Finance and Risk Analytics

Project (Milestone-1)

The report is based on a Data that is available includes information from the financial statement of the companies for the previous year (2015). Based on this data we have tried understanding different companies and build a model to check the likelihood of their default.

University of Texas at Austin

Great Learning

Submitted by- Gunjar Fuley

Contact- 9938126651

10/17/2021

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.

Solution

1) We have received a dataset of the companies in .xlsx, Microsoft Excel Worksheet (Company_Data2015-1.xlsx) format and therefore we have converted it to the .CSV format for better execution.
2) We have imported all the necessary Python libraries for in the Jupyter Notebook for analysis.
3) We have imported the converted the CSV file for the analysis.
4) We checked the head as well tail of the dataset and understood that it has 5063 rows and 80 columns. Each row refers to data of each company. The data has columns which contains the other critical information regarding the company like Networth, Capital Employed, Current Assets, Net Working Capital and so on. These parameters tells us how the performance of the company last year.

   Head:

	Co_Code	Co_Name	Networth Next Year	Equity Paid Up	Networth	Capital Employed	Total Debt	Gross Block	Net Working Capital	Current Assets	...	Unnamed: 70	Unnamed: 71	Unnamed: 72	Unnamed: 73	Unna
0	16974.0	Hind.Cables	-8021.60	419.36	-7,027.48	-1,007.24	5,936.03	474.3	-1,076.34	40.5	...	NaN	NaN	NaN	NaN	
1	21214.0	Tata Tele. Mah.	-3986.19	1,954.93	-2,968.08	4,458.20	7,410.18	9,070.86	-1,098.88	486.86	...	NaN	NaN	NaN	NaN	
2	14852.0	ABG Shipyard	-3192.58	53.84	506.86	7,714.68	6,944.54	1,281.54	4,496.25	9,097.64	...	NaN	NaN	NaN	NaN	
3	2439.0	GTL	-3054.51	157.3	-623.49	2,353.88	2,326.05	1,033.69	-2,612.42	1,034.12	...	NaN	NaN	NaN	NaN	
4	23505.0	Bharati Defence	-2967.36	50.3	-1,070.83	4,675.33	5,740.90	1,084.20	1,836.23	4,685.81	...	NaN	NaN	NaN	NaN	

   5 rows × 80 columns

   Tail:

	Co_Code	Co_Name	Networth Next Year	Equity Paid Up	Networth	Capital Employed	Total Debt	Gross Block	Net Working Capital	Current Assets	...	Unnamed: 70	Unnamed: 71	Unnamed: 72	Unnamed: 73	Unnamed: 74	U
5058	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
5059	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
5060	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
5061	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
5062	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	

5 rows × 80 columns

Also we can understand here that many of the columns as well rows in the dataset are null values. In many cases, we can see that the rows and columns are completely null values.

5) It was observed that most of the variables names had spaces in between as well special characters like (,),% etc. which may create problem ahead during the analysis. Therefore, we have replaced the spaces with '_' and removed special characters.
Example- Variable 'ROG-Cost of Production (%)' was converted to 'ROG_Cost_of_Prod_perc'

