Financial Risk Analysis

Milestone 1

Project Report

Mr. Akash Kamble

June_A Batch

Program: DSBA

❖ Table of Contents:

❖ List of Figures:

❖ List of Tables:

• Predicting the Defaulters

Executive Summary

Intend of the study is to use the concept of linear regression technique to detect the companies have high probability to default. Prediction of the defaulters is an important analysis performed by the money lending institutes to limit the credit risks..

Introduction

We will use various parameters/ratios derived from the financial statements of the companies to predict the defaulters.

Data Description (only for Top 12 important parameters)

1. **Equity Paid Up:** Amount that has been received by the company through the issue of shares to the shareholders
2. **Networth:** Value of a company as on 2015 - Current Year
3. **Current liabilities and provisions:** Short-term financial obligations that are due within one year (includes amount that is set aside cover a future liability)
4. **Other Income:** Income realized from non-business activities (e.g. sale of long term asset)
5. **CEPS (annualised) (Unit Curr):** Cash Earnings per Share, profitability ratio that measures the financial performance of a company by calculating cash flows on a per share basis
6. **ROG-Net Worth (%):** Rate of Growth - Networth
7. **ROG-Capital Employed (%):** Rate of Growth - Capital Employed
8. **ROG-Total Assets (%):** Rate of Growth - Total Assets
9. **Current Ratio[Latest]:** Liquidity ratio, company's ability to pay short-term obligations or those due within one year
10. **Fixed Assets Ratio[Latest]:** Solvency ratio, the capacity of a company to discharge its obligations towards long-term lenders indicating
11. **Interest Cover Ratio[Latest]:** Determines how easily a company can pay interest on its outstanding debt
12. **Value of Output/Gross Block:** Ratio of Value of Output (market value) to Gross Block

Sample of 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]	Debtors Velocity (Days)	Creditors Velocity (Days)	Inventory Velocity (Days)	Value of Output/Total Assets	Value of Output/Gross Block
0	16974	Hind.Cables	-8021.60	419.36	-7027.48	-1007.24	5936.03	474.30	-1076.34	40.50	...	0.00	0.00	0.00	0.00	0.00	0	0	45.0	0.00	0.00
1	21214	Tata Tele. Mah.	-3986.19	1954.93	-2968.08	4458.20	7410.18	9070.86	-1098.88	486.86	...	-10.30	-39.74	-57.74	-57.74	-87.18	29	101	2.0	0.31	0.24
2	14852	ABG Shipyard	-3192.58	53.84	506.86	7714.68	6944.54	1281.54	4496.25	9097.64	...	-5279.14	-5516.98	-7780.25	-7723.67	-7961.51	97	558	0.0	-0.03	-0.26
3	2439	GTL	-3054.51	157.30	-623.49	2353.88	2326.05	1033.69	-2612.42	1034.12	...	-3.33	-7.21	-48.13	-47.70	-51.58	93	63	2.0	0.24	1.90
4	23505	Bharati Defence	-2967.36	50.30	-1070.83	4675.33	5740.90	1084.20	1836.23	4685.81	...	-295.55	-400.55	-845.88	379.79	274.79	3887	346	0.0	0.01	0.05

Table 1.1: Data Sample

- Dataset has records of 3586 companies with 67 different features.

Exploratory Data Analysis (EDA)

- <u>Let’s check for data types of the variables and missing values in the data frame</u>

```
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/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 (%)                      3586 non-null   float64
 36  ROG-Capital Employed (%)               3586 non-null   float64
 37  ROG-Gross Block (%)                    3586 non-null   float64
 38  ROG-Gross Sales (%)                    3586 non-null   float64
 39  ROG-Net Sales (%)                      3586 non-null   float64
 40  ROG-Cost of Production (%)             3586 non-null   float64
 41  ROG-Total Assets (%)                   3586 non-null   float64
 42  ROG-PBIDT (%)                          3586 non-null   float64
 43  ROG-PBDT (%)                           3586 non-null   float64
 44  ROG-PBIT (%)                           3586 non-null   float64
 45  ROG-PBT (%)                            3586 non-null   float64
 46  ROG-PAT (%)                            3586 non-null   float64
 47  ROG-CP (%)                             3586 non-null   float64
 48  ROG-Revenue earnings in forex (%)      3586 non-null   float64
 49  ROG-Revenue expenses in forex (%)      3586 non-null   float64
 50  ROG-Market Capitalisation (%)          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 (%)[Latest]                     3585 non-null   float64
 58  PBITM (%)[Latest]                      3585 non-null   float64
 59  PBDTM (%)[Latest]                      3585 non-null   float64
 60  CPM (%)[Latest]                        3585 non-null   float64
 61  APATM (%)[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/Total Assets           3586 non-null   float64
 66  Value of Output/Gross Block            3586 non-null   float64
dtypes: float64(63), int64(3), object(1)
```

<u>Table 1.2: Data Summary</u>

- There are total 3586 entries and 67 columns present in the dataset.
- 63 features are of float data type while 1 with object and 3 with integer data type.
- There are comparatively more missing values present in ‘Inventory Velocity (Days)’ column.

