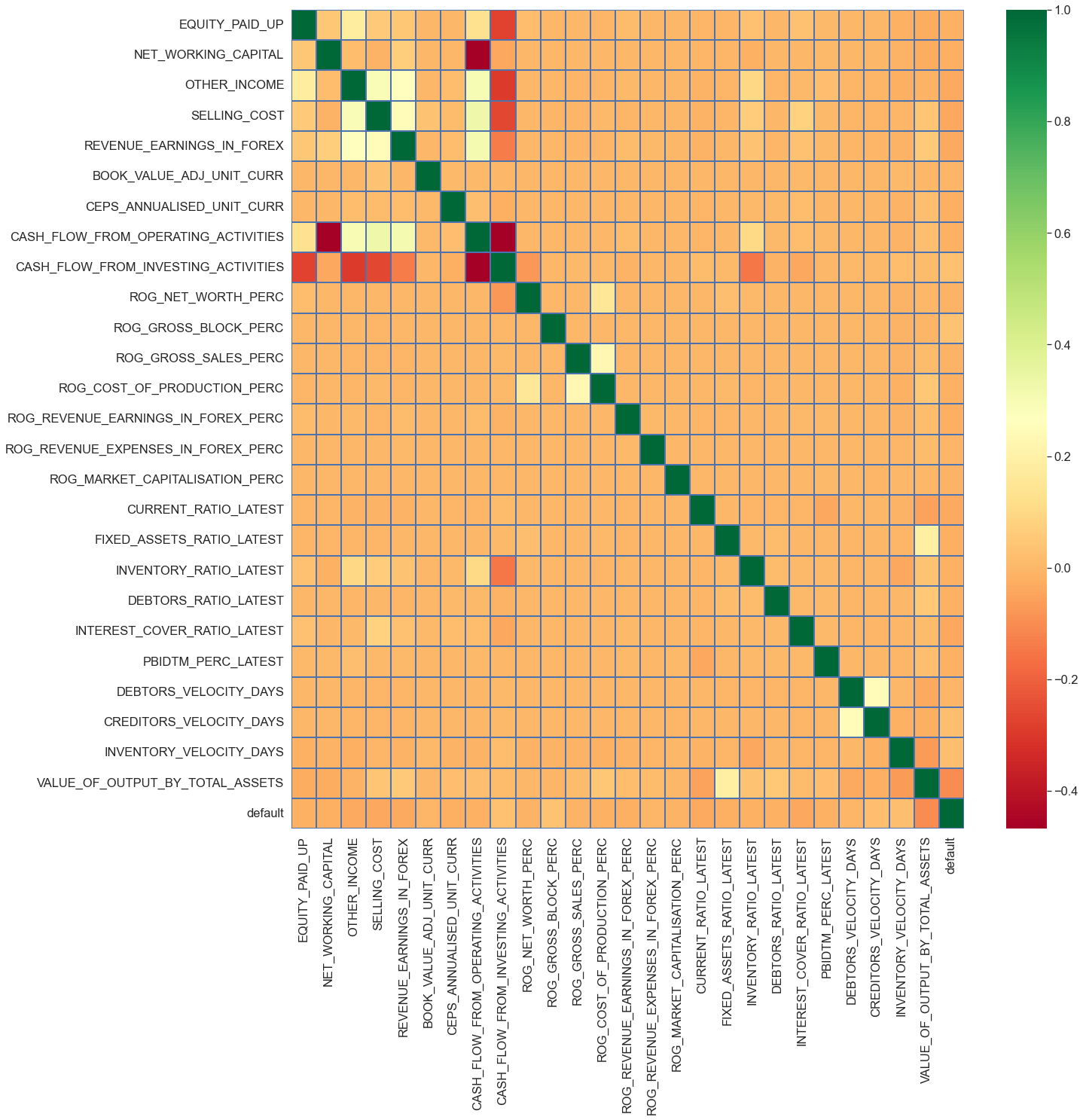
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **NETWORTH\_NEXT\_YEAR** | 3586.0 | 725.05 | 4769.68 | -8021.60 | 3.98 | 19.02 | 123.80 | 111729.10 |
| **EQUITY\_PAID\_UP** | 3586.0 | 62.97 | 778.76 | 0.00 | 3.75 | 8.29 | 19.52 | 42263.46 |
| **NETWORTH** | 3586.0 | 649.75 | 4091.99 | -7027.48 | 3.89 | 18.58 | 117.30 | 81657.35 |
| **CAPITAL\_EMPLOYED** | 3586.0 | 2799.61 | 26975.14 | -1824.75 | 7.60 | 39.09 | 226.60 | 714001.25 |
| **TOTAL\_DEBT** | 3586.0 | 1994.82 | 23652.84 | -0.72 | 0.03 | 7.49 | 72.35 | 652823.81 |
| **GROSS\_BLOCK** | 3586.0 | 594.18 | 4871.55 | -41.19 | 0.57 | 15.87 | 131.90 | 128477.59 |
| **NET\_WORKING\_CAPITAL** | 3586.0 | 410.81 | 6301.22 | -13162.42 | 0.94 | 10.14 | 61.18 | 223257.56 |
| **CURRENT\_ASSETS** | 3586.0 | 1960.35 | 22577.57 | -0.91 | 4.00 | 24.54 | 135.28 | 721166.00 |
| **CURRENT\_LIABILITIES\_AND\_PROVISIONS** | 3586.0 | 391.99 | 2675.00 | -0.23 | 0.73 | 9.23 | 65.65 | 83232.98 |
| **TOTAL\_ASSETS\_BY\_LIABILITIES** | 3586.0 | 1778.45 | 11437.57 | -4.51 | 10.56 | 52.01 | 310.54 | 254737.22 |
| **GROSS\_SALES** | 3586.0 | 1123.74 | 10603.70 | -62.59 | 1.44 | 31.21 | 242.25 | 474182.94 |
| **NET\_SALES** | 3586.0 | 1079.70 | 9996.57 | -62.59 | 1.44 | 30.44 | 234.44 | 443775.16 |
| **OTHER\_INCOME** | 3586.0 | 48.73 | 426.04 | -448.72 | 0.02 | 0.45 | 3.64 | 14143.40 |
| **VALUE\_OF\_OUTPUT** | 3586.0 | 1077.19 | 9843.88 | -119.10 | 1.41 | 30.90 | 235.84 | 435559.09 |
| **COST\_OF\_PRODUCTION** | 3586.0 | 798.54 | 9076.70 | -22.65 | 0.94 | 25.99 | 189.55 | 419913.50 |
| **SELLING\_COST** | 3586.0 | 25.55 | 194.24 | 0.00 | 0.00 | 0.16 | 3.88 | 5283.91 |
| **PBIDT** | 3586.0 | 248.18 | 1949.59 | -4655.14 | 0.04 | 2.04 | 23.52 | 42059.26 |
| **PBDT** | 3586.0 | 116.27 | 956.20 | -5874.53 | 0.00 | 0.80 | 12.94 | 23215.00 |
| **PBIT** | 3586.0 | 217.66 | 1850.97 | -4812.95 | 0.00 | 1.15 | 16.67 | 41402.96 |
| **PBT** | 3586.0 | 85.75 | 799.93 | -6032.34 | -0.06 | 0.31 | 7.42 | 16798.00 |
| **PAT** | 3586.0 | 61.22 | 620.30 | -6032.34 | -0.06 | 0.26 | 5.54 | 13383.39 |
| **ADJUSTED\_PAT** | 3586.0 | 60.06 | 580.43 | -4418.72 | -0.09 | 0.21 | 5.34 | 13384.11 |
| **CP** | 3586.0 | 91.73 | 780.79 | -5874.53 | 0.00 | 0.74 | 10.91 | 20760.20 |