6) In this step we have checked the most of the variables has datatype of 'object' but the values are in float. Therefore, we need to change the datatype to float. This has been done in next steps.

```
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Co_Code                            3586 non-null   float64
 1   Co_Name                            3586 non-null   object
 2   Networth Next Year                 3586 non-null   float64
 3   Equity Paid Up                     3586 non-null   object
 4   Networth                           3586 non-null   object
 5   Capital Employed                   3586 non-null   object
 6   Total Debt                         3586 non-null   object
 7   Gross Block                        3586 non-null   object
 8   Net Working Capital                3586 non-null   object
 9   Current Assets                     3586 non-null   object
 10  Current Liabilities and Provisions 3586 non-null   object
 11  Total Assets/Liabilities           3586 non-null   object
 12  Gross Sales                        3586 non-null   object
 13  Net Sales                          3586 non-null   object
 14  Other Income                       3586 non-null   object
 15  Value Of Output                    3586 non-null   object
 16  Cost of Production                 3586 non-null   object
 17  Selling Cost                       3586 non-null   object
 18  PBIDT                              3586 non-null   object
 19  PBDT                               3586 non-null   object
 20  PBIT                               3586 non-null   object
 21  PBT                                3586 non-null   object
 22  PAT                                3586 non-null   object
```

```
23  Adjusted PAT                            3586 non-null   object
24  CP                                      3586 non-null   object
25  Revenue earnings in forex               3586 non-null   object
26  Revenue expenses in forex               3586 non-null   object
27  Capital expenses in forex               3586 non-null   object
28  Book Value (Unit Curr)                  3586 non-null   object
29  Book Value (Adj.) (Unit Curr)           3582 non-null   object
30  Market Capitalisation                   3586 non-null   object
31  CEPS (annualised) (Unit Curr)           3586 non-null   object
32  Cash Flow From Operating Activities     3586 non-null   object
33  Cash Flow From Investing Activities     3586 non-null   object
34  Cash Flow From Financing Activities     3586 non-null   object
35  ROG-Net Worth (%)                       3586 non-null   object
36  ROG-Capital Employed (%)                3586 non-null   object
37  ROG-Gross Block (%)                     3586 non-null   object
38  ROG-Gross Sales (%)                     3586 non-null   object
39  ROG-Net Sales (%)                       3586 non-null   object
40  ROG-Cost of Production (%)              3586 non-null   object
41  ROG-Total Assets (%)                    3586 non-null   object
42  ROG-PBIDT (%)                           3586 non-null   object
43  ROG-PBDT (%)                            3586 non-null   object
44  ROG-PBIT (%)                            3586 non-null   object
45  ROG-PBT (%)                             3586 non-null   object
46  ROG-PAT (%)                             3586 non-null   object
47  ROG-CP (%)                              3586 non-null   object
48  ROG-Revenue earnings in forex (%)       3586 non-null   object
49  ROG-Revenue expenses in forex (%)       3586 non-null   object
50  ROG-Market Capitalisation (%)        3586 non-null   object
51  Current Ratio[Latest]                3585 non-null   object
52  Fixed Assets Ratio[Latest]           3585 non-null   object
53  Inventory Ratio[Latest]              3585 non-null   object
54  Debtors Ratio[Latest]                3585 non-null   object
55  Total Asset Turnover Ratio[Latest]   3585 non-null   float64
56  Interest Cover Ratio[Latest]         3585 non-null   object
57  PBIDTM (%)[Latest]                   3585 non-null   object
58  PBITM (%)[Latest]                    3585 non-null   object
59  PBDTM (%)[Latest]                    3585 non-null   object
60  CPM (%)[Latest]                      3585 non-null   object
61  APATM (%)[Latest]                    3585 non-null   object
62  Debtors Velocity (Days)              3586 non-null   object
63  Creditors Velocity (Days)            3586 non-null   object
64  Inventory Velocity (Days)            3483 non-null   float64
65  Value of Output/Total Assets         3586 non-null   float64
66  Value of Output/Gross Block          3586 non-null   object
67  Unnamed: 67                          0 non-null      float64
68  Unnamed: 68                          0 non-null      float64
69  Unnamed: 69                          0 non-null      float64
70  Unnamed: 70                          0 non-null      float64
71  Unnamed: 71                          0 non-null      float64
72  Unnamed: 72                          0 non-null      float64
73  Unnamed: 73                          0 non-null      float64
74  Unnamed: 74                          0 non-null      float64
75  Unnamed: 75                          0 non-null      float64
75  Unnamed: 75                          0 non-null      float64
76  Unnamed: 76                          0 non-null      float64
77  Unnamed: 77                          0 non-null      float64
78  Unnamed: 78                          0 non-null      float64
79  Unnamed: 79                          0 non-null      float64
```

