FRA PROJECT
Project I

DSBA

BUSINESS REPORT

Sayyed Abdul Khaliq
Email : abdulkhaliq01112001@gmail.com

CONTENTS

List of Figures

List of Tables

Problem 1

Problem statement:

As a data scientist, your task is to analyse the provided financial data and develop a predictive model using machine learning techniques to identify whether a company will be tagged as a defaulter based on its net worth next year. A company will be considered a defaulter if its net worth next year is negative; otherwise, it will not be tagged as a defaulter. This predictive model will help anticipate potential challenges in the financial performance of companies and enable proactive risk mitigation strategies.

Context

- In modern finance, effective debt management is crucial for businesses to maintain a favourable credit standing and foster sustainable growth. Investors scrutinize companies that navigate financial complexities while ensuring stability and profitability. The balance sheet, offering insights into a company's financial health and operational efficiency, becomes a key instrument in this evaluation process. Leveraging historical financial data is imperative for informed decision-making and strategic planning.
- A group of venture capitalists seeks to develop a Financial Health Assessment Tool to evaluate the financial well-being and creditworthiness of companies. By harnessing machine learning techniques, the tool will analyse historical financial statements to extract insights, enabling informed decision-making. Specifically, the tool aims to:

- Debt Management Analysis: Identify patterns and trends in debt management to assess businesses' ability to fulfil financial obligations and identify potential default cases.
- Credit Risk Evaluation: Evaluate credit risk exposure by analysing liquidity ratios, debt-to-equity ratios, and other key financial indicators to ascertain the likelihood of default and inform investment decisions.

Data Description

The data includes financial metrics from the balance sheets of different companies, as detailed in the data dictionary.

Data Dictionary

1. Networth Next Year: Net worth of the company in the next year
2. Total assets: Total assets of the company
3. Net worth: Net worth of the company for the present year
4. Total income: Total income of the company
5. Change in stock: Difference between the current stock value and the value of stock on the last trading day
6. Total expenses: Total expenses incurred by the company
7. Profit after tax (PAT): Profit after tax deduction
8. PBDITA: Profit before depreciation, income tax, and amortization

9. PBT: Profit before tax deduction
10. Cash profit: Total cash profit
11. PBDITA as % of total income: PBDITA / Total income
12. PBT as % of total income: PBT / Total income
13. PAT as % of total income: PAT / Total income
14. Cash profit as % of total income: Cash Profit / Total income
15. PAT as % of net worth: PAT / Net worth
16. Sales: Sales made by the company
17. Income from financial services: Income from financial services
18. Other income: Income from other sources
19. Total capital: Total capital of the company
20. Reserves and funds: Total reserves and funds of the company
21. Borrowings: Total amount borrowed by the company
22. Current liabilities & provisions: Current liabilities of the company
23. Deferred tax liability: Future income tax payable due to current transactions
24. Shareholders funds: Amount of equity in the company belonging to shareholders
25. Cumulative retained profits: Total cumulative profit retained by the company
26. Capital employed: Current assets minus current liabilities
27. TOL/TNW: Total liabilities divided by Total net worth
28. Total term liabilities / tangible net worth: Short + long term liabilities divided by tangible net worth
29. Contingent liabilities / Net worth (%): Contingent liabilities / Net worth
30. Contingent liabilities: Liabilities due to uncertain events
31. Net fixed assets: Purchase price of all fixed assets
32. Investments: Total invested amount
33. Current assets: Assets expected to be converted to cash within a year
34. Net working capital: Difference between current liabilities and current assets
35. Quick ratio (times): Total cash divided by current liabilities
36. Current ratio (times): Current assets divided by current liabilities
37. Debt to equity ratio (times): Total liabilities divided by shareholder equity
38. Cash to current liabilities (times): Total liquid cash divided by current liabilities
39. Cash to average cost of sales per day: Total cash divided by the average cost of sales
40. Creditors turnover: Net credit purchase divided by average trade creditors
41. Debtors turnover: Net credit sales divided by average accounts receivable
42. Finished goods turnover: Annual sales divided by average inventory