| **REVENUE\_EARNINGS\_IN\_FOREX** | 3586.0 | 131.17 | 1150.73 | 0.00 | 0.00 | 0.00 | 7.20 | 46158.00 |
| **REVENUE\_EXPENSES\_IN\_FOREX** | 3586.0 | 256.33 | 4132.34 | 0.00 | 0.00 | 0.00 | 6.99 | 193979.73 |
| **CAPITAL\_EXPENSES\_IN\_FOREX** | 3586.0 | 7.66 | 111.43 | 0.00 | 0.00 | 0.00 | 0.00 | 3722.10 |
| **BOOK\_VALUE\_UNIT\_CURR** | 3586.0 | 157.24 | 1622.66 | -3371.57 | 7.96 | 21.66 | 71.67 | 75790.00 |
| **BOOK\_VALUE\_ADJ\_UNIT\_CURR** | 3582.0 | 2243.15 | 128283.73 | -33715.70 | 7.06 | 18.92 | 60.01 | 7677600.29 |
| **MARKET\_CAPITALISATION** | 3586.0 | 1664.09 | 12805.17 | 0.00 | 0.00 | 8.37 | 111.46 | 260865.08 |
| **CEPS\_ANNUALISED\_UNIT\_CURR** | 3586.0 | 36.02 | 828.42 | -1808.00 | 0.00 | 1.14 | 8.77 | 45438.44 |
| **CASH\_FLOW\_FROM\_OPERATING\_ACTIVITIES** | 3586.0 | 65.77 | 1455.05 | -25469.23 | -0.31 | 0.45 | 12.65 | 44529.40 |
| **CASH\_FLOW\_FROM\_INVESTING\_ACTIVITIES** | 3586.0 | -60.87 | 701.97 | -23843.45 | -5.12 | -0.12 | 0.12 | 3732.98 |
| **CASH\_FLOW\_FROM\_FINANCING\_ACTIVITIES** | 3586.0 | 11.44 | 1272.26 | -38374.04 | -5.85 | 0.00 | 0.46 | 28846.00 |
| **ROG\_NET\_WORTH\_PERC** | 3586.0 | 1237.62 | 41041.93 | -14485.71 | -1.49 | 1.84 | 11.36 | 2144020.00 |
| **ROG\_CAPITAL\_EMPLOYED\_PERC** | 3586.0 | 2988.88 | 126472.87 | -8614.63 | -3.84 | 1.38 | 12.59 | 7412700.00 |
| **ROG\_GROSS\_BLOCK\_PERC** | 3586.0 | 37.55 | 893.62 | -116.12 | 0.00 | 0.25 | 6.72 | 47400.00 |
| **ROG\_GROSS\_SALES\_PERC** | 3586.0 | 242.67 | 6103.53 | -5503.70 | -8.08 | 3.31 | 21.52 | 320200.00 |
| **ROG\_NET\_SALES\_PERC** | 3586.0 | 242.59 | 6103.49 | -5503.70 | -8.12 | 3.20 | 21.57 | 320200.00 |
| **ROG\_COST\_OF\_PRODUCTION\_PERC** | 3586.0 | 310.49 | 5573.22 | -2130.23 | -7.24 | 4.42 | 23.12 | 267150.00 |
| **ROG\_TOTAL\_ASSETS\_PERC** | 3586.0 | 2793.28 | 125941.65 | -136.13 | -3.97 | 1.48 | 12.50 | 7422120.00 |
| **ROG\_PBIDT\_PERC** | 3586.0 | 375.85 | 23278.40 | -52200.00 | -23.36 | 4.57 | 47.88 | 1386200.00 |
| **ROG\_PBDT\_PERC** | 3586.0 | 336.38 | 20353.40 | -52200.00 | -30.60 | 3.36 | 52.92 | 1208700.00 |
| **ROG\_PBIT\_PERC** | 3586.0 | 374.70 | 22462.79 | -58500.00 | -31.35 | 2.13 | 50.14 | 1338000.00 |
| **ROG\_PBT\_PERC** | 3586.0 | 224.07 | 19659.23 | -78900.00 | -41.24 | 0.02 | 61.96 | 1160500.00 |
| **ROG\_PAT\_PERC** | 3586.0 | 112.23 | 13480.52 | -114500.00 | -43.73 | 0.00 | 65.35 | 774200.00 |