- <u>Checking for outliers present in the data (only for Top 12 important features)</u>

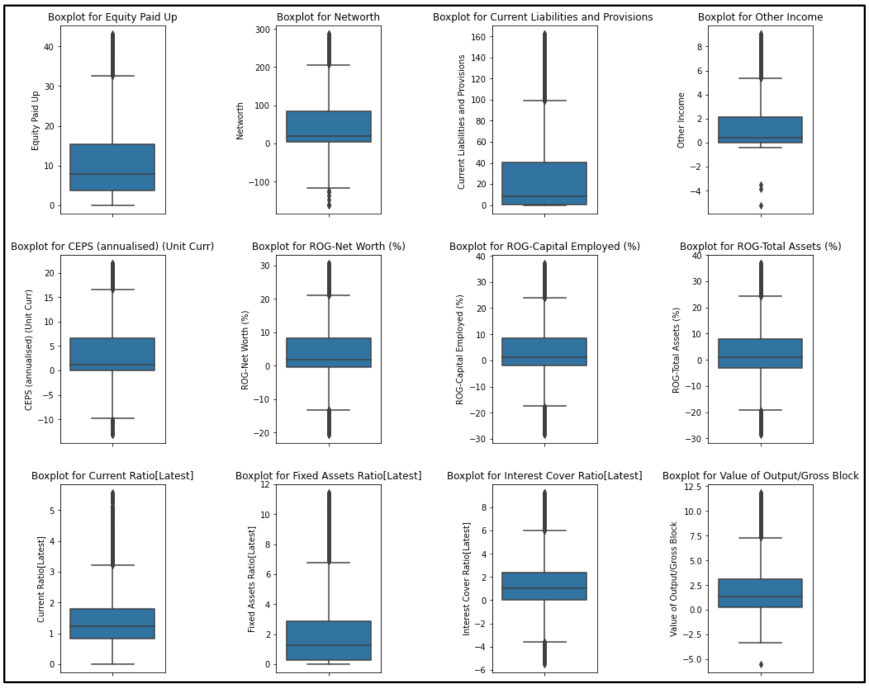

Fig 1.1: Boxplot for imp. features

From above boxplots,

- There are outliers present in all the features (top 12 important features).
- The outliers will be imputed using KNN Imputer in later part of the study.

- Distribution of features (only for Top 12 important features)