43. WIP turnover: Cost of goods sold divided by average inventory for the period
44. Raw material turnover: Cost of goods sold divided by average inventory for the same period
45. Shares outstanding: Number of issued shares minus shares held in the company
46. Equity face value: Cost of the equity at the time of issuing
47. EPS: Net income divided by the total number of outstanding shares
48. Adjusted EPS: Adjusted net earnings divided by the weighted average number of common shares outstanding on a diluted basis
49. Total liabilities: Sum of all types of liabilities
50. PE on BSE: Company's current stock price divided by its earnings per share

Data Overview

Read the data as an appropriate time series data

Data is loaded into dataframe using pandas library and first 5 were printed.

	Num	Networth Next Year	Total assets	Net worth	Total income	Change in stock	Total expenses	Profit after tax	PBDITA	PBT	...	Debtors turnover	Finished goods turnover	WIP turnover	Raw material turnover	Shares outstanding	Equity face value	EPS	A
0	1	395.3	827.6	336.5	534.1	13.5	508.7	38.9	124.4	64.6	...	5.65	3.99	3.37	14.87	8760056.0	10.0	4.44	
1	2	36.2	67.7	24.3	137.9	-3.7	131.0	3.2	5.5	1.0	...	NaN	NaN	NaN	NaN	NaN	NaN	0.00	
2	3	84.0	238.4	78.9	331.2	-18.1	309.2	3.9	25.8	10.5	...	2.51	17.67	8.76	8.35	NaN	NaN	0.00	
3	4	2041.4	6883.5	1443.3	8448.5	212.2	8482.4	178.3	418.4	185.1	...	1.91	18.14	18.62	11.11	10000000.0	10.0	17.60	
4	5	41.8	90.9	47.0	388.6	3.4	392.7	-0.7	7.2	-0.6	...	68.00	45.87	28.67	19.93	107315.0	100.0	-6.52	

5 rows × 51 columns

Table 1 –Dataset – Rows of Data

Check the structure of the data

- The number of rows (observations) is 4256
- The number of columns (variables) is 51