| **ROG\_CP\_PERC** | 3586.0 | 221.09 | 13980.20 | -52200.00 | -29.51 | 4.62 | 52.91 | 822400.00 |
| **ROG\_REVENUE\_EARNINGS\_IN\_FOREX\_PERC** | 3586.0 | 37.23 | 658.67 | -100.00 | 0.00 | 0.00 | 0.00 | 29084.77 |
| **ROG\_REVENUE\_EXPENSES\_IN\_FOREX\_PERC** | 3586.0 | 364.86 | 15233.64 | -100.00 | 0.00 | 0.00 | 0.00 | 894591.69 |
| **ROG\_MARKET\_CAPITALISATION\_PERC** | 3586.0 | 63.68 | 1047.93 | -98.05 | 0.00 | 0.00 | 47.52 | 61865.26 |
| **CURRENT\_RATIO\_LATEST** | 3585.0 | 12.06 | 108.41 | 0.00 | 0.88 | 1.36 | 2.77 | 4813.00 |
| **FIXED\_ASSETS\_RATIO\_LATEST** | 3585.0 | 51.54 | 681.15 | 0.00 | 0.27 | 1.56 | 4.74 | 22172.00 |
| **INVENTORY\_RATIO\_LATEST** | 3585.0 | 37.80 | 458.19 | 0.00 | 0.00 | 3.56 | 8.94 | 15472.00 |
| **DEBTORS\_RATIO\_LATEST** | 3585.0 | 33.03 | 489.56 | 0.00 | 0.42 | 3.82 | 8.52 | 22992.67 |
| **TOTAL\_ASSET\_TURNOVER\_RATIO\_LATEST** | 3585.0 | 1.24 | 2.67 | 0.00 | 0.07 | 0.60 | 1.55 | 57.75 |
| **INTEREST\_COVER\_RATIO\_LATEST** | 3585.0 | 16.39 | 351.74 | -5450.00 | 0.00 | 1.08 | 3.71 | 18639.40 |
| **PBIDTM\_PERC\_LATEST** | 3585.0 | -51.16 | 1795.13 | -78870.45 | 0.00 | 8.07 | 18.99 | 19233.33 |
| **PBITM\_PERC\_LATEST** | 3585.0 | -109.21 | 3057.64 | -141600.00 | 0.00 | 5.23 | 14.29 | 19195.70 |
| **PBDTM\_PERC\_LATEST** | 3585.0 | -311.57 | 10921.59 | -590500.00 | 0.00 | 4.69 | 14.11 | 15640.00 |
| **CPM\_PERC\_LATEST** | 3585.0 | -307.01 | 10676.15 | -572000.00 | 0.00 | 3.89 | 11.39 | 15640.00 |
| **APATM\_PERC\_LATEST** | 3585.0 | -365.06 | 12500.05 | -688600.00 | 0.00 | 1.59 | 7.41 | 15266.67 |
| **DEBTORS\_VELOCITY\_DAYS** | 3586.0 | 603.89 | 10636.76 | 0.00 | 8.00 | 49.00 | 106.00 | 514721.00 |
| **CREDITORS\_VELOCITY\_DAYS** | 3586.0 | 2057.85 | 54169.48 | 0.00 | 8.00 | 39.00 | 89.00 | 2034145.00 |
| **INVENTORY\_VELOCITY\_DAYS** | 3483.0 | 79.64 | 137.85 | -199.00 | 0.00 | 35.00 | 96.00 | 996.00 |
| **VALUE\_OF\_OUTPUT\_BY\_TOTAL\_ASSETS** | 3586.0 | 0.82 | 1.20 | -0.33 | 0.07 | 0.48 | 1.16 | 17.63 |
| **VALUE\_OF\_OUTPUT\_BY\_GROSS\_BLOCK** | 3586.0 | 61.88 | 976.82 | -61.00 | 0.27 | 1.53 | 4.91 | 43404.00 |
| **default** | 3586.0 | 0.11 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Let us study the data more to see the correlation  **Heat Map correlation:** |  |  |  |  |  |  |  |  |