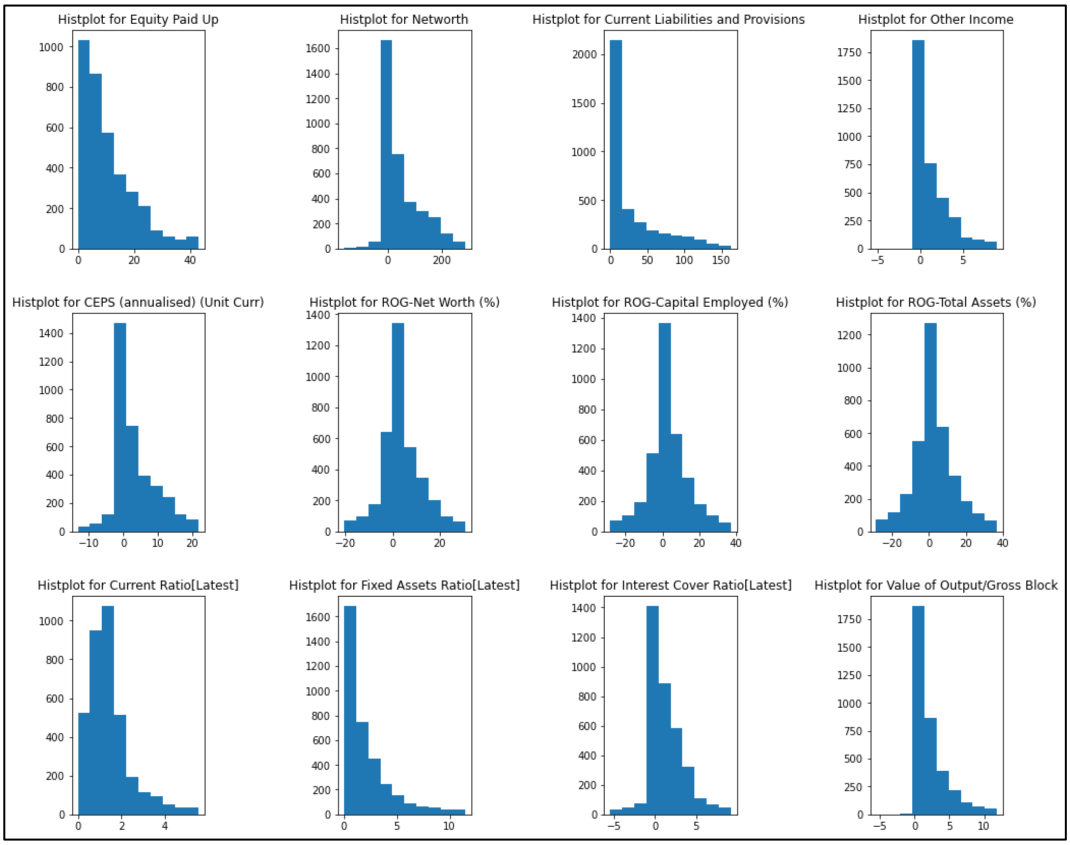

Fig 1.2: Distribution Graph for imp. features

- In the above graph, 1st & 3rd row features are right skewed whereas 2nd row features are more or less normally distributed.

\

Problem Statement 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.

Q1.1. & Q1.2. Outlier & missing value Treatment

- Step 1: Estimate number of outliers present in the dataset

APATM (%)[Latest]	933
Adjusted PAT	954
Book Value (Adj.) (Unit Curr)	486
Book Value (Unit Curr)	485
CEPS (annualised) (Unit Curr)	602
	...
Total Assets/Liabilities	574
Total Debt	583
Value Of Output	559
Value of Output/Gross Block	481
Value of Output/Total Assets	150

Table 1.3: Number of Outliers

- In above snap we can see the number of outliers for corresponding columns. There are total 42322 outliers present in the dataset.

- Step 2: Estimate number of NaN values present in the dataset
  - There are total 118 NaN values present in the dataset.

- Step 3: Convert the outliers into NA values
  - Now, there are total 42322 + 118 = 42440 NaN values present in the dataset.
  - Let’s visualize the missing/NaN values in the dataset:

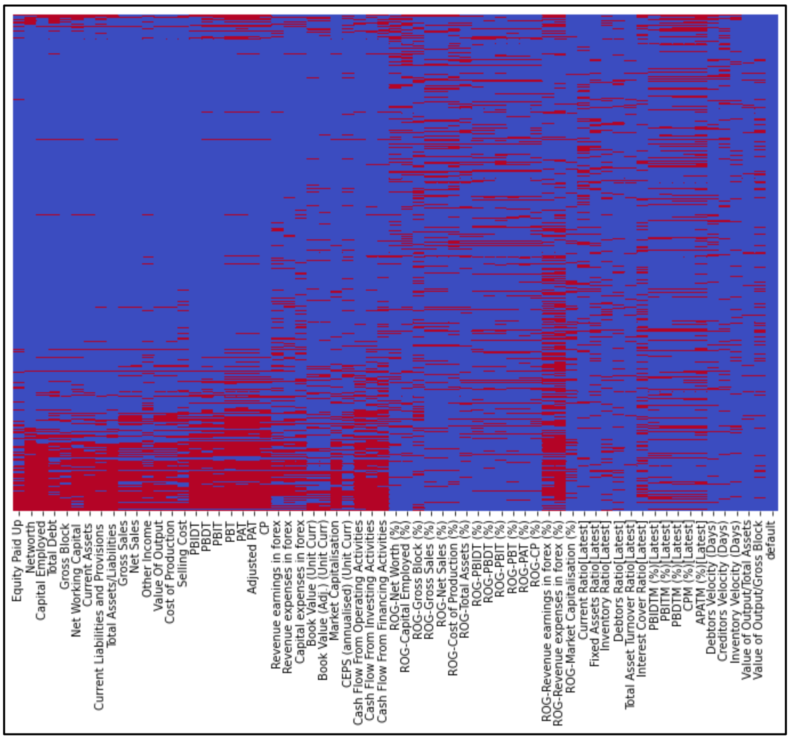

Fig 1.3: Visualizing the Outliers/NaN values

- Step 4: Drop the row & columns based upon NaN values
    - Drop the row whose NaN count is more than 6.
    - Drop the columns with more than 30% NaN values.
    - Let's visualize the results