Check the Datatypes:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4256 entries, 0 to 4255
Data columns (total 51 columns):
 #   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
 0   Num                                       4256 non-null   int64
 1   Networth Next Year                        4256 non-null   float64
 2   Total assets                              4256 non-null   float64
 3   Net worth                                 4256 non-null   float64
 4   Total income                              4025 non-null   float64
 5   Change in stock                           3706 non-null   float64
 6   Total expenses                            4091 non-null   float64
 7   Profit after tax                          4102 non-null   float64
 8   PBDITA                                    4102 non-null   float64
 9   PBT                                       4102 non-null   float64
 10  Cash profit                               4102 non-null   float64
 11  PBDITA as % of total income               4177 non-null   float64
 12  PBT as % of total income                  4177 non-null   float64
 13  PAT as % of total income                  4177 non-null   float64
 14  Cash profit as % of total income          4177 non-null   float64
 15  PAT as % of net worth                     4256 non-null   float64
 16  Sales                                     3951 non-null   float64
 17  Income from fincial services              3145 non-null   float64
 18  Other income                              2700 non-null   float64
 19  Total capital                             4251 non-null   float64
 20  Reserves and funds                        4158 non-null   float64
 21  Borrowings                                3825 non-null   float64
 22  Current liabilities & provisions          4146 non-null   float64
 23  Deferred tax liability                    2887 non-null   float64
 24  Shareholders funds                        4256 non-null   float64
 25  Cumulative retained profits               4211 non-null   float64
 26  Capital employed                          4256 non-null   float64
 27  TOL/TNW                                   4256 non-null   float64
 28  Total term liabilities / tangible net worth 4256 non-null  float64
 29  Contingent liabilities / Net worth (%)    4256 non-null   float64
