FRA Project: Milestone-1 Business Report

By

Ritusri Mohan

CONTENTS

PROBLEM 1

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.

ABOUT THE DATASET

	Co_Code	Co_Name	Networth Next Year	Equity Paid Up	Networth	Capital Employed	Total Debt	Gross Block	Net Working Capital	Current Assets	...	PBIDTM (%) [Latest]	PBITM (%) [Latest]	PBDTM (%) [Latest]	CPM (%) [Latest]	APATM (%) [Latest]
0	16974	Hind.Cables	-8021.60	419.36	-7,027.48	-1,007.24	5,936.03	474.3	-1,076.34	40.5	...	0	0	0	0	0
1	21214	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	...	-10.3	-39.74	-57.74	-57.74	-87.18
2	14852	ABG Shipyard	-3192.58	53.84	506.86	7,714.68	6,944.54	1,281.54	4,496.25	9,097.64	...	-5,279.14	-5,516.98	-7,780.25	-7,723.67	-7,961.51
3	2439	GTL	-3054.51	157.3	-623.49	2,353.88	2,326.05	1,033.69	-2,612.42	1,034.12	...	-3.33	-7.21	-48.13	-47.7	-51.58
4	23505	Bharati Defence	-2967.36	50.3	-1,070.83	4,675.33	5,740.90	1,084.20	1,836.23	4,685.81	...	-295.55	-400.55	-845.88	379.79	274.79

5 rows × 67 columns

- The head of the dataset is shown above.
- On checking the shape of the dataset, it is found that there are 3586 rows and 67 columns.
- The messy column names were changed for easy use.
- The following 16 columns were dropped from the dataset as they were raw values:
  - Co_Name as call of the organisation may be identified from Company code as well.
  - Networth as ROG-Net_Worth_perc is not anything but percent of Value of a organisation as on 2015 - Current Year.
  - Capital_Employed as ROG-Capital_Employed_perc is not anything but percent of Total amount of capital used for the purchase of profits with the aid of using a organisation.

- Gross Block as ROG-Gross_Block_perc is percent of Total price of all the property that a organisation owns i.e. Gross Block.
- Gross Sales as ROG-Gross_Sales_perc is percent of The grand overall of sale transactions inside the accounting duration i.e., Gross Sales.
- Net_Sales as ROG-Net_Sales_perc is percent of Gross income minus returns, allowances, and discounts
- i.e. Net Sales.
- Cost_of_Production as ROG-Cost_of_Production_perc is percent of Costs incurred with the aid of using a enterprise from production a product or imparting a provider i.e. Cost_of_Production.
- PBIDT as ROG-PBIDT_perc is percent of Profit Before Interest, Depreciation & Taxes i.e., PBIDT.
- PBDT as ROG-PBDT_perc is percent of Profit Before Depreciation and Tax i.e., PBDT.
- PBIT as ROG-PBIT_perc is percent of Profit earlier than hobby and taxes i.e., PBIT.
- PBT as ROG-PBT_perc is percent of Profit earlier than tax i.e., PBT.
- PAT as ROG-PAT_perc is percent of Profit After Tax i.e., PAT.
- CP as ROG-CP_perc is percent of Commercial paper, a short-time period debt tool to satisfy short-time period liabilities. i.e CP.
- Revenue_earnings_in_forex as ROG-Revenue_earnings_in_forex_perc is percent of Revenue earned in overseas foreign money i.e.,Revenue_earnings_in_forex .
- Revenue_expenses_in_forex as ROG-Revenue_expenses_in_forex_perc is percent of Expenses because of overseas foreign money transactions i.e., Revenue_expenses_in_forex.
- Market_Capitalisation as ROG-Market_Capitalisation_perc is percent of Product of the overall variety of a organisation's incredible stocks and the contemporary marketplace charge of 1 percentage i.e., Market_Capitalisation.

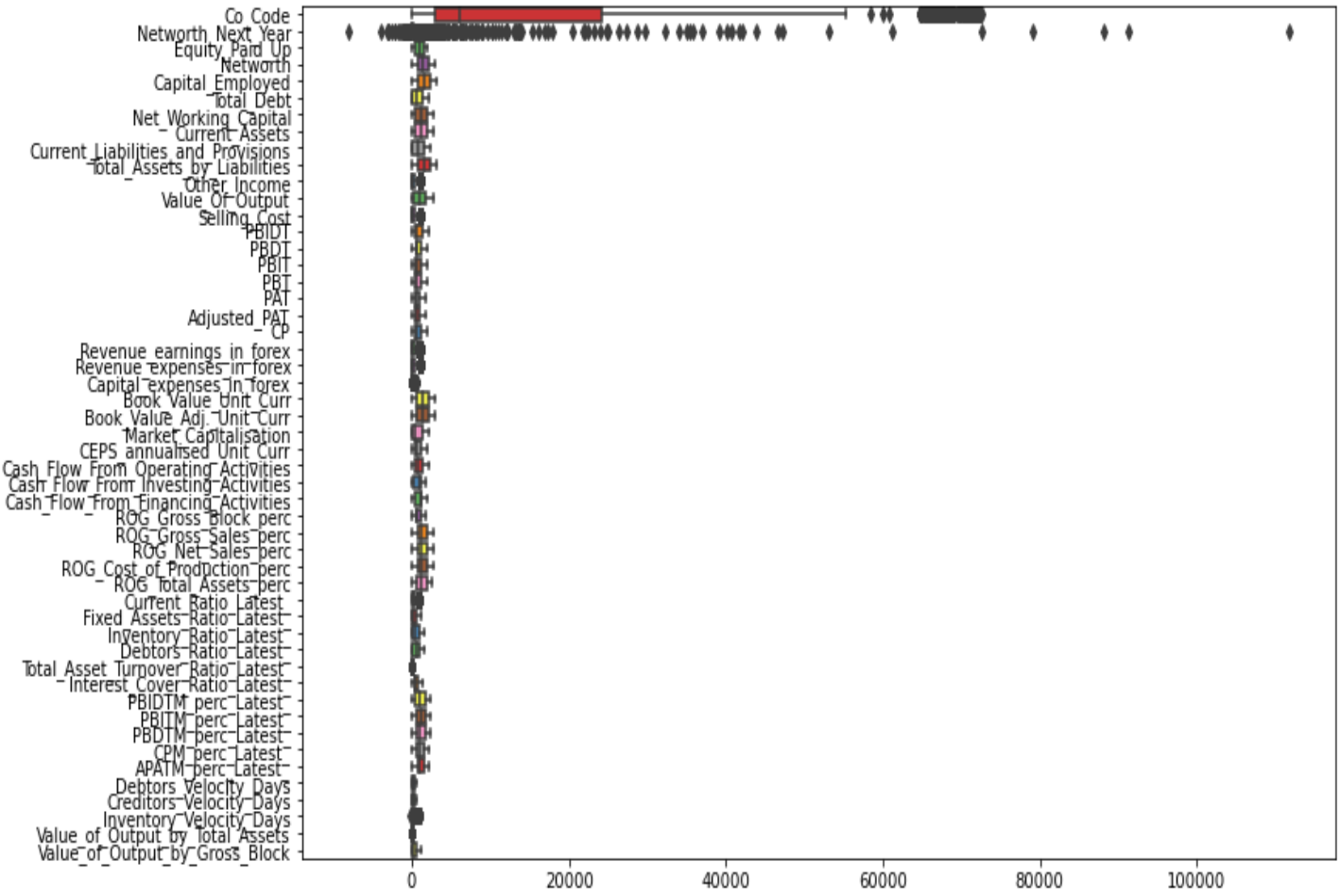
- No duplicate values were found in the dataset.
- After dropping the columns, dataset was found to have 3586 rows and 51 columns
- The columns with object data types were converted to int type and below is the descriptive summary of the dataset.