7) We have confirmed the fact that many of columns as well as rows have complete 100% null values. In the below image it is confirmed.

```
Co_Code               1477
Co_Name               1477
Networth_Next_Year    1477
Equity_Paid_Up        1477
Networth              1477
                      ...
Unnamed: 75           5063
Unnamed: 76           5063
Unnamed: 77           5063
Unnamed: 78           5063
Unnamed: 79           5063
```

8) Now we have deleted all the columns and rows with null values.
After deletion, now the data have some columns left which has null values like variable 'Inventory_Vel_Days' which shall be treated later.

```
Co_Code                              0
Co_Name                              0
Networth_Next_Year                   0
Equity_Paid_Up                       0
Networth                             0
                                  ...
Debtors_Vel_Days                     0
Creditors_Vel_Days                   0
Inventory_Vel_Days                 103
Value_of_Output_to_Total_Assets      0
Value_of_Output_to_Gross_Block       0
```

9) After removal of rows and columns with 100% null values, we checked the number of rows and columns in the dataset. It was found that the number of rows (observations) is 3586 and the number of columns (variables) is 67.
10) According to the finance domain understanding, we know that companies with positive net worth next year are not the defaulters where are the companies with negative net worth next year are the defaulters. Therefore, 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. The default variable was created and checked subsequently.

	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

Also we have checked that out of 3586 companies, the number of defaulters are 388. Defaulters are approximately 11% of all the companies.

```
0    0.891801
1    0.108199
```

<u>Cleaning the dataset</u>

11) We have seen in the above steps that the data base has values in the columns with ‘,’ notation. E.g. 10000 is mentioned as 10,000. For such data, the datatype is string. Therefore for analysis purpose we need to remove the commas in the dataset.
12) The comma among the dataset was removed except the columns like Company name and the datatype for the variables was converted to float.

    Below we can see that the datatypes for the variables with decimal and numeric values has been converted to float. However, the company name remained as object only.

```
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Equity_Paid_Up                     3586 non-null   float64
 1   Networth                           3586 non-null   float64
 2   Capital_Employed                   3586 non-null   float64
 3   Total_Debt                         3586 non-null   float64
 4   Gross_Block                        3586 non-null   float64
 5   Net_Working_Capital                3586 non-null   float64
 6   Curr_Assets                        3586 non-null   float64
 7   Curr_Liab_and_Prov                 3586 non-null   float64
 8   Total_Assets_to_Liab               3586 non-null   float64
 9   Gross_Sales                        3586 non-null   float64
 10  Net_Sales                          3586 non-null   float64
 11  Other_Income                       3586 non-null   float64
 12  Value_Of_Output                    3586 non-null   float64
 13  Cost_of_Prod                       3586 non-null   float64
 14  Selling_Cost                       3586 non-null   float64
 15  PBIDT                              3586 non-null   float64
 16  PBDT                               3586 non-null   float64
 17  PBIT                               3586 non-null   float64
 18  PBT                                3586 non-null   float64
 19  PAT                                3586 non-null   float64
 20  Adjusted_PAT                       3586 non-null   float64
 21  CP                                 3586 non-null   float64
 22  Rev_earn_in_forex                  3586 non-null   float64
 23  Rev_exp_in_forex                   3586 non-null   float64
 24  Capital_exp_in_forex               3586 non-null   float64
```

```
25   Book_Value_Unit_Curr                  3586 non-null   float64
26   Book_Value_Adj_Unit_Curr              3582 non-null   float64
27   Market_Capitalisation                 3586 non-null   float64
28   CEPS_annualised_Unit_Curr             3586 non-null   float64
29   Cash_Flow_From_Opr                    3586 non-null   float64
30   Cash_Flow_From_Inv                    3586 non-null   float64
31   Cash_Flow_From_Fin                    3586 non-null   float64
32   ROG_Net_Worth_perc                    3586 non-null   float64
33   ROG_Capital_Employed_perc             3586 non-null   float64
34   ROG_Gross_Block_perc                  3586 non-null   float64
35   ROG_Gross_Sales_perc                  3586 non-null   float64
36   ROG_Net_Sales_perc                    3586 non-null   float64
37   ROG_Cost_of_Prod_perc                 3586 non-null   float64
38   ROG_Total_Assets_perc                 3586 non-null   float64
39   ROG_PBIDT_perc                        3586 non-null   float64
40   ROG_PBDT_perc                         3586 non-null   float64
41   ROG_PBIT_perc                         3586 non-null   float64
42   ROG_PBT_perc                          3586 non-null   float64
43   ROG_PAT_perc                          3586 non-null   float64
44   ROG_CP_perc                           3586 non-null   float64
45   ROG_Rev_earn_in_forex_perc            3586 non-null   float64
46   ROG_Rev_exp_in_forex_perc             3586 non-null   float64
47   ROG_Market_Capitalisation_perc        3586 non-null   float64
48   Curr_Ratio_Latest                     3585 non-null   float64
49   Fixed_Assets_Ratio_Latest             3585 non-null   float64
50   Inventory_Ratio_Latest                3585 non-null   float64
51   Debtors_Ratio_Latest                  3585 non-null   float64
52   Interest_Cover_Ratio_Latest           3585 non-null   float64
53   PBIDTM_perc_Latest                    3585 non-null   float64
54   PBITM_perc_Latest                     3585 non-null   float64
55   PBDTM_perc_Latest                     3585 non-null   float64
56   CPM_perc_Latest                       3585 non-null   float64
57   APATM_perc_Latest                     3585 non-null   float64
58   Debtors_Vel_Days                      3586 non-null   float64
59   Creditors_Vel_Days                    3586 non-null   float64
60   Value_of_Output_to_Gross_Block        3586 non-null   float64
61   Co_Name                               3586 non-null   object
62   Co_Code                               3586 non-null   float64
63   Networth_Next_Year                    3586 non-null   float64
64   Total_Asset_Turnover_Ratio_Latest     3585 non-null   float64
65   Inventory_Vel_Days                    3483 non-null   float64
66   Value_of_Output_to_Total_Assets       3586 non-null   float64
67   default                               3586 non-null   float64
```