 30  Contingent liabilities                    2854 non-null   float64
 31  Net fixed assets                          4124 non-null   float64
 32  Investments                               2541 non-null   float64
 33  Current assets                            4176 non-null   float64
 34  Net working capital                       4219 non-null   float64
 35  Quick ratio (times)                       4151 non-null   float64
 36  Current ratio (times)                     4151 non-null   float64
 37  Debt to equity ratio (times)              4256 non-null   float64
 38  Cash to current liabilities (times)       4151 non-null   float64
 39  Cash to average cost of sales per day     4156 non-null   float64
 40  Creditors turnover                        3865 non-null   float64
 41  Debtors turnover                          3871 non-null   float64
 42  Finished goods turnover                   3382 non-null   float64
 43  WIP turnover                              3492 non-null   float64
 44  Raw material turnover                     3828 non-null   float64
 45  Shares outstanding                        3446 non-null   float64
 46  Equity face value                         3446 non-null   float64
 47  EPS                                       4256 non-null   float64
 48  Adjusted EPS                              4256 non-null   float64
 49  Total liabilities                         4256 non-null   float64
 50  PE on BSE                                 1629 non-null   float64
dtypes: float64(50), int64(1)
memory usage: 1.7 MB
```

Table 2 –Dataset– Info

There are 49 float datatypes. Additionally, there is 1 datatype which is redundant for analysis.

Check for and treat (if needed) missing values –

Num	0
Networth Next Year	0
Total assets	0
Net worth	0
Total income	231
Change in stock	550
Total expenses	165
Profit after tax	154
PBDITA	154
PBT	154
Cash profit	154
PBDITA as % of total income	79
PBT as % of total income	79
PAT as % of total income	79
Cash profit as % of total income	79
PAT as % of net worth	0
Sales	305
Income from fincial services	1111
Other income	1556
Total capital	5
Reserves and funds	98
Borrowings	431
Current liabilities & provisions	110
Deferred tax liability	1369
Shareholders funds	0
Cumulative retained profits	45
Capital employed	0
TOL/TNW	0
Total term liabilities / tangible net worth	0
Contingent liabilities / Net worth (%)	0