We can see that the data is highly correlated.

We have dropped highly correlated features.

The correlations for the selected 24 features are checked again.



The above Data set of 24 features was further cleaned by finding out the outliers in each feature.

The outliers are replaced by Nan values

The total Nan values in the data set were close to 18 %.

Features 'ROG\_REVENUE\_EARNINGS\_IN\_FOREX\_PERC', and 'ROG\_REVENUE\_EXPENSES\_IN\_FOREX\_PERC' were also dropped as they had more than 30 % records missing.

The Data was subsequently split in the ratio of 67 : 33 in Train and Test data set. Care was taken that the proportion of ‘default’ present in the full data set at 89:11 was maintained for the Train and Test data as well.

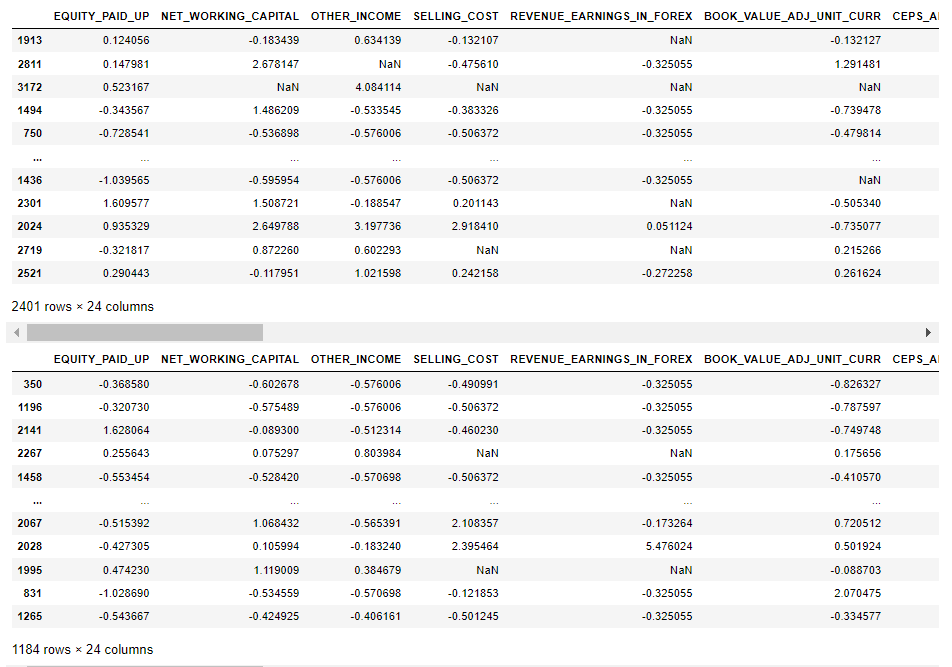
The Train data was scaled using the Standard Scaler tool.

Test Data was scaled using the mean and standard deviations of the train data features.

After scaling the Null values of the Train Data were imputed using the K Nearest Neighbor tool, using 10 neighbors as the parameter.

The null values in the Test data were imputed by values as ascertained by nearest neighbors of the Train Data

**Sample train and test dataset:**



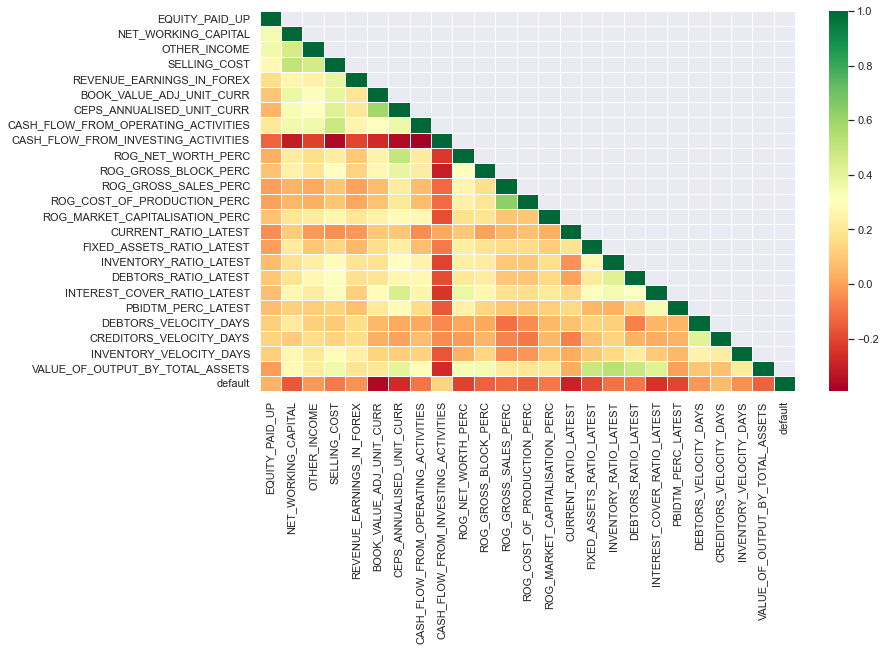
Target variable has binary values of 1 and 0. 1 means net worth next year decreases and 0 means net worth next year increases or is equal.

X\_train dataset and X\_test dataset is concatenated. Now above rows are X\_train and below them is X\_test.

Also, y\_train and y\_test is also concatenated.

The data is now scaled and imputed.

Let us check again correlation of our cleaned dataset.