Checking the missing values and the outliers

13) The size of the total dataset is (number of rows* number of columns) 243848.
    The total number of missing values in the entire dataset is 118.
14) For checking the outliers we first removed the variable default because it has all the values in binary format only.

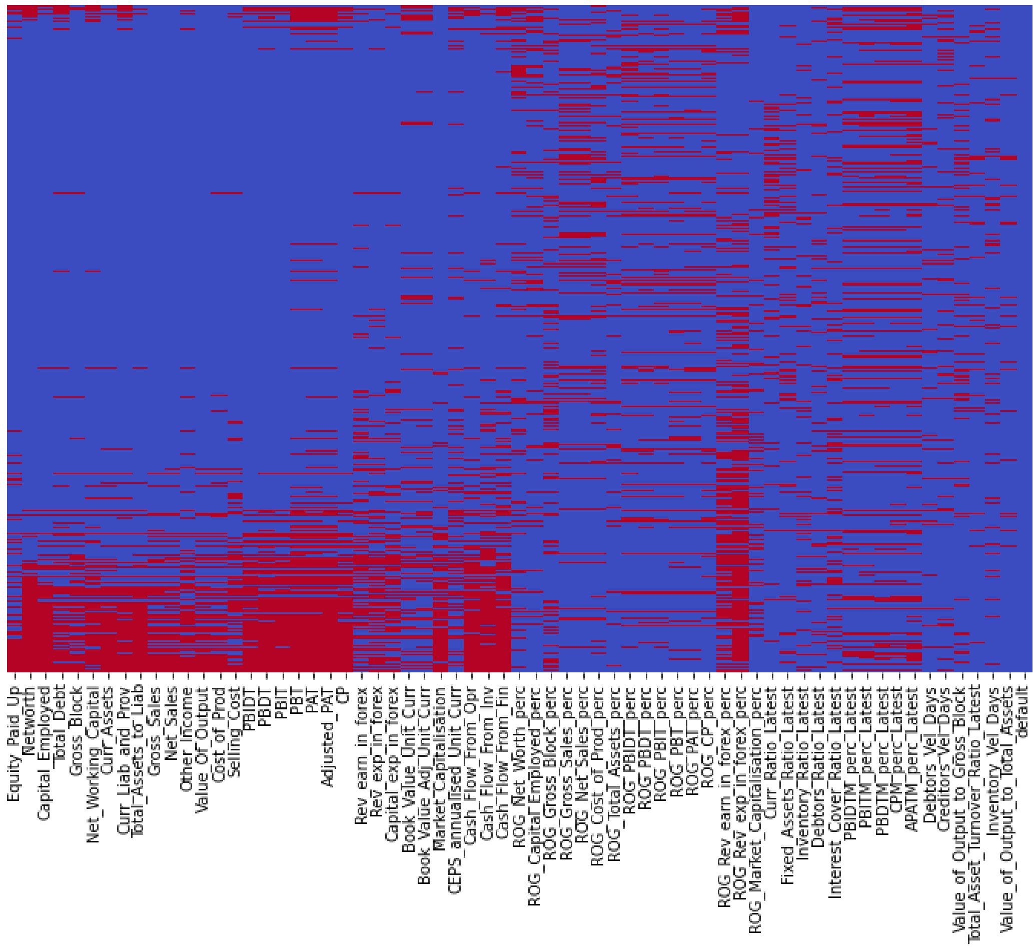
15) The numbers of outliers were significantly present in the dataset. We can understand that almost every column variable has data points which need to be worked upon.

```
APATM_perc_Latest                  933
Adjusted_PAT                       954
Book_Value_Adj_Unit_Curr           486
Book_Value_Unit_Curr               485
CEPS_annualised_Unit_Curr          602
                                  ...
Total_Assets_to_Liab               574
Total_Debt                         583
Value_Of_Output                    559
Value_of_Output_to_Gross_Block     481
Value_of_Output_to_Total_Assets    150
Length: 67, dtype: int64
```

Outliers and Missing values treatment

16) Here we have decided to follow a non- conventional method of converting all the outliers to NaN or Null values instead of imputing them with mean or median. The argument here is the quality of data might degrade.
17) Finally the number of null values in the dataset after converting outliers to NaN in totally becomes 42440. However, the shape of the data is 3586 rows and 67 columns.
18) Now we have to remove columns 'Networth_Next_Year','Co_Code' and 'Co_Name'. 'Networth_Next_Year' is converted into ‘default’ therefore they are highly correlated & might affect the analysis. The other two variables are merely indicators therefore we need to remove them.