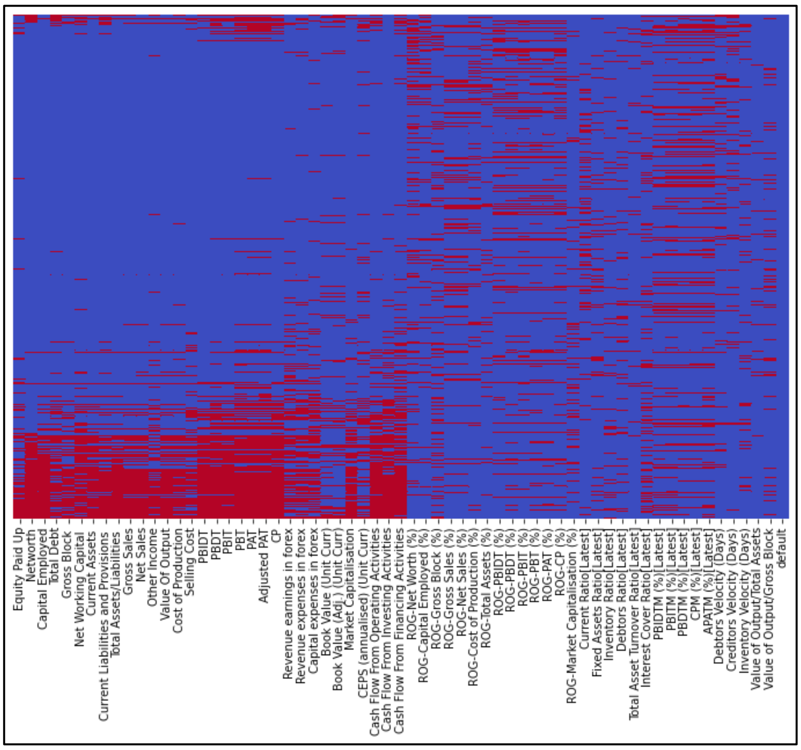

Fig 1.4: Visualizing the Outliers/NaN values after dropping certain values

The red color in the heatmap above is reduce certainly.

- Step 5: Treat the outliers and NaNs values
  - Import the KNNImputer from sklearn.impute library.
  - Define th K-neighbours, k=20
  - Apply the imputer to the dataset and check for NaN values. So, there are no more any NaN values present in the dataset.

Q1.3. Transform Target variable into 0 and 1.

- We will use 'Networth Next Year' variable which stands for Networth for 2016 for transforming target variable into binary target variable i.e., 0 & 1.
- Any value of 'Networth Next Year' < 0 is considered to be defaulter.
- Any value of 'Networth Next Year' > 0 is considered as a non-defaulter.
- To execute the above, we will be using np.where function.

	Co_Code	Co_Name	Networth Next Year	Equity Paid Up	Networth	Capital Employed	Total Debt	Gross Block	Net Working Capital	Current Assets	...	PBITM (%) [Latest]	PBDTM (%) [Latest]	CPM (%) [Latest]	APATM (%) [Latest]	Debtors Velocity (Days)	Creditors Velocity (Days)	Inventory Velocity (Days)	Value of Output/Total Assets	Value of Output/Gross Block	default
0	16974	Hind.Cables	-8021.60	419.36	-7027.48	-1007.24	5936.03	474.30	-1076.34	40.50	...	0.00	0.00	0.00	0.00	0	0	45.0	0.00	0.00	1
1	21214	Tata Tele. Mah.	-3986.19	1954.93	-2968.08	4458.20	7410.18	9070.86	-1098.88	486.86	...	-39.74	-57.74	-57.74	-87.18	29	101	2.0	0.31	0.24	1
2	14852	ABG Shipyard	-3192.58	53.84	506.86	7714.68	6944.54	1281.54	4496.25	9097.64	...	-5516.98	-7780.25	-7723.67	-7961.51	97	558	0.0	-0.03	-0.26	1
3	2439	GTL	-3054.51	157.30	-623.49	2353.88	2326.05	1033.69	-2612.42	1034.12	...	-7.21	-48.13	-47.70	-51.58	93	63	2.0	0.24	1.90	1
4	23505	Bharati Defence	-2967.36	50.30	-1070.83	4675.33	5740.90	1084.20	1836.23	4685.81	...	-400.55	-845.88	379.79	274.79	3887	346	0.0	0.01	0.05	1
5	2484	Usha Ispat	-2519.40	179.35	-2519.39	-1824.75	694.64	0.02	-1843.74	0.00	...	0.00	0.00	0.00	0.00	0	0	0.0	0.00	0.00	1
6	23633	Hanung Toys	-2125.05	30.82	-1031.57	1536.08	2567.65	949.98	804.82	834.86	...	-987.73	-396.67	-672.36	-1264.22	456	12	392.0	0.00	-0.01	1
7	3226	K S Oils	-2100.56	45.92	-1945.45	979.13	2664.04	920.67	263.95	705.76	...	-596.97	-456.40	-461.06	-610.80	828	622	799.0	-0.02	-0.03	1
8	1541	Quadrant Tele.	-1695.75	61.23	-1560.94	-613.79	597.82	1700.27	-1121.96	117.67	...	-20.43	-3.58	-3.58	-25.91	34	145	2.0	0.92	0.31	1
9	2334	ITI	-1677.18	288.00	-1947.85	86.35	1220.83	1329.82	-390.53	2536.78	...	18.18	9.76	9.76	8.71	1112	913	62.0	0.54	1.16	1

Table 1.4: Tranforming traget variables

Q1.4. Univariate & Bivariate analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building)