There isn’t much collinearity now.

|  | **VIF** |
| --- | --- |
| **CEPS\_ANNUALISED\_UNIT\_CURR** | 2.342458 |
| **VALUE\_OF\_OUTPUT\_BY\_TOTAL\_ASSETS** | 2.327682 |
| **SELLING\_COST** | 2.001325 |
| **NET\_WORKING\_CAPITAL** | 1.781558 |
| **ROG\_GROSS\_SALES\_PERC** | 1.746198 |
| **ROG\_COST\_OF\_PRODUCTION\_PERC** | 1.713397 |
| **BOOK\_VALUE\_ADJ\_UNIT\_CURR** | 1.710451 |
| **INTEREST\_COVER\_RATIO\_LATEST** | 1.612414 |
| **OTHER\_INCOME** | 1.566357 |
| **CASH\_FLOW\_FROM\_OPERATING\_ACTIVITIES** | 1.565049 |
| **INVENTORY\_RATIO\_LATEST** | 1.551107 |
| **DEBTORS\_RATIO\_LATEST** | 1.539226 |
| **ROG\_NET\_WORTH\_PERC** | 1.480845 |
| **FIXED\_ASSETS\_RATIO\_LATEST** | 1.433282 |
| **CASH\_FLOW\_FROM\_INVESTING\_ACTIVITIES** | 1.380912 |
| **DEBTORS\_VELOCITY\_DAYS** | 1.375988 |
| **ROG\_GROSS\_BLOCK\_PERC** | 1.319728 |
| **CREDITORS\_VELOCITY\_DAYS** | 1.286420 |
| **EQUITY\_PAID\_UP** | 1.283740 |
| **PBIDTM\_PERC\_LATEST** | 1.279328 |
| **REVENUE\_EARNINGS\_IN\_FOREX** | 1.261746 |
| **INVENTORY\_VELOCITY\_DAYS** | 1.243324 |
| **ROG\_MARKET\_CAPITALISATION\_PERC** | 1.167552 |
| **CURRENT\_RATIO\_LATEST** | 1.124366 |

When we check the with VIF we find VIF values are less than 5.

We are splitting the data again into test and train from the same row from where we concatenated.

Our dependent variable is default.

**Logistic Regression Model 1:**

|  |  |  |  |
| --- | --- | --- | --- |
| Logit Regression Results | | | |
| **Dep. Variable:** | default | **No. Observations:** | 2401 |
| **Model:** | Logit | **Df Residuals:** | 2376 |
| **Method:** | MLE | **Df Model:** | 24 |
| **Date:** | Sun, 20 Nov 2022 | **Pseudo R-squ.:** | 0.5288 |
| **Time:** | 13:25:04 | **Log-Likelihood:** | -387.99 |
| **converged:** | True | **LL-Null:** | -823.35 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 2.303e-168 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | -4.8395 | 0.240 | -20.127 | 0.000 | -5.311 | -4.368 |
| **EQUITY\_PAID\_UP** | 0.1146 | 0.107 | 1.068 | 0.286 | -0.096 | 0.325 |
| **NET\_WORKING\_CAPITAL** | -0.2913 | 0.159 | -1.829 | 0.067 | -0.603 | 0.021 |
| **OTHER\_INCOME** | 0.3743 | 0.128 | 2.935 | 0.003 | 0.124 | 0.624 |
| **SELLING\_COST** | 0.4902 | 0.149 | 3.280 | 0.001 | 0.197 | 0.783 |
| **REVENUE\_EARNINGS\_IN\_FOREX** | -0.0241 | 0.115 | -0.210 | 0.833 | -0.249 | 0.201 |
| **BOOK\_VALUE\_ADJ\_UNIT\_CURR** | -3.2946 | 0.278 | -11.858 | 0.000 | -3.839 | -2.750 |
| **CEPS\_ANNUALISED\_UNIT\_CURR** | -0.2280 | 0.181 | -1.259 | 0.208 | -0.583 | 0.127 |
| **CASH\_FLOW\_FROM\_OPERATING\_ACTIVITIES** | -0.1293 | 0.140 | -0.924 | 0.355 | -0.404 | 0.145 |
| **CASH\_FLOW\_FROM\_INVESTING\_ACTIVITIES** | 0.1717 | 0.148 | 1.163 | 0.245 | -0.118 | 0.461 |
| **ROG\_NET\_WORTH\_PERC** | -0.1415 | 0.121 | -1.168 | 0.243 | -0.379 | 0.096 |
| **ROG\_GROSS\_BLOCK\_PERC** | -0.0819 | 0.157 | -0.522 | 0.602 | -0.389 | 0.226 |
| **ROG\_GROSS\_SALES\_PERC** | 0.1560 | 0.134 | 1.168 | 0.243 | -0.106 | 0.418 |
| **ROG\_COST\_OF\_PRODUCTION\_PERC** | -0.4813 | 0.134 | -3.579 | 0.000 | -0.745 | -0.218 |
| **ROG\_MARKET\_CAPITALISATION\_PERC** | -0.0020 | 0.106 | -0.019 | 0.985 | -0.210 | 0.206 |
| **CURRENT\_RATIO\_LATEST** | -1.5085 | 0.181 | -8.356 | 0.000 | -1.862 | -1.155 |
| **FIXED\_ASSETS\_RATIO\_LATEST** | -0.4430 | 0.203 | -2.186 | 0.029 | -0.840 | -0.046 |
| **INVENTORY\_RATIO\_LATEST** | -0.0386 | 0.129 | -0.299 | 0.765 | -0.292 | 0.215 |
| **DEBTORS\_RATIO\_LATEST** | -0.0445 | 0.126 | -0.354 | 0.724 | -0.291 | 0.202 |
| **INTEREST\_COVER\_RATIO\_LATEST** | -0.3723 | 0.157 | -2.369 | 0.018 | -0.680 | -0.064 |
| **PBIDTM\_PERC\_LATEST** | -0.1880 | 0.116 | -1.623 | 0.105 | -0.415 | 0.039 |
| **DEBTORS\_VELOCITY\_DAYS** | 0.1628 | 0.108 | 1.503 | 0.133 | -0.050 | 0.375 |
| **CREDITORS\_VELOCITY\_DAYS** | 0.1500 | 0.100 | 1.497 | 0.134 | -0.046 | 0.346 |
| **INVENTORY\_VELOCITY\_DAYS** | 0.0018 | 0.117 | 0.016 | 0.988 | -0.227 | 0.231 |
| **VALUE\_OF\_OUTPUT\_BY\_TOTAL\_ASSETS** | 0.3834 | 0.165 | 2.328 | 0.020 | 0.061 | 0.706 |

Possibly complete quasi-separation: A fraction 0.14 of observations can be  
perfectly predicted. This might indicate that there is complete  
quasi-separation. In this case some parameters will not be identified.