Contingent liabilities	1402
Net fixed assets	132
Investments	1715
Current assets	80
Net working capital	37
Quick ratio (times)	105
Current ratio (times)	105
Debt to equity ratio (times)	0
Cash to current liabilities (times)	105
Cash to average cost of sales per day	100
Creditors turnover	391
Debtors turnover	385
Finished goods turnover	874
WIP turnover	764
Raw material turnover	428
Shares outstanding	810
Equity face value	810
EPS	0
Adjusted EPS	0
Total liabilities	0
PE on BSE	2627
dtype: int64	

Table 3 –Null values

There are null values in the dataset.

Treating null values is very important to do further analysis.

In this approach, instead of taking means, we used KNN neighbours to impute the missing data and dependent columns are recalculated again using the imputed variables.

There are few infinite values created in the process of recalculating dependent columns which are converted to zeroes.

KNN Imputation:

Imputing missing values using K-Nearest Neighbours (KNN) is a robust method that leverages the similarity between data points.

This method is effective for datasets where the missing values are randomly distributed and the dataset size is manageable for computational resources.

Num	0
Networth Next Year	0
Total assets	0
Net worth	0
Total income	0
Change in stock	0
Total expenses	0
Profit after tax	0
PBDITA	0
PBT	0
Cash profit	0
PBDITA as % of total income	2
PBT as % of total income	2
PAT as % of total income	2
Cash profit as % of total income	2
PAT as % of net worth	0
Sales	0
Income from fincial services	0
Other income	0
Total capital	0
Reserves and funds	0
Borrowings	0
Current liabilities & provisions	0
Deferred tax liability	0
Shareholders funds	0
Cumulative retained profits	0
Capital employed	0
TOL/TNW	0
Total term liabilities / tangible net worth	0
Contingent liabilities / Net worth (%)	0
Contingent liabilities	0
Net fixed assets	0
Investments	0
Current assets	0
Net working capital	0
Quick ratio (times)	0
Current ratio (times)	0
Debt to equity ratio (times)	0
Cash to current liabilities (times)	0
Cash to average cost of sales per day	0
Creditors turnover	0
Debtors turnover	0
Finished goods turnover	0
WIP turnover	0
Raw material turnover	0
Shares outstanding	0
Equity face value	0
EPS	0
Adjusted EPS	0
Total liabilities	0
PE on BSE	0

Table 4 – After treating Null values