- **Univariate Analysis**:
  - Distplot for all numeric features-

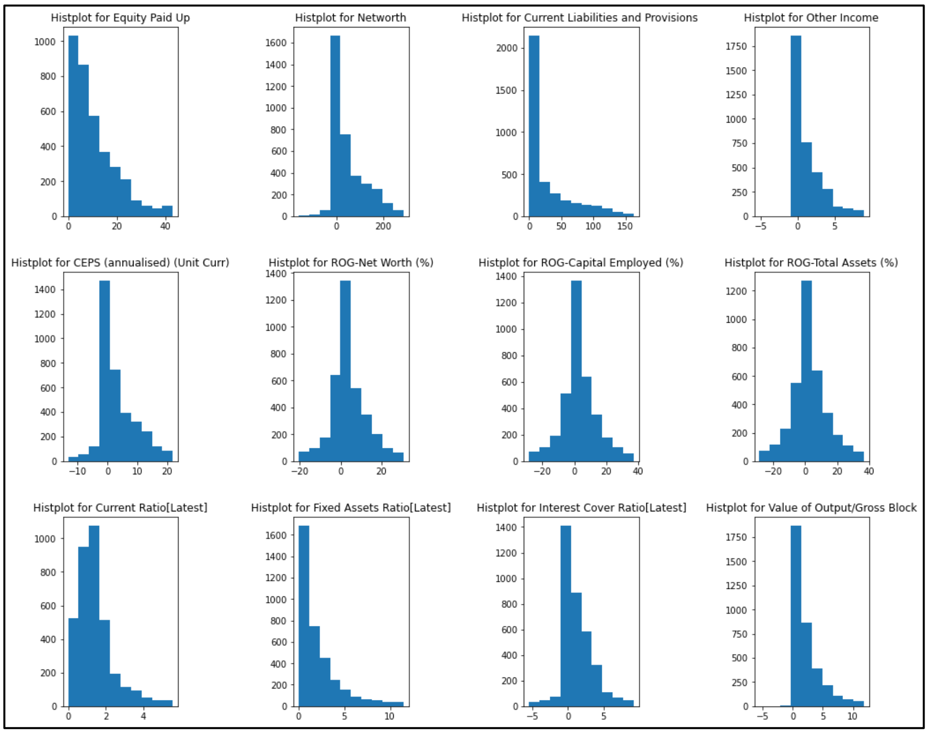

Fig 1.5: Distribution graph for imp. features

- In the above graph, 1st & 3rd row features are right skewed whereas 2nd row features are more or less normally distributed.