	count	mean	std	min	25%	50%	75%	max
Co_Code	3586	16065.39	19776.82	4	3029.25	6077.5	24269.5	72493
Networth_Next_Year	3586	725.05	4769.68	-8021.6	3.98	19.02	123.8	111729.1
Equity_Paid_Up	3586	963.22	604.3	0	399.25	1058	1468.75	2027
Networth	3586	1468.1	853.56	0	706	1462.5	2202.75	2980
Capital_Employed	3586	1527.7	899.62	0	739.25	1547	2294	3104
Total_Debt	3586	716.45	704.05	0	5	545.5	1328.75	2138
Net_Working_Capital	3586	1241.72	788.93	0	484.25	1205.5	1912.75	2698
Current_Assets	3586	1226.84	859.33	0	415.25	1192.5	1977.75	2774
Current_Liabilities_and_Provisions	3586	838.92	737.16	0	76.25	740.5	1486.75	2277
Total_Assets_by_Liabilities	3586	1543.45	918.73	0	745.25	1561.5	2319.75	3137
Other_Income	3586	237.34	320.1	0	10	53	422.75	1114
Value_Of_Output	3586	1060.56	851.33	0	193.25	983	1801.75	2654
Selling_Cost	3586	218.16	326.97	0	0	16	386.75	1112
PBIDT	3586	790.03	611.33	0	336	536	1304.75	2082
PBDT	3586	782.58	546.61	0	489	568.5	1195.75	1979
PBIT	3586	789.2	581.63	0	440	571	1262.75	2039
PBT	3586	759.78	508.45	0	470.25	635	1116.75	1888
PAT	3586	735.57	483.94	0	472.25	628.5	1069.75	1822
Adjusted_PAT	3586	725.19	486.18	0	429.25	634	1046.75	1815
CP	3586	750.73	530.64	0	469	543	1136.75	1932
Revenue_earnings_in_forex	3586	196.61	332.67	0	0	0	299.75	1156
Revenue_expenses_in_forex	3586	221.84	356.12	0	0	0	377.75	1206
Capital_expenses_in_forex	3586	38.41	103.54	0	0	0	0	490
Book_Value_Unit_Curr	3586	1475.19	876.21	0	677	1441.5	2235	3019
Book_Value_Adj._Unit_Curr	3586	1439.54	859.67	-1	660.25	1397.5	2186.75	2962
Market_Capitalisation	3586	678.61	699.72	0	0	473.5	1286.75	2106
CEPS_annualised_Unit_Curr	3586	766.75	526.91	0	464	582	1195.75	1899
Cash_Flow_From_Operating_Activities	3586	853.48	617.21	0	355.25	703	1344.75	2129
Cash_Flow_From_Investing_Activities	3586	830.13	534.97	0	271.25	1027.5	1210	1769
Cash_Flow_From_Financing_Activities	3586	926.98	562.65	0	425.25	1200	1245.75	1968
ROG_Gross_Block_perc	3586	784.95	464.85	0	556	580	1111	1826
ROG_Gross_Sales_perc	3586	1283.2	734.54	0	747.25	1143.5	1897.75	2652
ROG_Net_Sales_perc	3586	1279.94	732.6	0	748.25	1138.5	1894	2649
ROG_Cost_of_Production_perc	3586	1291.83	730.64	0	740.25	1177.5	1913	2649
ROG_Total_Assets_perc	3586	1237.08	736.42	0	631.25	1154	1876	2575
Current_Ratio_Latest_	3586	249.97	249.97	-1	88	136	388	980
Fixed_Assets_Ratio_Latest_	3586	328.16	352.03	-1	27	164.5	584.75	1198
Inventory_Ratio_Latest_	3586	514.77	504.85	-1	0	401.5	976	1496
Debtors_Ratio_Latest_	3586	574.38	491.33	-1	39.25	571	1002	1516
Total_Asset_Turnover_Ratio_Latest_	3585	1.24	2.67	0	0.07	0.6	1.55	57.75
Interest_Cover_Ratio_Latest_	3586	583.88	344.73	-1	372	471	822	1431
PBIDTM_perc_Latest_	3586	1125.01	675.97	-1	453	1059.5	1720.75	2369
PBITM_perc_Latest_	3586	1131.02	642.02	-1	575	1078.5	1676.75	2311
PBDTM_perc_Latest_	3586	1144.83	645.68	-1	619	1072.5	1718.75	2326
CPM_perc_Latest_	3586	1086.44	602.03	-1	608	1016	1610	2192
APATM_perc_Latest_	3586	1046.47	545.06	-1	754	911.5	1502	2106
Debtors_Velocity_Days	3586	249.91	194.35	0	60.25	255	425.75	572
Creditors_Velocity_Days	3586	228.04	171.88	0	60	237	388	516
Inventory_Velocity_Days	3483	79.64	137.85	-199	0	35	96	996
Value_of_Output_by_Total_Assets	3586	0.82	1.2	-0.33	0.07	0.48	1.16	17.63
Value_of_Output_by_Gross_Block	3586	346.93	353	0	46	181.5	607.75	1209

From the summary of the dataset above, it is seen that there is difference in the scale of the values across the columns. There is a chance for the outliers to be present as there is a huge difference between the 75% value and max value for some variables.

1.1 Outlier Treatment

A visual representation off the outliers can be seen below.



The IQR approach was applied to identify outliers which involves creating a new range known as a decision range, and any data point that lies outside of it is regarded as an outlier. The scope is provided below.

$$IQR = Q3 - Q1$$
$$\text{Lower Bound} = Q1 - 1.5*IQR$$
$$\text{Upper Bound} = Q3 + 1.5*IQR$$

The number of outliers in are shown below

```
Co Code                                  291
Networth_Next_Year                       676
Equity Paid Up                             0
Networth                                   0
Capital Employed                           0
Total_Debt                                 0
Net Working Capital                        0
Current_Assets                             0
Current Liabilities and Provisions         0
Total Assets by Liabilities                0
Other Income                              79
Value Of Output                            0
Selling Cost                             168
PBIDT                                      0
PBDT                                       0
PBIT                                       0
PBT                                        0
PAT                                        0
Adjusted_PAT                               0
CP                                         0
Revenue_earnings_in_forex                418
Revenue expenses in forex                278
Capital expenses in forex                694
Book Value Unit Curr                       0
Book Value Adj. Unit Curr                  0
Market Capitalisation                      0
CEPS annualised Unit Curr                  0
Cash_Flow_From_Operating_Activities        0
Cash Flow From Investing Activities        0
Cash_Flow_From_Financing_Activities        0
ROG Gross Block perc                       0
ROG_Gross_Sales_perc                       0
ROG Net Sales perc                         0
ROG_Cost_of_Production_perc                0
ROG Total Assets perc                      0
Current Ratio Latest                     160
Fixed Assets Ratio Latest                  0
Inventory Ratio Latest                     0
Debtors_Ratio_Latest_                      0
Total Asset Turnover Ratio Latest        201
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_Velocity_Days                      0
Creditors Velocity Days                    0
Inventory Velocity Days                  262
Value of Output by Total Assets          150
Value of Output by Gross Block             0
dtype: int64
```