There are many features with P\_value more than 0.05. Hence those features are dropped and the model is built again.

**Model 2:**

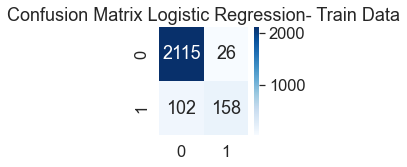
|  |  |  |  |
| --- | --- | --- | --- |
| Logit Regression Results | | | |
| **Dep. Variable:** | default | **No. Observations:** | 2401 |
| **Model:** | Logit | **Df Residuals:** | 2392 |
| **Method:** | MLE | **Df Model:** | 8 |
| **Date:** | Sun, 20 Nov 2022 | **Pseudo R-squ.:** | 0.5150 |
| **Time:** | 13:25:05 | **Log-Likelihood:** | -399.33 |
| **converged:** | True | **LL-Null:** | -823.35 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 9.034e-178 |

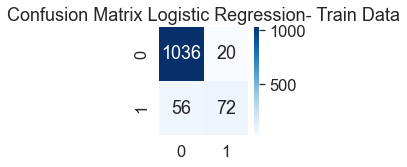
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | -4.7205 | 0.229 | -20.634 | 0.000 | -5.169 | -4.272 |
| **NET\_WORKING\_CAPITAL** | -0.3027 | 0.149 | -2.028 | 0.043 | -0.595 | -0.010 |
| **OTHER\_INCOME** | 0.4031 | 0.118 | 3.424 | 0.001 | 0.172 | 0.634 |
| **SELLING\_COST** | 0.4977 | 0.130 | 3.838 | 0.000 | 0.244 | 0.752 |
| **BOOK\_VALUE\_ADJ\_UNIT\_CURR** | -3.4553 | 0.270 | -12.793 | 0.000 | -3.985 | -2.926 |
| **ROG\_COST\_OF\_PRODUCTION\_PERC** | -0.4675 | 0.110 | -4.256 | 0.000 | -0.683 | -0.252 |
| **CURRENT\_RATIO\_LATEST** | -1.4824 | 0.164 | -9.021 | 0.000 | -1.804 | -1.160 |
| **INTEREST\_COVER\_RATIO\_LATEST** | -0.4515 | 0.138 | -3.261 | 0.001 | -0.723 | -0.180 |
| **PBIDTM\_PERC\_LATEST** | -0.2307 | 0.111 | -2.077 | 0.038 | -0.448 | -0.013 |

Possibly complete quasi-separation: A fraction 0.13 of observations can be  
perfectly predicted. This might indicate that there is complete  
quasi-separation. In this case some parameters will not be identified.

WE have completed building a model with only 8 features and discarding others to build an efficient model.

Confusion Matrix:





Classification Report:

Classification Report - Logistic Regression- Train Data

precision recall f1-score support

0 0.95 0.99 0.97 2141

1 0.86 0.61 0.71 260

accuracy 0.95 2401

macro avg 0.91 0.80 0.84 2401

weighted avg 0.94 0.95 0.94 2401

Classifiaction Report - Logistic Regression- Test Data

precision recall f1-score support

0 0.95 0.98 0.96 1056

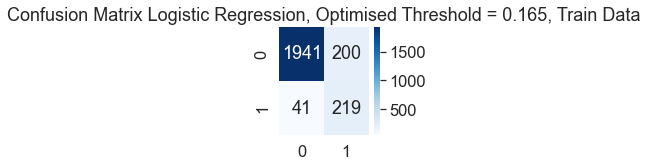
1 0.78 0.56 0.65 128

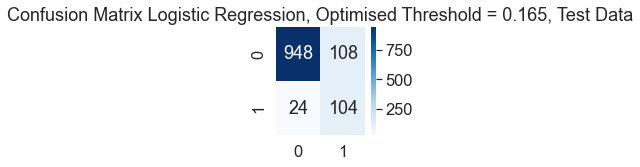
accuracy 0.94 1184

macro avg 0.87 0.77 0.81 1184

weighted avg 0.93 0.94 0.93 1184

We are further improving the model by optimising the threshold level to 0.165.





Classification Report - Logistic Regression, Optimised Threshold = 0.165, Train Data

precision recall f1-score support

0 0.98 0.91 0.94 2141

1 0.52 0.84 0.65 260

accuracy 0.90 2401

macro avg 0.75 0.87 0.79 2401

weighted avg 0.93 0.90 0.91 2401

Classification Report - Logistic Regression, Optimised Threshold = 0.165, Test Data

precision recall f1-score support

0 0.98 0.90 0.93 1056

1 0.49 0.81 0.61 128

accuracy 0.89 1184

macro avg 0.73 0.86 0.77 1184

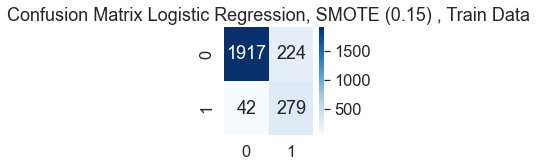
weighted avg 0.92 0.89 0.90 1184

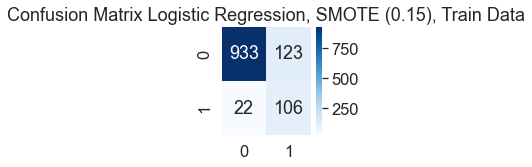
We find that recall for 1 has improved but precision has gone down.