  - Boxplot for all numeric features-

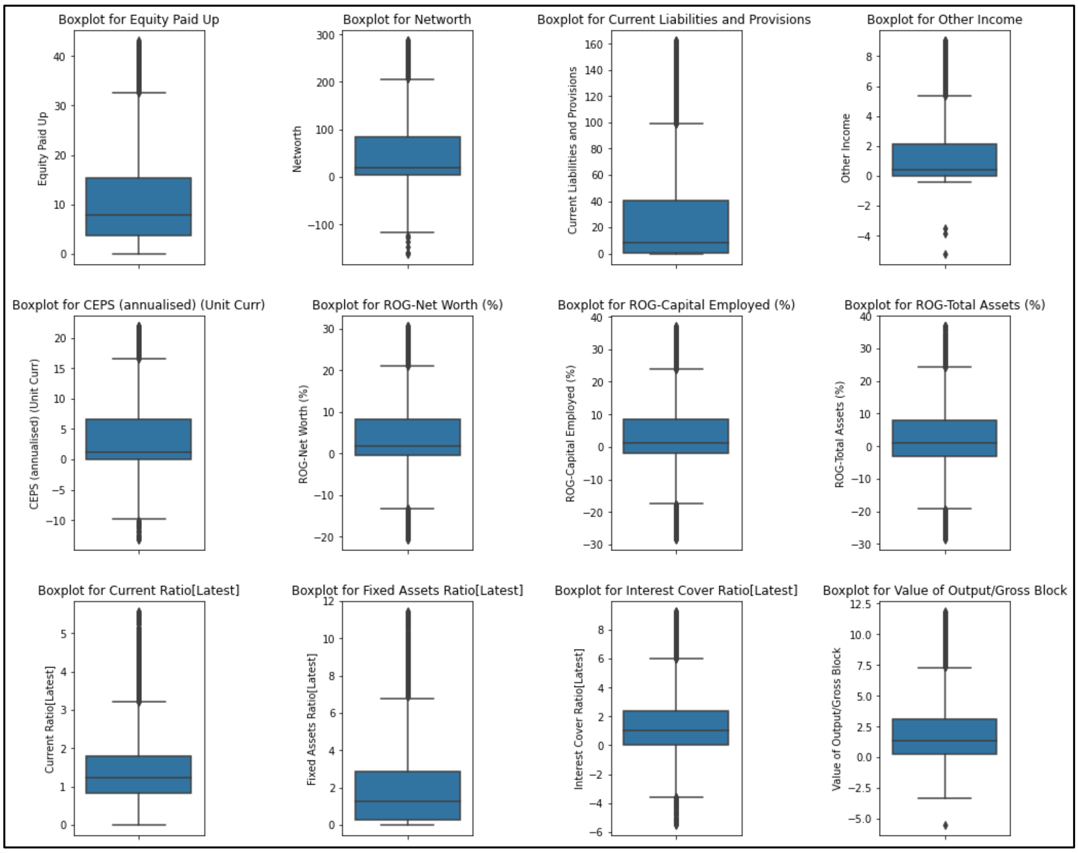

Fig 1.6: Boxplot for imp. features

- There are outliers present in all the features (top 12 important features).
- The outliers are imputed using KNN Imputer as seen earlier.

  - Count Plot:
- Defaulters & Non-defaulters:

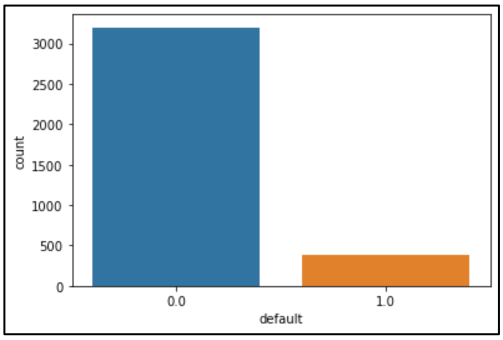

Fig 1.7: Count plot for Binary Target Variables

- There are significantly low defaulters incomparison with non-defaulters.
- The defaulters are around 10% and non-defaulters are around 90%.

- Bivariate Analysis:
  - Bar Plots:

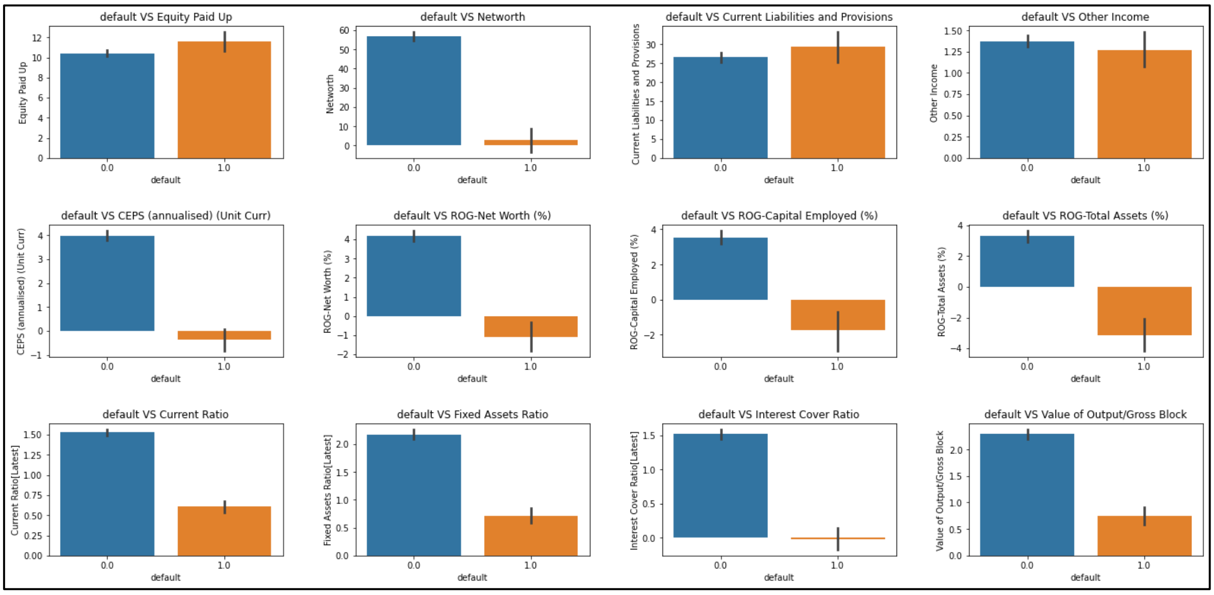

Fig 1.8: Bar plot for important features

- In the above bar plots, parameters clearly differentiate between the defaulters & non-defaulters.
- **Net worth** shows that non-defaulters have significantly higher net worth than defaulters. So this variable is good predictor of defaulters & non-defaulters.

- **CEPS** for defaulters is –ve whereas for non-defaulters its +ve.
- Similarly, **ROG-Net worth, ROG-Capital Employed and ROG –Total Assets** are –vr for defaulters and +ve for non-defaulters.
- Defaulters have comparatively low current ratio, fixed asset ratio, interest cover ratio & ratio of value of output & gross block.