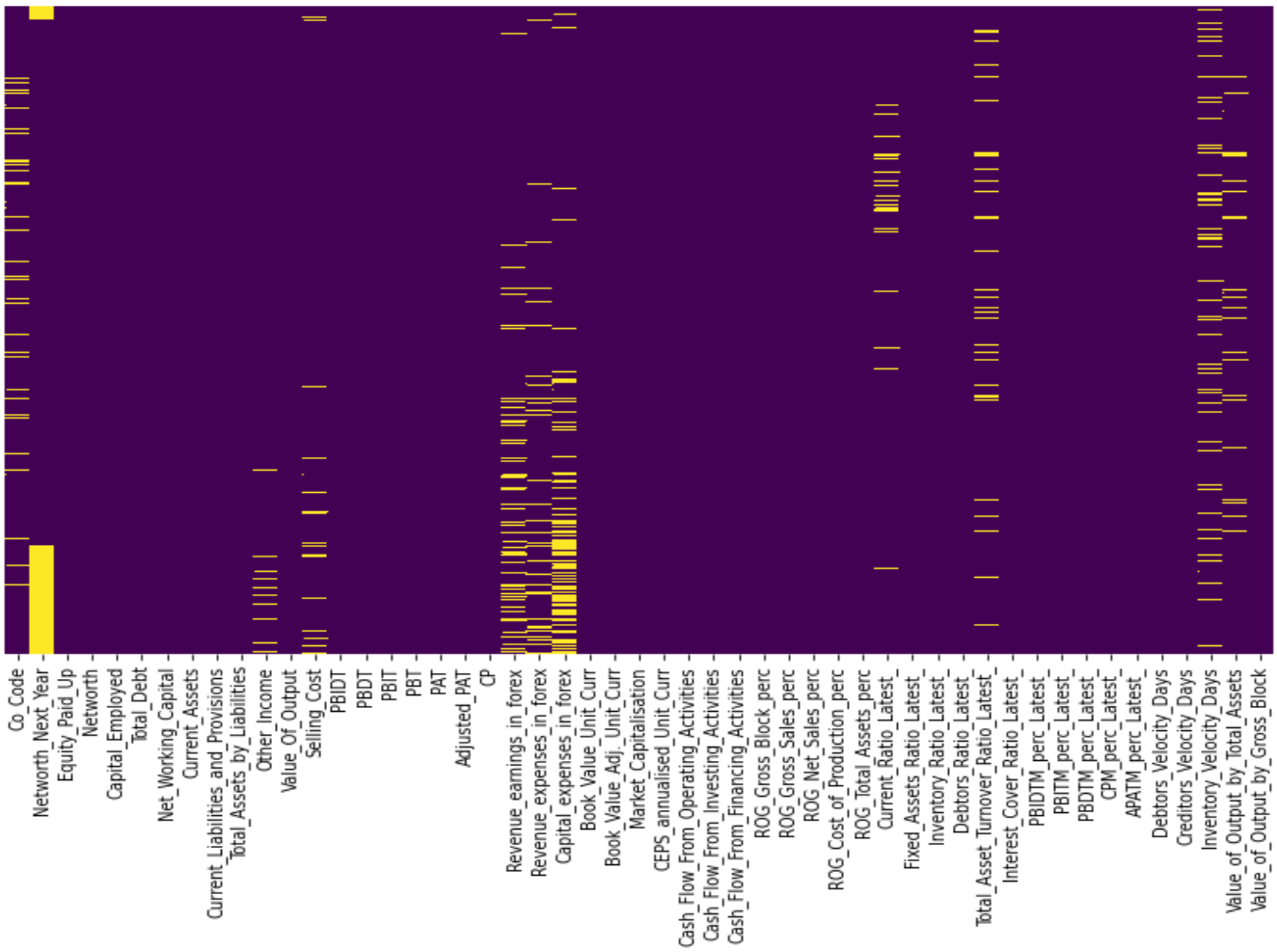
The outliers were replaced using the Nan values :

```
Company[((Company> UL) | (Company< LL))]= np.nan
```

```
Co_Code                                  291
Networth_Next_Year                       676
Equity_Paid_Up                             0
Networth                                   0
Capital Employed                           0
Total Debt                                 0
Net Working Capital                        0
Current Assets                             0
Current Liabilities and Provisions         0
Total_Assets_by_Liabilities                0
Other_Income                              79
Value_Of_Output                            0
Selling_Cost                             168
PBIDT                                      0
PBDT                                       0
PBIT                                       0
PBT                                        0
PAT                                        0
Adjusted_PAT                               0
CP                                         0
Revenue_earnings_in_forex                418
Revenue expenses in forex                278
Capital expenses in forex                694
Book Value Unit Curr                       0
Book Value Adj. Unit Curr                  0
Market Capitalisation                      0
CEPS_annualised_Unit_Curr                  0
Cash_Flow_From_Operating_Activities        0
Cash_Flow_From_Investing_Activities        0
Cash Flow From Financing Activities        0
ROG Gross Block perc                       0
ROG Gross Sales perc                       0
ROG Net Sales perc                         0
ROG Cost of Production perc                0
ROG Total Assets perc                      0
Current_Ratio_Latest_                    160
Fixed_Assets_Ratio_Latest_                 0
Inventory_Ratio_Latest_                    0
Debtors Ratio Latest                       0
Total Asset Turnover Ratio Latest        202
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 Velocity Days                      0
Creditors Velocity Days                    0
Inventory Velocity Days                  365
Value of Output by Total Assets          150
Value of Output by Gross Block             0
dtype: int64
```

1.2 Missing Value Treatment?

The number of missing values after treating the outliers was found to be 3481.The visual representation of the missing values can be seen below:



- Some variables have noticeable missing values, which can be seen. Yellow signifies missing values in the data, while the heatmap's purple colour denotes occupied cells.
- Some values for the variable "Networth Next Year" may be entirely absent.
- The variables "ROG-Revenue expenses in forex (percent)" and "Revenue expenses in forex" are missing their maximum values (which is expected, since ROG is the percentage represtation of of revenue values). Additionally, there are some missing values in the fields "Inventory Velocity (Days)," "Debtors Ratio[Latest]","Capital expenses in forex," "Selling cost," and "Other Income."