Statistical Summary:

	count	mean	std	min	25%	50%	75%	max
Num	4256.00000	2128.50000	1228.74570	1.00000	1064.75000	2128.50000	3192.25000	4256.00000
Networth Next Year	4256.00000	1344.74088	15936.74317	-74265.60000	3.97500	72.10000	330.82500	805773.40000
Total assets	4256.00000	3573.61715	30074.44344	0.10000	91.30000	315.50000	1120.80000	1176509.20000
Net worth	4256.00000	1351.94960	12961.31165	0.00000	31.47500	104.80000	389.85000	613151.60000
Total income	4025.00000	4688.18979	53918.94661	0.00000	107.10000	455.10000	1485.00000	2442828.20000
Change in stock	3706.00000	43.70248	436.91505	-3029.40000	-1.80000	1.60000	18.40000	14185.50000
Total expenses	4091.00000	4356.30110	51398.08712	-0.10000	96.80000	426.80000	1395.70000	2366035.30000
Profit after tax	4102.00000	295.05059	3079.90207	-3908.30000	0.50000	9.00000	53.30000	119439.10000
PBDITA	4102.00000	605.94064	5646.23063	-440.70000	6.92500	36.90000	158.70000	208576.50000
PBT	4102.00000	410.25904	4217.41531	-3894.80000	0.80000	12.60000	74.17500	145292.60000
Cash profit	4102.00000	408.26748	4143.92639	-2245.70000	2.90000	19.40000	96.25000	176911.80000
PBDITA as % of total income	4177.00000	3.17989	172.25656	-6400.00000	4.97000	9.68000	16.47000	100.00000
PBT as % of total income	4177.00000	-18.19683	419.91109	-21340.00000	0.56000	3.34000	8.94000	100.00000
PAT as % of total income	4177.00000	-20.03367	423.57619	-21340.00000	0.35000	2.37000	6.42000	150.00000
Cash profit as % of total income	4177.00000	-9.02128	299.95743	-15020.00000	2.00000	5.66000	10.73000	100.00000
PAT as % of net worth	4256.00000	10.16786	61.53240	-748.72000	0.00000	8.04000	20.20250	2466.67000
Sales	3951.00000	4645.68454	53080.90330	0.10000	113.35000	468.60000	1481.20000	2384984.40000
Income from fincial services	3145.00000	81.36006	1042.75868	0.00000	0.50000	1.90000	9.80000	51938.20000
Other income	2700.00000	55.95289	1178.41526	0.00000	0.40000	1.50000	6.20000	42856.70000
Total capital	4251.00000	224.55766	1684.95129	0.10000	13.20000	42.60000	103.15000	78273.20000
Reserves and funds	4158.00000	1210.56193	12816.22922	-6525.90000	5.30000	55.15000	282.52500	625137.80000
Borrowings	3825.00000	1176.24808	8581.24892	0.10000	24.40000	99.80000	358.30000	278257.30000
Current liabilities & provisions	4146.00000	960.63143	9140.53613	0.10000	17.50000	70.30000	265.92500	352240.30000