  - Heatmap:

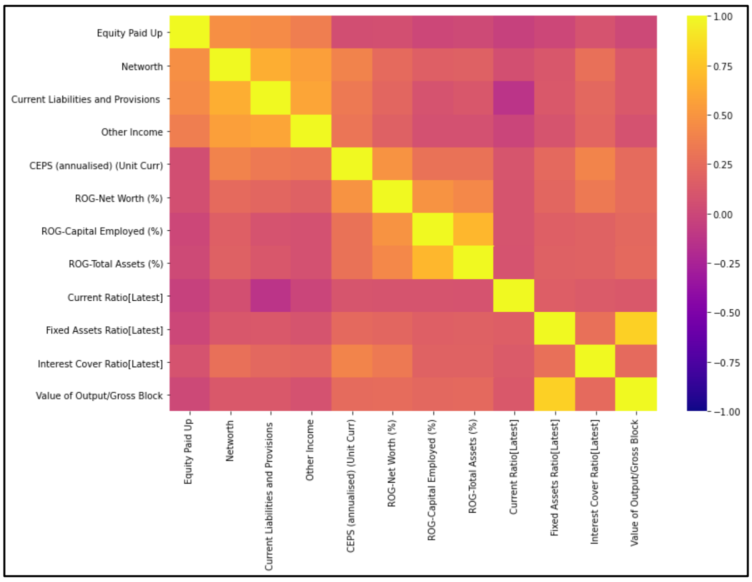

Fig 1.9: Heatmap for important features

- Value of output/Gross block seems to be positively correlated with fixed asset ratio.
- There doesn't seem to be any other strong correlations among the variables.

Q1.5. Train Test Split.

- We will use train_test_split from sklearn.model_selection library.
- Data is split into train & test dataset in proportion of 67:33.
- Use random state of 42 for data split.
- Following is the data shape for train & test data:

```
X_train:  (2402, 33)
X_test:  (1184, 33)
train_labels:  (2402,)
test_labels:  (1184,)
```

Table 1.5: Train test data size

Q1.6. 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.

- Step 1: Cheking the Variance Inflation Factor (VIF)
  - By using user defined function, list down the VIF for corresponding variables.

	variables	VIF
10	Net Sales	422.149768
12	Value Of Output	239.270424
9	Gross Sales	181.135712
35	ROG-Gross Sales (%)	74.485182
36	ROG-Net Sales (%)	74.080296
...	...	...
30	Cash Flow From Investing Activities	1.587149
45	ROG-Market Capitalisation (%)	1.478617
22	Revenue earnings in forex	1.445804
62	default	1.401539
24	Capital expenses in forex	NaN

Table 1.6: VIF values

- Step 2: Drop the variables with VIF > 5
  - We drop variables such as ‘Net Sales’, ‘Value Of Output’, ‘Gross Sales’, ‘ROG-Gross Sales (%)’, ‘ROG-Net Sales (%)’ etc. There are other such 25 variables.
  - We retained Fixed assets ratio and Value of Output/Gross Block which have VIF > 5 beacause these are important variables as far as financial statements are concerned.
  - Following is the list of retained variables for the model building activity.

	variables	VIF
23	Fixed Assets Ratio[Latest]	5.410343
32	Value of Output/Gross Block	5.184700
1	Networth	4.453249
4	Current Liabilities and Provisions	4.443206
0	Equity Paid Up	3.140773
27	PBITM (%)[Latest]	3.124548
7	Adjusted PAT	3.108448
28	APATM (%)[Latest]	2.991166
3	Net Working Capital	2.932643
2	Total Debt	2.825057
5	Other Income	2.748757
22	Current Ratio[Latest]	2.697516
29	Debtors Velocity (Days)	2.674274
12	CEPS (annualised) (Unit Curr)	2.644163
25	Debtors Ratio[Latest]	2.597147
30	Creditors Velocity (Days)	2.559007
6	Selling Cost	2.480497
11	Market Capitalisation	2.473724
17	ROG-Capital Employed (%)	2.374157
24	Inventory Ratio[Latest]	2.372920
31	Inventory Velocity (Days)	2.197108
20	ROG-Total Assets (%)	2.155819
26	Interest Cover Ratio[Latest]	2.154373
16	ROG-Net Worth (%)	2.111237
15	Cash Flow From Financing Activities	2.095659
13	Cash Flow From Operating Activities	1.927601
9	Revenue expenses in forex	1.773079
18	ROG-Gross Block (%)	1.681722
14	Cash Flow From Investing Activities	1.505245
21	ROG-Market Capitalisation (%)	1.432300
8	Revenue earnings in forex	1.425148
33	default	1.329690
19	ROG-Cost of Production (%)	1.174239
10	Capital expenses in forex	NaN