By dividing the total number of missing values by the total number of suitable rows,the proportion of missing data was ranked. Notably, none of the columns were eliminated as the value is not more than 30%.

```
Capital expenses in forex              0.19
Networth Next Year                     0.19
Revenue earnings in forex              0.12
Inventory_Velocity_Days                0.10
Co_Code                                0.08
Revenue_expenses_in_forex              0.08
Total Asset Turnover Ratio Latest      0.06
Selling Cost                           0.05
Current Ratio Latest                   0.04
Value of Output by Total Assets        0.04
Other Income                           0.02
Current_Liabilities_and_Provisions     0.00
Total_Assets_by_Liabilities            0.00
Networth                               0.00
Value_Of_Output                        0.00
Current Assets                         0.00
Total Debt                             0.00
PBIDT                                  0.00
PBDT                                   0.00
PBIT                                   0.00
PBT                                    0.00
PAT                                    0.00
Adjusted_PAT                           0.00
CP                                     0.00
Equity Paid Up                         0.00
Net Working Capital                    0.00
Capital Employed                       0.00
Value of Output by Gross Block         0.00
Book_Value_Unit_Curr                   0.00
Inventory_Ratio_Latest_                0.00
Creditors_Velocity_Days                0.00
Debtors_Velocity_Days                  0.00
APATM perc Latest                      0.00
CPM perc Latest                        0.00
PBDTM perc Latest                      0.00
PBITM perc Latest                      0.00
PBIDTM perc Latest                     0.00
Interest_Cover_Ratio_Latest_           0.00
Debtors_Ratio_Latest_                  0.00
Fixed_Assets_Ratio_Latest_             0.00
Book Value Adj. Unit Curr              0.00
ROG Total Assets perc                  0.00
ROG Cost of Production perc            0.00
ROG Net Sales perc                     0.00
ROG Gross Sales perc                   0.00
ROG Gross Block perc                   0.00
Cash_Flow_From_Financing_Activities    0.00
Cash_Flow_From_Investing_Activities    0.00
Cash_Flow_From_Operating_Activities    0.00
CEPS annualised Unit Curr              0.00
Market Capitalisation                  0.00
dtype: float64
```

- The data was split into ‘response’ and ‘predictors’ variables and Scikit-Learn’s Standard Scaler method was used which scaled our variables to have values between 0 and 1.
- Since the missing values were of numeric type, KNN Imputer was used. This imputer utilizes the k-Nearest Neighbors method to replace the missing values in the datasets by finding the nearest neighbors with the Euclidean distance matrix.
- The missing value was imputed by making predictions based on the values of the variable’s 10 nearest neighbours, replacing all missing values with those values.
- Below we can see that there are no missing values.

```
Co_Code                                  0
Networth_Next_Year                       0
Equity_Paid_Up                           0
Networth                                 0
Capital Employed                         0
Total Debt                               0
Net Working Capital                      0
Current Assets                           0
Current Liabilities and Provisions       0
Total_Assets_by_Liabilities              0
Other_Income                             0
Value_Of_Output                          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                0
Market Capitalisation                    0
CEPS_annualised_Unit_Curr                0
Cash_Flow_From_Operating_Activities      0
Cash_Flow_From_Investing_Activities      0
Cash Flow From Financing Activities      0
ROG Gross Block perc                     0
ROG Gross Sales perc                     0
ROG Net Sales perc                       0
ROG Cost of Production perc              0
ROG Total Assets perc                    0
Current_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 Velocity Days                    0
Creditors Velocity Days                  0
Inventory Velocity Days                  0
Value of Output by Total Assets          0
Value of Output by Gross Block           0
dtype: int64
```

1.3 Transform Target variable into 0 and 1

- The target variable was not defined, but since the goal was to build a model for the investor to decipher the company in which it invests, the variable Networth_Next_Year was used to convert it to the target variable.
- A default variable was created that took the following values:
  - 1 if next year's net worth is negative, which means company is unlikely to return a reasonable investment to the investor and transform as Default.
  - 0 if next year's net worth is positive, which means the company would continue to return the appropriate investment to the investor and thus be transformed as Non-default.

	default	Networth_Next_Year
0	0	3.65
1	0	5.22
2	0	32.47
3	1	-24.03
4	1	-18.12
5	1	-6.58
6	0	7.64
7	0	6.10
8	1	-28.47
9	1	-1.13

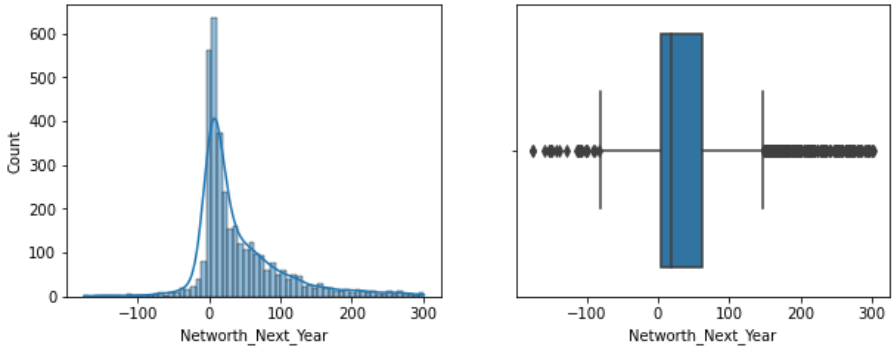
On checking the proportion of default it was found that about 10% of the companies in the dataset are likely to default, and these are the companies that investors are unlikely to invest in.

```
0   0.90
1   0.10
Name: default, dtype: float64
```

1.4 Univariate (4 marks) & Bivariate (6marks) analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building)

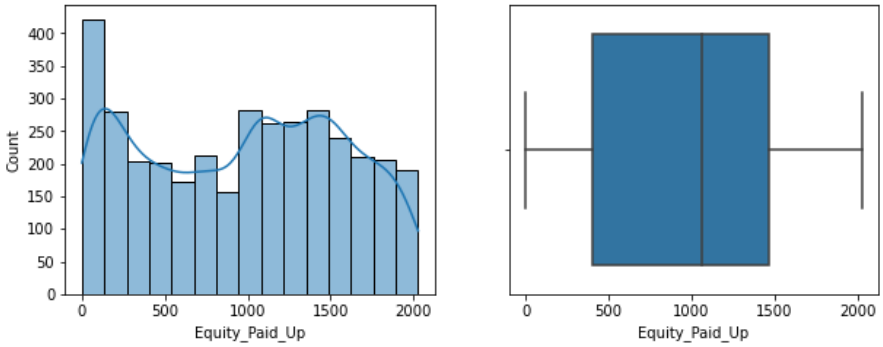
Univariate Analysis

- Networth Next Year



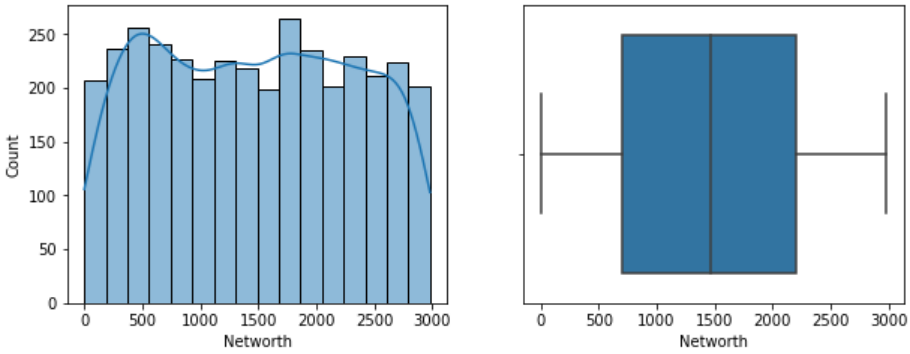
The distribution seems to be right skewed and there are still outliers present.