Deferred tax liability	2887.00000	234.49512	2106.25316	0.10000	3.20000	13.50000	51.30000	72796.60000
Shareholders funds	4256.00000	1376.48672	13010.69116	0.00000	32.30000	107.60000	408.90000	613151.60000
Cumulative retained profits	4211.00000	937.18198	9853.09609	-6534.30000	1.10000	37.40000	206.20000	390133.80000
Capital employed	4256.00000	2433.61758	20496.40388	0.00000	61.30000	221.20000	790.30000	891408.90000
TOL/TNW	4256.00000	4.02534	20.87909	-350.48000	0.60000	1.42000	2.83000	473.00000
Total term liabilities / tangible net worth	4256.00000	1.85429	15.87506	-325.60000	0.05000	0.34500	1.00000	456.00000
Contingent liabilities / Net worth (%)	4256.00000	55.70750	369.16567	0.00000	0.00000	5.36000	31.01250	14704.27000
Contingent liabilities	2854.00000	948.55224	12056.73758	0.10000	6.00000	37.85000	195.32500	559506.80000
Net fixed assets	4124.00000	1209.48652	12502.39664	0.00000	26.20000	93.85000	352.82500	636604.60000
Investments	2541.00000	721.86588	6793.85987	0.00000	1.00000	8.20000	63.80000	199978.60000
Current assets	4176.00000	1350.36001	10155.57275	0.10000	36.60000	148.35000	515.00000	354815.20000
Net working capital	4219.00000	162.87424	3182.02996	-63839.00000	-1.10000	16.70000	86.50000	85782.80000
Quick ratio (times)	4151.00000	1.49735	9.32752	0.00000	0.41000	0.67000	1.03000	341.00000
Current ratio (times)	4151.00000	2.25740	12.47829	0.00000	0.93000	1.23000	1.72000	505.00000
Debt to equity ratio (times)	4256.00000	2.87156	15.59997	0.00000	0.22000	0.79000	1.75000	456.00000
Cash to current liabilities (times)	4151.00000	0.52842	4.79634	0.00000	0.02000	0.07000	0.19000	165.00000
Cash to average cost of sales per day	4156.00000	145.15793	2521.99181	0.00000	2.88000	8.04000	21.97000	128040.76000
Creditors turnover	3865.00000	16.81226	75.67492	0.00000	3.72000	6.17000	11.69000	2401.00000
Debtors turnover	3871.00000	17.92903	90.16443	0.00000	3.81000	6.47000	11.85000	3135.20000
Finished goods turnover	3382.00000	84.36999	562.63736	-0.09000	8.19000	17.32000	40.01250	17947.60000
WIP turnover	3492.00000	28.68451	169.65092	-0.18000	5.10000	9.86000	20.24000	5651.40000
Raw material turnover	3828.00000	17.73393	343.12586	-2.00000	3.02000	6.41000	11.82250	21092.00000
Shares outstanding	3446.00000	23764909.55543	170979041.32987	-2147483647.00000	1308382.50000	4750000.00000	10906020.00000	4130400545.00000