19) Below we have visually inspected the missing values in our data



20) For the better analysis, we shall be going ahead with data points were the missing values are less than or equal to five. Because if the number of variables or features available shall be greater than 5 then it might give wrong interpretation. But after removal of these rows we found that the 70% of total defaulters were missed out. Therefore, we decided not to move ahead with this treatment.

21) Further we have checked the columns in decreasing order of the missing values.
Below are list of variables in order of percentage of missing values in them-

ROG_Rev_exp_in_forex_perc	0.450363
ROG_Rev_earn_in_forex_perc	0.367262
Cash_Flow_From_Fin	0.280257
PAT	0.267429
Adjusted_PAT	0.266035
PBT	0.262409
APATM_perc_Latest	0.260457
Cash_Flow_From_Inv	0.244283
ROG_Gross_Block_perc	0.231456
CP	0.227552
PBDT	0.227273
Cash_Flow_From_Opr	0.223369
ROG_Net_Worth_perc	0.208310
Rev_earn_in_forex	0.205800
Interest_Cover_Ratio_Latest	0.202454
CPM_perc_Latest	0.201060
PBIT	0.200781
PBITM_perc_Latest	0.200223
PBDTM_perc_Latest	0.194088
Capital_exp_in_forex	0.193530
Rev_exp_in_forex	0.193252
ROG_Cost_of_Prod_perc	0.188232
ROG_Gross_Sales_perc	0.187117
PBIDT	0.187117
ROG_Net_Sales_perc	0.186001
Networth	0.181260

Here we can see that there are 7 column variables where more than 25% of the data is NULL. Therefore we shall get rid of the variables 'ROG_Rev_exp_in_forex_perc', 'ROG_Rev_earn_in_forex_perc', 'Cash_Flow_From_Fin', 'PAT', 'Adjusted_PAT', 'PBT' and 'APATM_perc_Latest'.