Table 1.7: Retained variables basis VIF

- Step 3: Split Test Train Dataset
  - We will use train_test_split from sklearn.model_selection library.
  - Data is split into train & test dataset in proportion of 67:33.
  - Use random state of 42 for data split.
  - Following is the data shape for train & test data:

```
X_train:  (2402, 33)
X_test:  (1184, 33)
train_labels:  (2402,)
test_labels:  (1184,)
```

Table 1.5: Train test data size

- Step 4: Model Building
  - For model building, we will use Logistic regression with recursive feature elimination (RFE).
  - Following are the functions in model building exercise:
    A. LogR = LogisticRegression()
    B. selector = RFE(estimator = LogR, n_features_to_select=12, step=1)
    C. selector = selector.fit(X_train, y_train)
    D. pred_train = selector.predict(X_train)
    E. pred_test = selector.predict(X_test)
  - We are using top 12 features to build LR model using RFE. Following are the feature list used in the model:

	Feature	Rank
0	Equity Paid Up	1
1	Networth	1
4	Current Liabilities and Provisions	1
5	Other Income	1
12	CEPS (annualised) (Unit Curr)	1
16	ROG-Net Worth (%)	1
19	ROG-Cost of Production (%)	1
22	Current Ratio[Latest]	1
26	Interest Cover Ratio[Latest]	1
27	PBITM (%)[Latest]	1
28	APATM (%)[Latest]	1
32	Value of Output/Gross Block	1

Table 1.8: Top 12 important variables considered for Model Building

  - Using above model, we get following classification report:

Train Classification Report

	precision	recall	f1-score	support
0.0	0.95	0.99	0.97	2157
1.0	0.84	0.49	0.62	245
accuracy			0.94	2402
macro avg	0.89	0.74	0.79	2402
weighted avg	0.93	0.94	0.93	2402

Test Classification Report

	precision	recall	f1-score	support
0.0	0.94	0.98	0.96	1041
1.0	0.82	0.51	0.63	143
accuracy			0.93	1184
macro avg	0.88	0.75	0.79	1184
weighted avg	0.92	0.93	0.92	1184

Table 1.9: Classification reports for Initial model

- Step 4: Improving the model
  - As you can see, the model in above exercise is overfitting as well as doesn't posses good recall score.
  - To improve the model we will use SMOTE technique to balance the binary traget variable.
  - Following are the functions in improved model building exercise:

A. from imblearn.over_sampling import SMOTE
B. sm = SMOTE(random_state=42)
C. X_res, y_res = sm.fit_resample(X_train, y_train)
D. selector_smote = selector.fit(X_res, y_res)
E. pred_train_smote = selector_smote.predict(X_res)
F. pred_test_smote = selector_smote.predict(X_test)

- Following is the classification report for the improved model

Train Classification Report

	precision	recall	f1-score	support
0.0	0.88	0.85	0.87	2157
1.0	0.86	0.89	0.87	2157
accuracy			0.87	4314
macro avg	0.87	0.87	0.87	4314
weighted avg	0.87	0.87	0.87	4314

Test Classification Report

	precision	recall	f1-score	support
0.0	0.98	0.82	0.89	1041
1.0	0.41	0.89	0.56	143
accuracy			0.83	1184
macro avg	0.69	0.85	0.73	1184
weighted avg	0.91	0.83	0.85	1184

Table 1.10: Classification reports for Improved model

Q1.7. Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model.

- **Step 1: Accuracy**
  1. Accuracy on test dataset is 83% which is good for a model.

- **Step 2:** Confusion Matrix

```
[[855 186]
 [ 16 127]]
```

SMOTE Test CM

Fig 1.11: Confusion matrix for Improved Model

- **Step 3:** Classification Reports

Train Classification Report

	precision	recall	f1-score	support
0.0	0.88	0.85	0.87	2157
1.0	0.86	0.89	0.87	2157
accuracy			0.87	4314
macro avg	0.87	0.87	0.87	4314
weighted avg	0.87	0.87	0.87	4314

Test Classification Report

	precision	recall	f1-score	support
0.0	0.98	0.82	0.89	1041
1.0	0.41	0.89	0.56	143
accuracy			0.83	1184
macro avg	0.69	0.85	0.73	1184
weighted avg	0.91	0.83	0.85	1184

Table 1.10: Classification reports for Improved model

1. In terms of recall, train & test model have performed well.

Interpretations from the model:

1. The model which is build above is much of practical approach.
2. We 1st treated the NaN & outliers in a proper way and not just assigning the central tendencies.

3. Eliminate the variables basis upon the VIF values which will ensure the good variables are retained.
4. Using the most important variables is necessary for model to perform well which are identified using RFE.
5. SMOTE is the key when it comes to unbalanced binary target variables. The performance of the model is lifted significantly as seen from the above exercise.
6. We obtain same recall score in train & test dataset which is balance between overfitting and under fitting.

The End of the Report