- Equity paid up



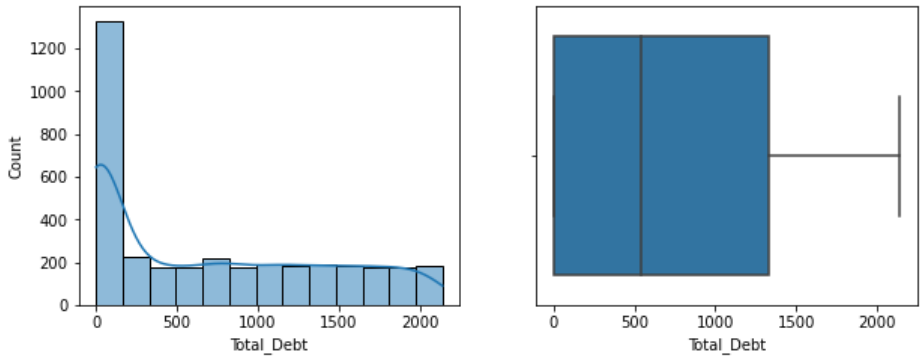
The distribution seems to be symmetrical and it look like 50% of the times, Equity paid up i.e., amount that has been received by the company through the issue of shares to the shareholders is in positive.

- Networth



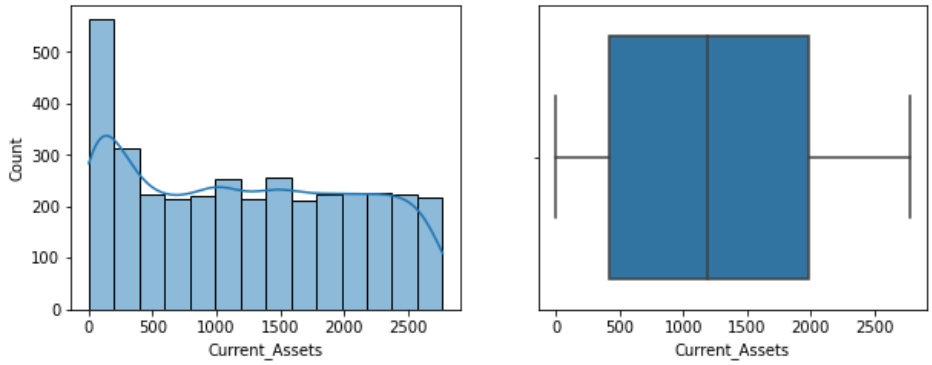
The distribution seems to be normal and there are no outliers present.

• Total Debt



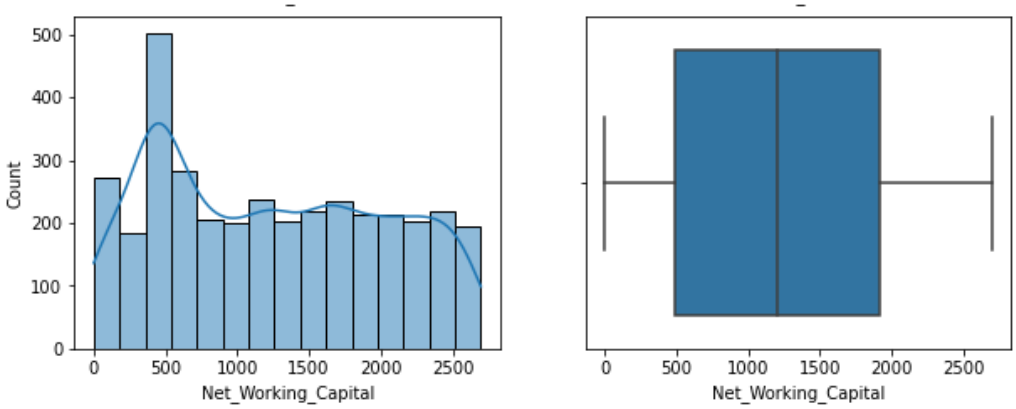
There is one company which is borrowed a highest sum from market which is a little more than 1200 units, and some companies have taken debt of around 500-750. Most if the companies have debt lower than 250. Companies with high debt may deploy high risk of not making profit next year.

• Current Assets



The distribution seems to be moderately skewed positively without outliers.

• Net working capital



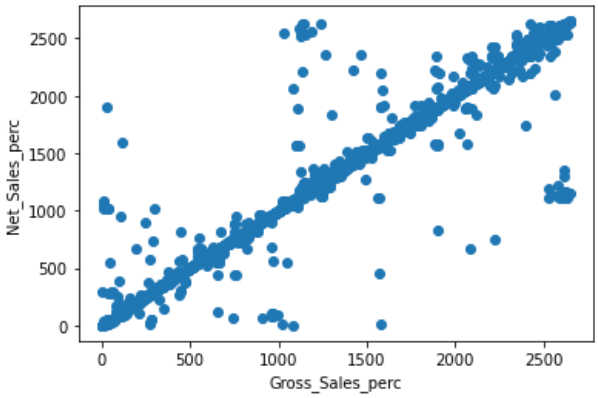
The distribution seems to be moderately skewed positively without outliers.

Skew values of various columns:

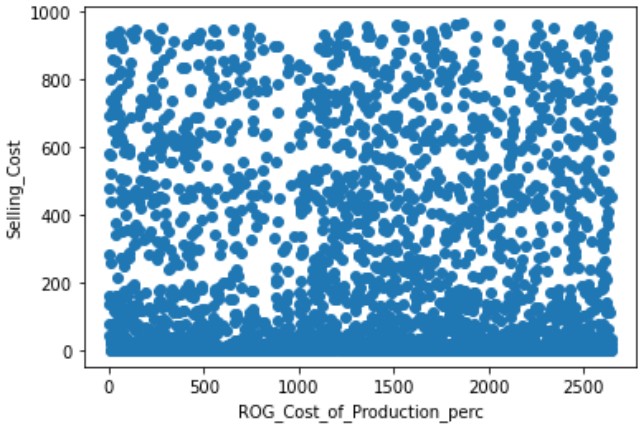
```
Revenue earnings in forex                 1.87
Revenue expenses in forex                 1.60
Networth_Next_Year                        1.59
Selling Cost                              1.40
Co_Code                                   1.36
Other Income                              1.33
Inventory Velocity Days                   1.27
Current_Ratio_Latest_                     1.26
Total Asset Turnover Ratio Latest         1.06
Value_of_Output_by_Total_Assets           0.90
Fixed Assets Ratio Latest                 0.89
Value of Output by Gross Block            0.89
Interest_Cover_Ratio_Latest_              0.74
CP                                        0.60
PBIDT                                     0.58
ROG Gross Block perc                      0.56
Market_Capitalisation                     0.56
PBIT                                      0.55
PBDT                                      0.54
Total_Debt                                0.51
CEPS annualised Unit Curr                 0.50
PBT                                       0.44
PAT                                       0.43
Adjusted PAT                              0.42
Inventory_Ratio_Latest_                   0.40
Current Liabilities and Provisions        0.38
Cash_Flow_From_Operating_Activities       0.37
Value Of Output                           0.27
Debtors Ratio Latest                      0.23
PBIDTM_perc_Latest_                       0.20
Net Working Capital                       0.18
PBDTM_perc_Latest_                        0.16
PBITM perc Latest                         0.15
APATM perc Latest                         0.14
CPM perc Latest                           0.14
ROG Net Sales perc                        0.13
ROG_Gross_Sales_perc                      0.13
Current Assets                            0.12
ROG Cost of Production perc               0.12
Book Value Adj. Unit Curr                 0.11
Book Value Unit Curr                      0.10
ROG_Total_Assets_perc                     0.07
Debtors Velocity Days                     0.06
Networth                                  0.03
Creditors Velocity Days                   0.01
Capital Employed                          0.00
Capital_expenses_in_forex                 0.00
Total Assets by Liabilities              -0.01
Equity_Paid_Up                           -0.05
Cash Flow From Financing Activities      -0.26
Cash Flow From Investing Activities      -0.27
dtype: float64
```

Bivariate Analysis

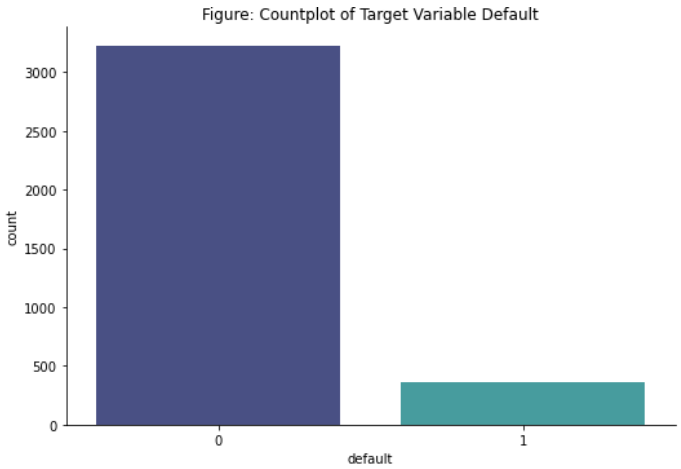
Scatterplot between Gross Sales and Net Sales shows direct relationship between them.



There is no proper relationship between ROG_Cost_of_production_perc and Selling_Cost.



The data has higher Non-default companies i.e., the companies which are expected to have a positive Net Worth next year (which is good for investors for decision making).



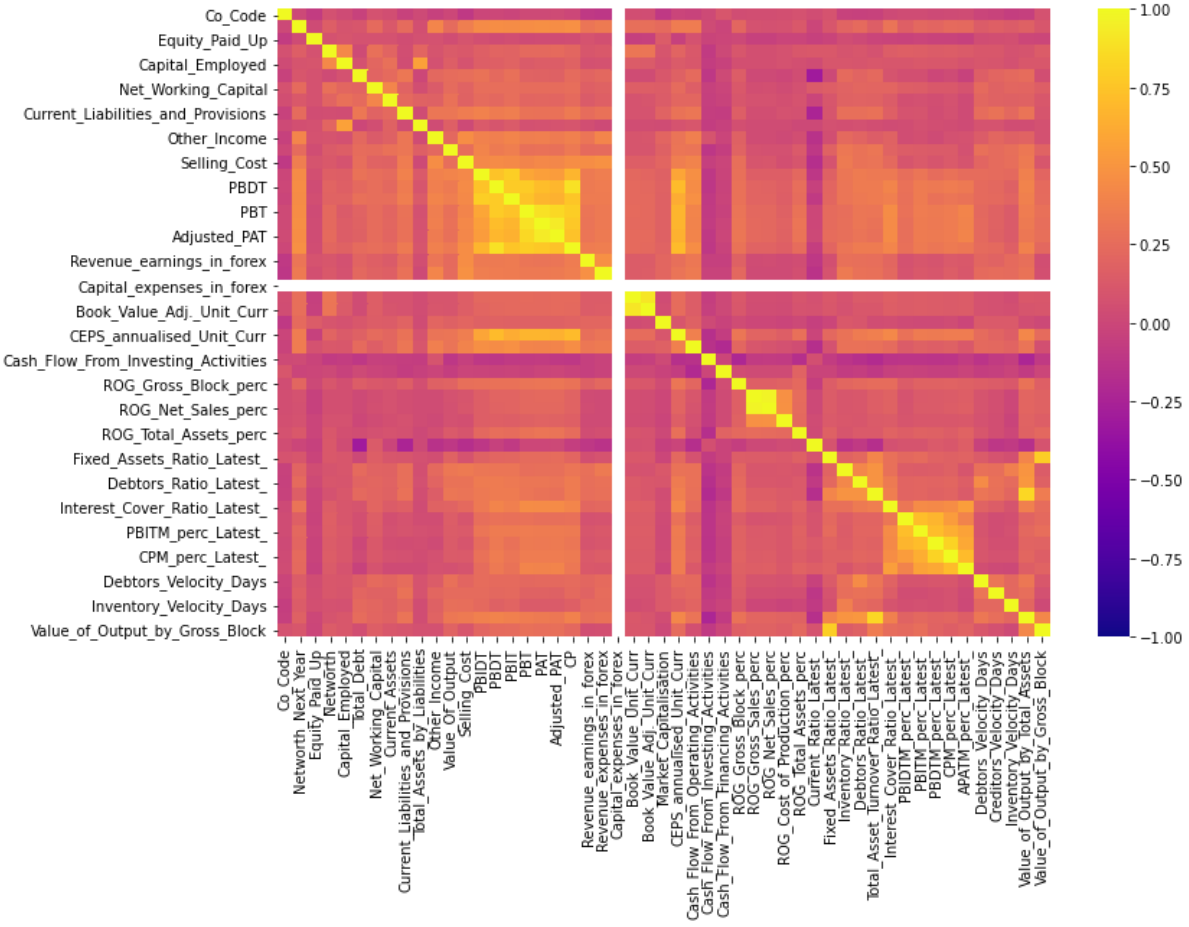
Multivariate Analysis

- Correlation Matrix

  It gives the values of correlation of between variables. Below is only a snippet of the matrix. Refer to the Jupyter notebook for the entire matrix.

	Co_Code	Networth_Next_Year	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Net_Working_Capital	C
Co_Code	1.00	0.01	-0.07	0.06	0.03	-0.03	0.08	
Networth_Next_Year	0.01	1.00	0.03	0.16	0.09	0.11	0.24	
Equity_Paid_Up	-0.07	0.03	1.00	0.08	0.10	0.07	0.07	
Networth	0.06	0.16	0.08	1.00	0.43	0.02	0.27	
Capital_Employed	0.03	0.09	0.10	0.43	1.00	0.08	0.21	
Total_Debt	-0.03	0.11	0.07	0.02	0.08	1.00	0.20	
Net_Working_Capital	0.08	0.24	0.07	0.27	0.21	0.20	1.00	

- Heatmap



- There is weak correlation between most of the variables.