22) Finally after the treatment we are left with 3586 rows and 58 columns.

Segregating and Scaling the predictors

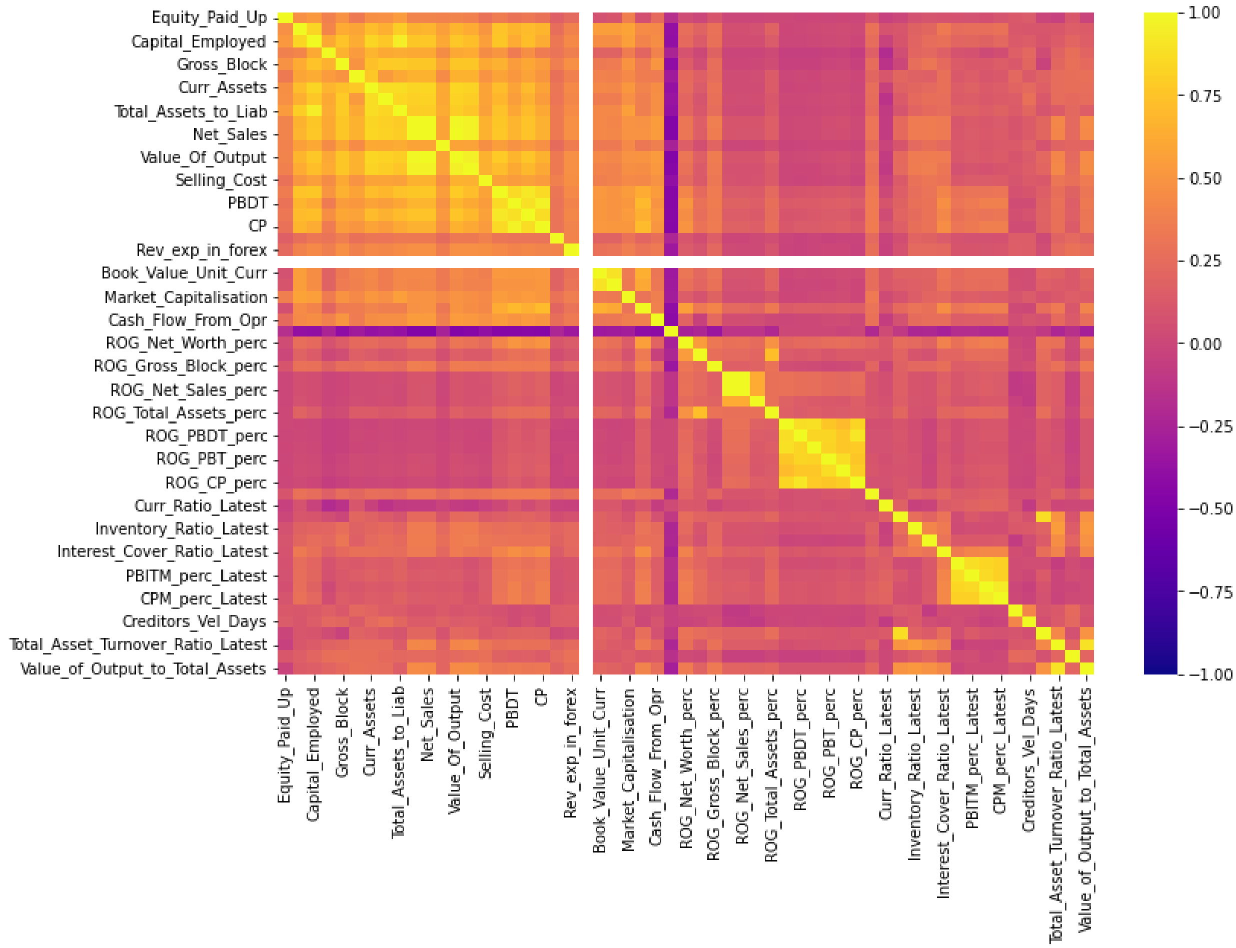
23) In this step we shall remove the predictors that the dataset with columns other than default. As default is the response i.e. it tells that the company defaulted or not.
24) Then we have scaled the predictors using StandardScaler library of python