We will now balance the data using SMOTE

Sampling Strategy Ratio - Initial proportion of ‘default’ in original data is 388 / 3197 = 0.12.

Using the SMOTE, we iterate the sampling strategy ratio and see that by increasing the proportion of default by engineering synthetic data, we get models which over fit. The most ideal ratio ascertained is 0.15, where there was an improvement in the model performance and the Recall and Precision readings for ‘1’ were reasonably together.





Classification Report- Logistic Regression, SMOTE (0.15) , Train Data

precision recall f1-score support

0 0.98 0.90 0.94 2141

1 0.55 0.87 0.68 321

accuracy 0.89 2462

macro avg 0.77 0.88 0.81 2462

weighted avg 0.92 0.89 0.90 2462

Classification Report- Logistic Regression, SMOTE (0.15) , Test Data

precision recall f1-score support

0 0.98 0.88 0.93 1056

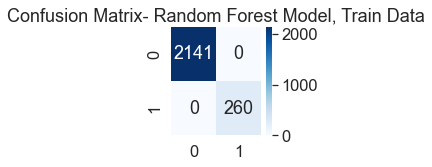
1 0.46 0.83 0.59 128

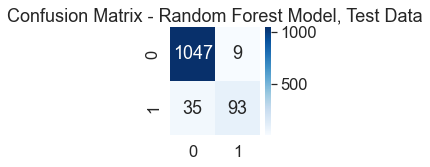
accuracy 0.88 1184

macro avg 0.72 0.86 0.76 1184

weighted avg 0.92 0.88 0.89 1184

**Random Forest Model:**





Classification Report- Random Forest Model, Train Data

precision recall f1-score support

0 1.00 1.00 1.00 2141

1 1.00 1.00 1.00 260

accuracy 1.00 2401

macro avg 1.00 1.00 1.00 2401

weighted avg 1.00 1.00 1.00 2401

Classification Report- Random Forest Model, Train Data

precision recall f1-score support

0 0.97 0.99 0.98 1056

1 0.91 0.73 0.81 128

accuracy 0.96 1184

macro avg 0.94 0.86 0.89 1184

weighted avg 0.96 0.96 0.96 1184

We have a over fit model here with random forest model.

We will use GridSearch module to improve the model.

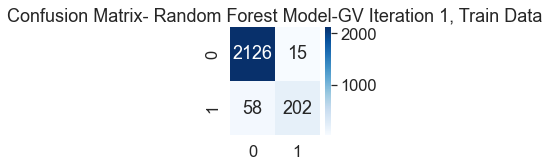
Parameters that we have with this module are:

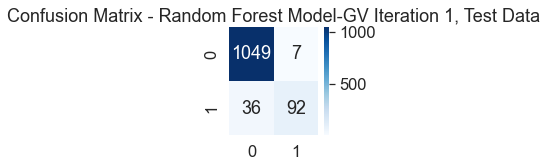
max\_depth': [3, 5, 7],

'min\_samples\_leaf': [5, 10, 15],

'min\_samples\_split': [15, 30, 15],

'n\_estimators': [25, 50]





Classification Report- Random Forest Model-GV Iteration 1, Train Data

precision recall f1-score support

0 0.97 0.99 0.98 2141

1 0.93 0.78 0.85 260

accuracy 0.97 2401

macro avg 0.95 0.88 0.92 2401

weighted avg 0.97 0.97 0.97 2401

Classification Report- Random Forest Model-GV Iteration 1 Train Data

precision recall f1-score support

0 0.97 0.99 0.98 1056

1 0.93 0.72 0.81 128

accuracy 0.96 1184

macro avg 0.95 0.86 0.90 1184

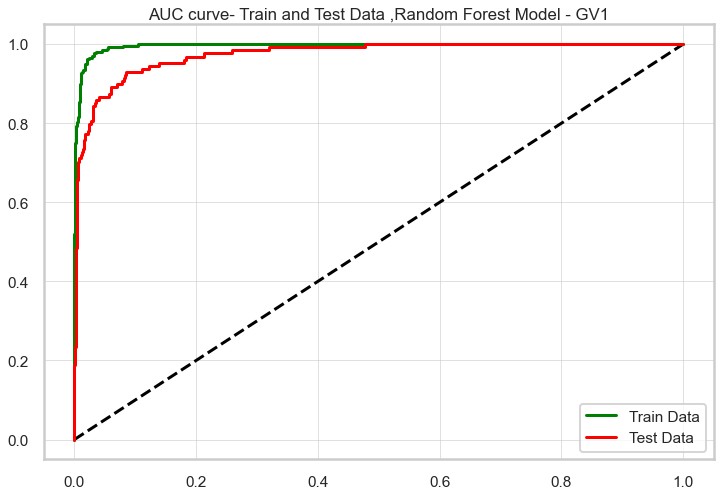
weighted avg 0.96 0.96 0.96 1184

We have a random forest model with a 71% recall and 93% precision. And these results are similar in train and test data.

This model is best random forest model.