1.5 Train Test Split

The data was split into predictors and response variables in the ratio of 67:33 and random_state of 42 was used and Logistic regression was applied.

```
X_train, X_test, y_train, y_test = train_test_split(predictors, response,
                                                    test_size = 0.33, random_state = 42)
```

X_train head:

	Co_Code	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Net_Working_Capital	Current_Assets	Current_Liabilities_and_Provisions
662	5271.00	1058.00	1302.00	1197.00	2.00	566.00	232.00	4.00
1373	5723.00	739.00	2916.00	3008.00	2.00	1225.00	1026.00	39.00
3268	27272.00	970.00	2740.00	1125.00	161.00	499.00	165.00	1402.00
3246	737.00	1211.00	2413.00	308.00	128.00	2562.00	134.00	1456.00
1456	6863.50	765.00	2773.00	821.00	1751.00	2535.00	2699.00	90.00

5 rows × 50 columns

y_train head:

	default
662	0
1373	0
3268	0
3246	0
1456	0

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 was applied using the statsmodel library.
Assumptions for logistic regression:

- It assumes that there is minimal, or no multi-collinearity among the independent variables.
- It assumes that independent variables are linearly related to log of odds.
- It assumes a large sample for good prediction.
- It assumes that the observations are independent of each other.
- There are no influential values(outliers) in the continuous predictors (independent variables).

Statsmodels provides a Logit() function for performing logistic regression. The Logit() function accepts y and X as parameters and returns the Logit object. The model is then fitted to the data. The logit function is simply the logarithm of the odds.

$$\text{logit}(x) = \log(x / (1 - x))$$

The train data was concatenated into a single set and test data was concatenated into another set followed by a check on the balance of the data. It was found that balance value was 0.095.

Multicollinearity was checked before proceeding to build the model. Multicollinearity occurs when two or more independent variables are highly correlated in the regression model.

The Variance Inflation Factor method was used for this problem. It was seen that the value of VIF is high for many variables. Hence, variables with VIF more than 5 (very high correlation) were dropped and the model was built.

VIF VALUES

	variables	VIF>5
31	ROG_Net_Sales_perc	80.59
30	ROG_Gross_Sales_perc	80.34
16	PAT	37.01
17	Adjusted_PAT	28.61
22	Book_Value_Unit_Curr	21.77
23	Book_Value_Adj._Unit_Curr	21.75
15	PBT	20.94
13	PBDT	20.72
18	CP	17.79
42	PBDTM_perc_Latest_	17.42
43	CPM_perc_Latest_	16.76
14	PBIT	13.31
44	APATM_perc_Latest_	12.45
41	PBITM_perc_Latest_	12.38
12	PBIDT	10.31
40	PBIDTM_perc_Latest_	10.11
48	Value_of_Output_by_Total_Assets	8.28
25	CEPS_annualised_Unit_Curr	8.16
3	Capital_Employed	7.31
38	Total_Asset_Turnover_Ratio_Latest_	7.2
39	Interest_Cover_Ratio_Latest_	6.91
49	Value_of_Output_by_Gross_Block	6.01
2	Networth	6
8	Total_Assets_by_Liabilities	5.85
35	Fixed_Assets_Ratio_Latest_	5.66
32	ROG_Cost_of_Production_perc	5.51
5	Net_Working_Capital	5.16
33	ROG_Total_Assets_perc	5.15

Selected for Model building

	variables	VIF< 5
29	ROG_Gross_Block_perc	4.86
26	Cash_Flow_From_Operating_Activities	4.51
6	Current_Assets	4.36
37	Debtors_Ratio_Latest_	4.1
28	Cash_Flow_From_Financing_Activities	3.82
45	Debtors_Velocity_Days	3.69
10	Value_Of_Output	3.63
1	Equity_Paid_Up	3.49
46	Creditors_Velocity_Days	3.48
7	Current_Liabilities_and_Provisions	3.39
36	Inventory_Ratio_Latest_	3.3
27	Cash_Flow_From_Investing_Activities	3.19
34	Current_Ratio_Latest_	2.8
4	Total_Debt	2.7
11	Selling_Cost	2.63
9	Other_Income	2.47
24	Market_Capitalisation	2.31
20	Revenue_expenses_in_forex	2.3
47	Inventory_Velocity_Days	2.26
0	Co_Code	2.1
19	Revenue_earnings_in_forex	2.03
21	Capital_expenses_in_forex	nan

Model 1

The following variables were used and written in the following manner to build the formula.

default~ Revenue_earnings_in_forex+Revenue_expenses_in_forex + ROG_Gross_Block_perc + Current_Ratio_Latest_+ Creditors_Velocity_Days + Inventory_Ratio_Latest_ + Inventory_Velocity_Days +Debtors_Velocity_Days + Debtors_Ratio_Latest_ + Cash_Flow_From_Financing_Activities + Capital_expenses_in_forex + Equity_Paid_Up + Selling_Cost + Other_Income +Cash_Flow_From_Investing_Activities + Market_Capitalisation + Total_Debt

In the P|z| column of the Logistic Regression Results table it can be seen that 10 variables have p value > 0.05 which mean that they are insignificant in nature and so, this model is not considered. Due to this Model 2 was built by dropping these variables.

Logit Regression Results

Dep. Variable:	default	**No. Observations:**	3586
Model:	Logit	**Df Residuals:**	3569
Method:	MLE	**Df Model:**	16
Date:	Fri, 08 Jul 2022	**Pseudo R-squ.:**	0.2266
Time:	20:31:22	**Log-Likelihood:**	-909.10
converged:	False	**LL-Null:**	-1175.4
Covariance Type:	nonrobust	**LLR p-value:**	4.350e-103

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6182	0.260	-2.377	0.017	-1.128	-0.108
Revenue_earnings_in_forex	-0.0006	0.000	-1.375	0.169	-0.002	0.000
Revenue_expenses_in_forex	-0.0004	0.000	-1.244	0.213	-0.001	0.000
ROG_Gross_Block_perc	-0.0008	0.000	-4.870	0.000	-0.001	-0.000
Current_Ratio_Latest_	-0.0090	0.001	-11.414	0.000	-0.011	-0.007
Creditors_Velocity_Days	0.0004	0.000	0.969	0.333	-0.000	0.001
Inventory_Ratio_Latest_	-0.0002	0.000	-1.361	0.173	-0.001	9.3e-05
Inventory_Velocity_Days	0.0012	0.001	0.950	0.342	-0.001	0.004
Debtors_Velocity_Days	-3.71e-05	0.000	-0.100	0.921	-0.001	0.001
Debtors_Ratio_Latest_	-0.0006	0.000	-3.832	0.000	-0.001	-0.000
Cash_Flow_From_Financing_Activities	-0.0001	0.000	-1.109	0.267	-0.000	9.34e-05
Capital_expenses_in_forex	0	2.07e+04	0	1.000	-4.05e+04	4.05e+04
Equity_Paid_Up	0.0002	0.000	1.835	0.066	-1.26e-05	0.000
Selling_Cost	-0.0006	0.000	-1.744	0.081	-0.001	6.87e-05
Other_Income	-0.0006	0.000	-2.207	0.027	-0.001	-6.27e-05
Cash_Flow_From_Investing_Activities	0.0005	0.000	4.162	0.000	0.000	0.001
Market_Capitalisation	-0.0002	9.52e-05	-2.522	0.012	-0.000	-5.36e-05
Total_Debt	0.0005	9.14e-05	5.317	0.000	0.000	0.001