Imputing the remaining missing values

25) The missing or null values in the dataset were imputed using the KNN (K- Nearest Neighbor) technique. Here we have opted 10 nearest values. The mean of those nearest values shall be imputed in the missing values. After imputation the null values were checked. Finally we have done treatment for outliers and null values for further model building.

Inspecting possible correlations between independent variables

26) We have checked the correlation between the variables using a heat map.



Here we can see the correlation between various variables but due to high number of dependent variables we cannot distinguish according to the color.

Splitting the data into train and test sets

27) We have split the data into in a ratio of 67:33

For modeling we will use Logistic Regression with recursive feature elimination

28) Here first we shall be selecting the initial first 20 features using the RFE function. Typically we select one third of the total number features to select the top features. We have total 58 features therefore we are going with 20 features. Below are the selected 20 top features. However, It has not treated multicollinearity.

	Feature	Rank
0	Equity_Paid_Up	1
1	Networth	1
2	Capital_Employed	1
4	Gross_Block	1
6	Curr_Assets	1
7	Curr_Liab_and_Prov	1
8	Total_Assets_to_Liab	1
11	Other_Income	1
12	Value_Of_Output	1
13	Cost_of_Prod	1
15	PBIDT	1
17	PBIT	1
22	Book_Value_Unit_Curr	1
23	Book_Value_Adj_Unit_Curr	1
25	CEPS_annualised_Unit_Curr	1
26	Cash_Flow_From_Opr	1
28	ROG_Net_Worth_perc	1
29	ROG_Capital_Employed_perc	1
42	Curr_Ratio_Latest	1
46	Interest_Cover_Ratio_Latest	1

Validating the model on train and test set

29) We have done validation of model on the train as well as test set and we found that the recall as 67% and 55% for train and test respectively. Since only ~11% of the total data had defaults, we will now try to balance the data before fitting the model.

Train:

```
              precision    recall  f1-score   support

         0.0       0.96      0.99      0.98      2151
         1.0       0.87      0.67      0.76       251

    accuracy                           0.96      2402
   macro avg       0.92      0.83      0.87      2402
weighted avg       0.95      0.96      0.95      2402
```

Test:

	precision	recall	f1-score	support
0.0	0.94	0.99	0.97	1047
1.0	0.88	0.55	0.68	137
accuracy			0.94	1184
macro avg	0.91	0.77	0.82	1184
weighted avg	0.94	0.94	0.93	1184

30) We were not giving enough exposure to the model on defaults to learn through the SMOTE therefore the results were not good. After doing the balancing on the same model through the SMOTE we get the different result which is better. For the new model the recall is now 95% for the train and 84% for the test. The precision for train set is 91% and for test set it is 62%.

Train:

	precision	recall	f1-score	support
0.0	0.95	0.91	0.93	2151
1.0	0.91	0.95	0.93	2151
accuracy			0.93	4302
macro avg	0.93	0.93	0.93	4302
weighted avg	0.93	0.93	0.93	4302

Test:

	precision	recall	f1-score	support
0.0	0.98	0.93	0.96	1047
1.0	0.62	0.84	0.71	137
accuracy			0.92	1184
macro avg	0.80	0.89	0.83	1184
weighted avg	0.94	0.92	0.93	1184

Conclusion

At last we were able to get a good recall value as well as precision without overfitting. In this case scenario, there were several challenges like outliers, missing values and correlated features.