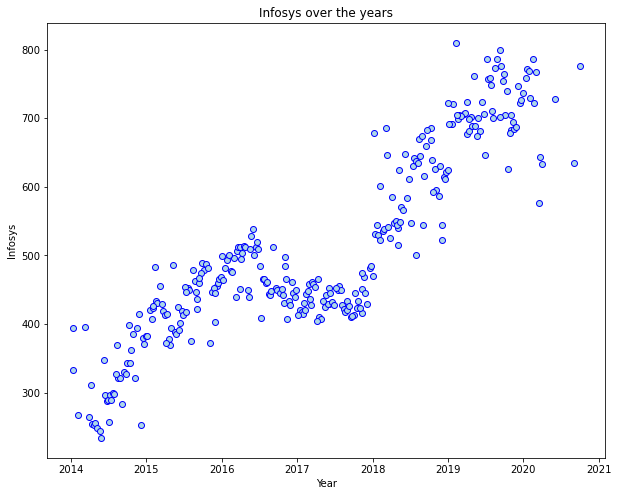
Now that we have completed model building we can compare all the models and we find that the best model is the random forest model with GridSearch parameters. Hence it should be chosen.

**ROC curve for random forest model:**



**Draw Stock Price Graph(Stock Price vs Time):**

Infosys

****

count 314.000000

mean 511.340764

std 135.952051

min 234.000000

25% 424.000000

50% 466.500000

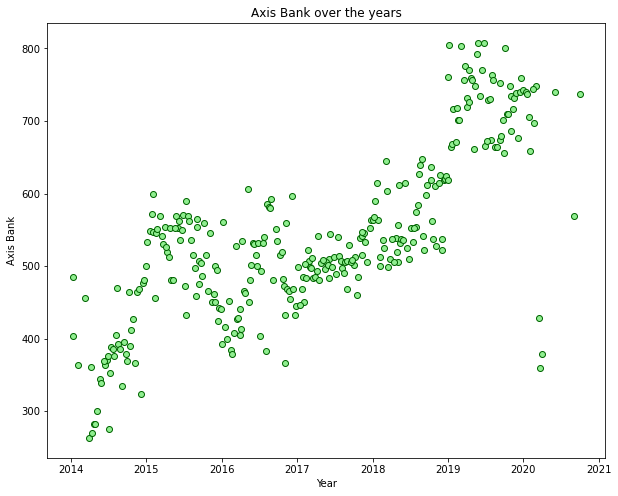
75% 630.750000

max 810.000000

Name: infosys, dtype: float64

The stocks are rising for Infosys.

Axis Bank



count 314.000000

mean 540.742038

std 115.835569

min 263.000000

25% 470.500000

50% 528.000000

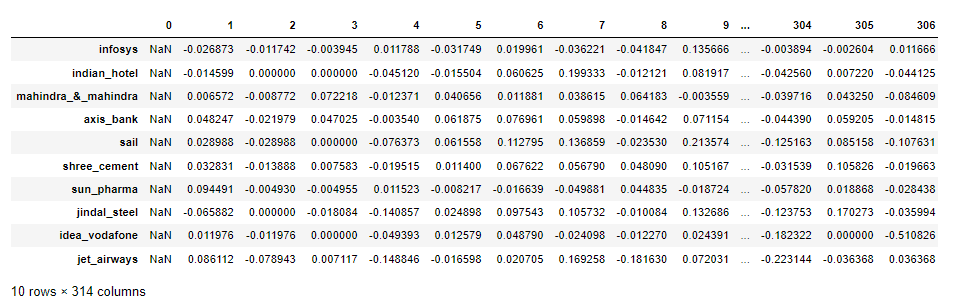
75% 605.250000

max 808.000000

Name: axis\_bank, dtype: float64

Stocks for axis bank has also been rising.

**Calculate Returns for all stocks with inference**



Weekly returns on stocks of all the companies can be see in the above chart.

**Calculate Stock Means and Standard Deviation for all stocks with inference**

infosys 0.035070

indian\_hotel 0.047131

mahindra\_&\_mahindra 0.040169

axis\_bank 0.045828

sail 0.062188

shree\_cement 0.039917

sun\_pharma 0.045033

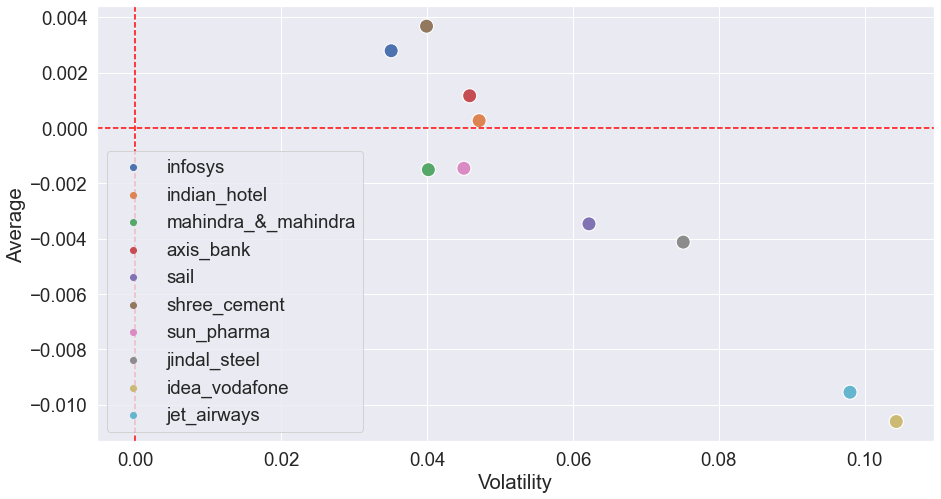
jindal\_steel 0.075108

idea\_vodafone 0.104315

jet\_airways 0.097972

dtype: float64

**Draw a plot of Stock Means vs Standard Deviation and state your inference**

****