Model 2

The following variables were used and written in the following manner to build the formula.

default~ Revenue_expenses_in_forex+ROG_Gross_Block_perc + Current_Ratio_Latest_ + Debtors_Ratio_Latest_ + Other_Income+ Cash_Flow_From_Investing_Activities + Market_Capitalisation + Total_Debt+Cash_Flow_From_Operating_Activities+Current_Assets+Debtors_ Ratio_Latest_+Cash_Flow_From_Financing_Activities+Debtors_Velocity_Days +Value_Of_Output'

In the P|z| column of the Logistic Regression Results table it can be seen that 5 variables have p value > 0.05 which mean that they are insignificant in nature and so, this model is not considered. Due to this Model 3 was built by dropping these variables.

Dep. Variable:	default	**No. Observations:**	3586
Model:	Logit	**Df Residuals:**	3572
Method:	MLE	**Df Model:**	13
Date:	Fri, 08 Jul 2022	**Pseudo R-squ.:**	0.2339
Time:	20:31:23	**Log-Likelihood:**	-900.50
converged:	False	**LL-Null:**	-1175.4
Covariance Type:	nonrobust	**LLR p-value:**	3.831e-109

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.0270	0.257	0.105	0.917	-0.477	0.531
Revenue_expenses_in_forex	-0.0005	0.000	-1.585	0.113	-0.001	0.000
ROG_Gross_Block_perc	-0.0008	0.000	-4.928	0.000	-0.001	-0.000
Current_Ratio_Latest_	-0.0087	0.001	-11.283	0.000	-0.010	-0.007
Debtors_Ratio_Latest_	-0.0006	0.000	-3.869	0.000	-0.001	-0.000
Other_Income	-0.0004	0.000	-1.504	0.133	-0.001	0.000
Cash_Flow_From_Investing_Activities	0.0005	0.000	3.571	0.000	0.000	0.001
Market_Capitalisation	-0.0002	9.52e-05	-1.791	0.073	-0.000	1.6e-05
Total_Debt	0.0006	9.44e-05	6.380	0.000	0.000	0.001
Cash_Flow_From_Operating_Activities	-0.0004	0.000	-3.611	0.000	-0.001	-0.000
Current_Assets	-0.0001	7.99e-05	-1.297	0.195	-0.000	5.3e-05
Cash_Flow_From_Financing_Activities	-0.0002	0.000	-1.990	0.047	-0.000	-3.39e-06
Debtors_Velocity_Days	0.0004	0.000	1.042	0.297	-0.000	0.001
Value_Of_Output	-0.0002	8.69e-05	-2.872	0.004	-0.000	-7.93e-05

Model 3

The following variables were used and written in the following manner to build the formula.

default~ ROG_Gross_Block_perc + Current_Ratio_Latest_ + Debtors_Ratio_Latest_+ Cash_Flow_From_Investing_Activities + Total_Debt+Cash_Flow_From_Operating_Activities+Debtors_Ratio_Latest_+ Cash_Flow_From_Financing_Activities+Value_Of_Output

In the P|z| column of the Logistic Regression Results table it can be seen that all variables have p value < 0.05 which mean that they are significant in nature and so, this model is chosen.

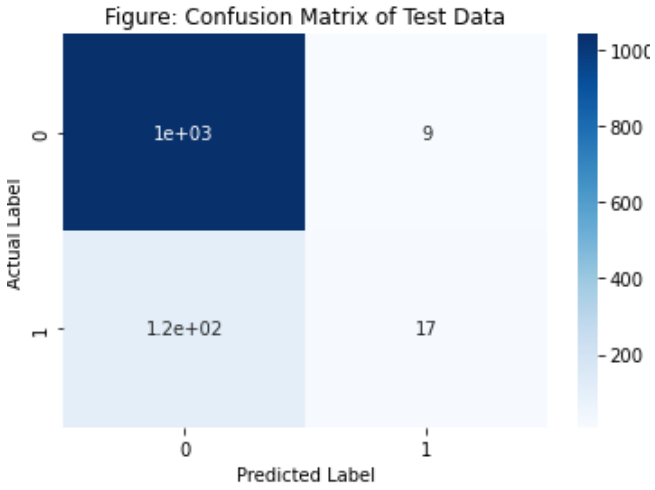
Dep. Variable:	default	**No. Observations:**	3586
Model:	Logit	**Df Residuals:**	3577
Method:	MLE	**Df Model:**	8
Date:	Fri, 08 Jul 2022	**Pseudo R-squ.:**	0.2278
Time:	20:31:23	**Log-Likelihood:**	-907.62
converged:	True	**LL-Null:**	-1175.4
Covariance Type:	nonrobust	**LLR p-value:**	1.669e-110

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.0876	0.250	0.351	0.726	-0.402	0.577
ROG_Gross_Block_perc	-0.0008	0.000	-5.190	0.000	-0.001	-0.001
Current_Ratio_Latest_	-0.0087	0.001	-11.432	0.000	-0.010	-0.007
Debtors_Ratio_Latest_	-0.0006	0.000	-4.293	0.000	-0.001	-0.000
Cash_Flow_From_Investing_Activities	0.0004	0.000	3.087	0.002	0.000	0.001
Total_Debt	0.0005	9e-05	5.850	0.000	0.000	0.001
Cash_Flow_From_Operating_Activities	-0.0006	0.000	-4.950	0.000	-0.001	-0.000
Cash_Flow_From_Financing_Activities	-0.0002	0.000	-2.195	0.028	-0.000	-2.62e-05
Value_Of_Output	-0.0003	8.2e-05	-3.823	0.000	-0.000	-0.000

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

Validation of Model_3 on Test Data

- Confusion Matrix



- True negatives (TN) = 1000
- True positives (TP) = 17
- False negatives (FN) = 120
- False positives (FP) = 9

- Classification Report

	precision	recall	f1-score	support
0	0.900	0.991	0.943	1051
1	0.654	0.128	0.214	133
accuracy			0.894	1184
macro avg	0.777	0.560	0.579	1184
weighted avg	0.872	0.894	0.861	1184

- Out of all the companies that the model predicted to transform as Non-default in 2016, only 90% actually did.

- Out of all the companies that actually did get transformed as Non-default, the model only predicted this outcome correctly for 99.1% of those companies.

- Since f1-score of 0.943 is very close to 1, it tells us that the model does a very good job of predicting whether or not the companies will default.

THE END