Equity face value	3446.00000	-1094.82867	34101.35864	-999998.90000	10.00000	10.00000	10.00000	100000.00000
EPS	4256.00000	-196.21747	13061.95342	-843181.82000	0.00000	1.49000	10.00000	34522.53000
Adjusted EPS	4256.00000	-197.52761	13061.92951	-843181.82000	0.00000	1.24000	7.61500	34522.53000
Total liabilities	4256.00000	3573.61715	30074.44344	0.10000	91.30000	315.50000	1120.80000	1176509.20000
PE on BSE	1629.00000	55.46229	1304.44530	-1116.64000	2.97000	8.69000	17.00000	51002.74000

Table 5– Summary Stats

Check for duplicates:

There are no duplicates in the dataset

Outlier treatment:

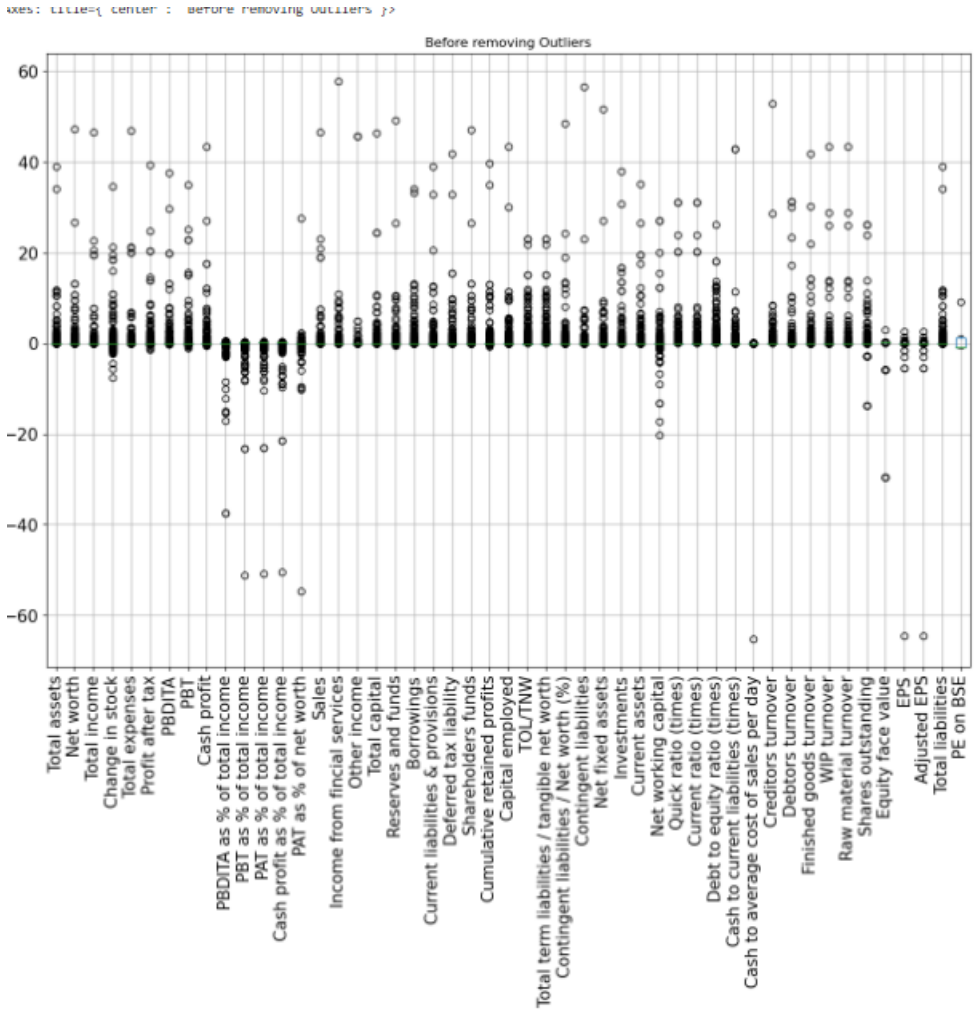


Fig.1 Before removing Outliers

There are various ways to handle outliers, such as removing them or capping them. Using the IQR method is effective for identifying and handling outliers in many datasets, ensuring that your analysis is not unduly influenced by extreme values.

- The IQR is the range between the first and third quartiles: IQR=Q3−Q1\text{IQR} = Q3 - Q1IQR=Q3−Q1.
- Outliers are typically defined as data points that fall below Q1−1.5xIQR or above Q3+1.5×IQR

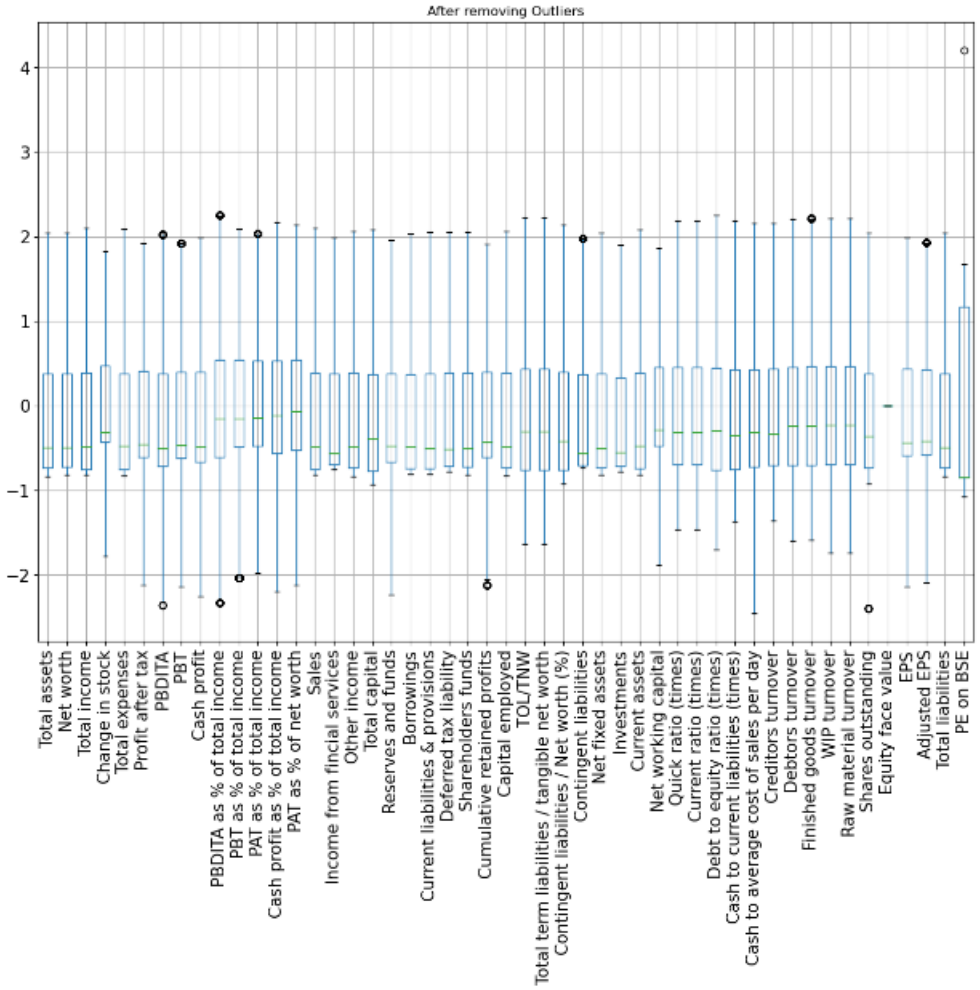


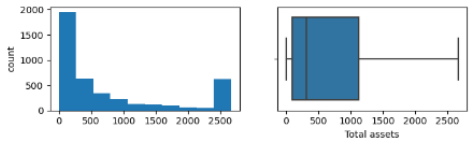
Fig.2 After removing Outliers

Perform EDA :

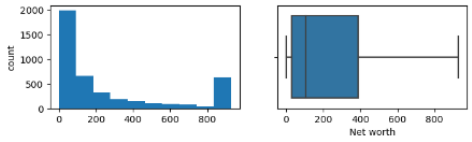
Univariate Analysis:

Let's check the distribution and skewness for each column in the data

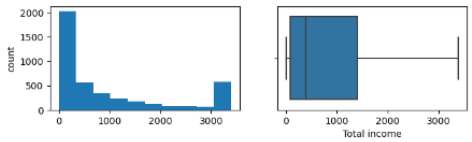
Total assets
Skew : 1.18



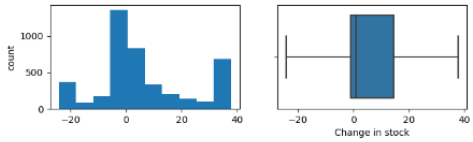
Net worth
Skew : 1.2



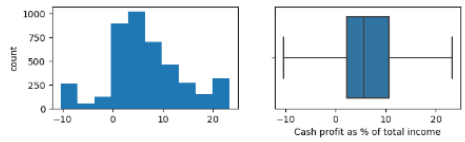
Total income
Skew : 1.2



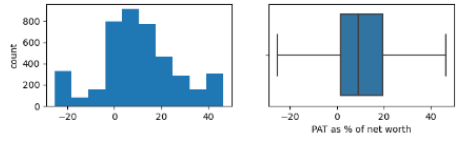
Change in stock
Skew : 0.47



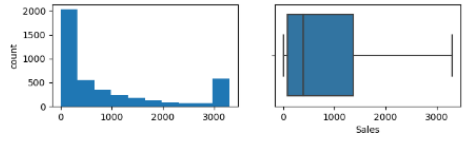
Cash profit as % of total income
Skew : 0.15



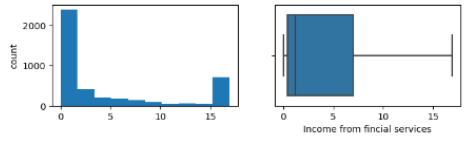
PAT as % of net worth
Skew : 0.01



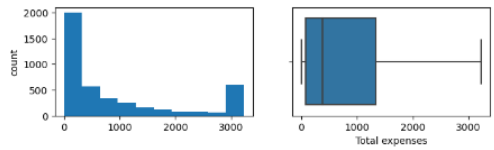
Sales
Skew : 1.2



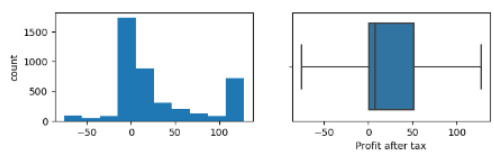
Income from fincial services
Skew : 1.22



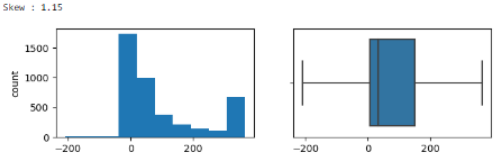
Total expenses
Skew : 1.19



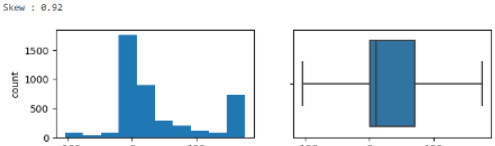
Profit after tax
Skew : 0.88



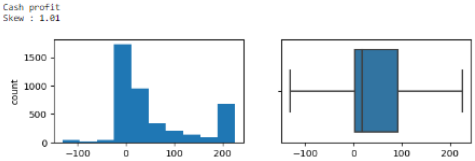
PBDITA
Skew : 1.15



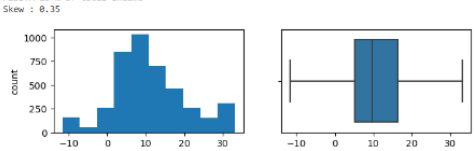
PBT
Skew : 0.92



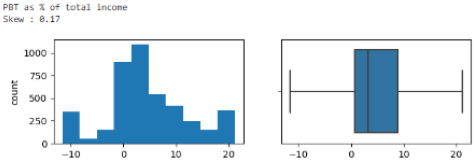
Cash profit
Skew : 1.01



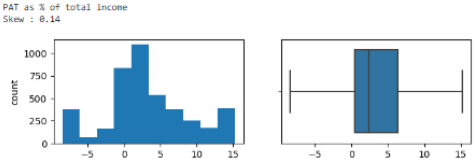
PBDITA as % of total income
Skew : 0.35

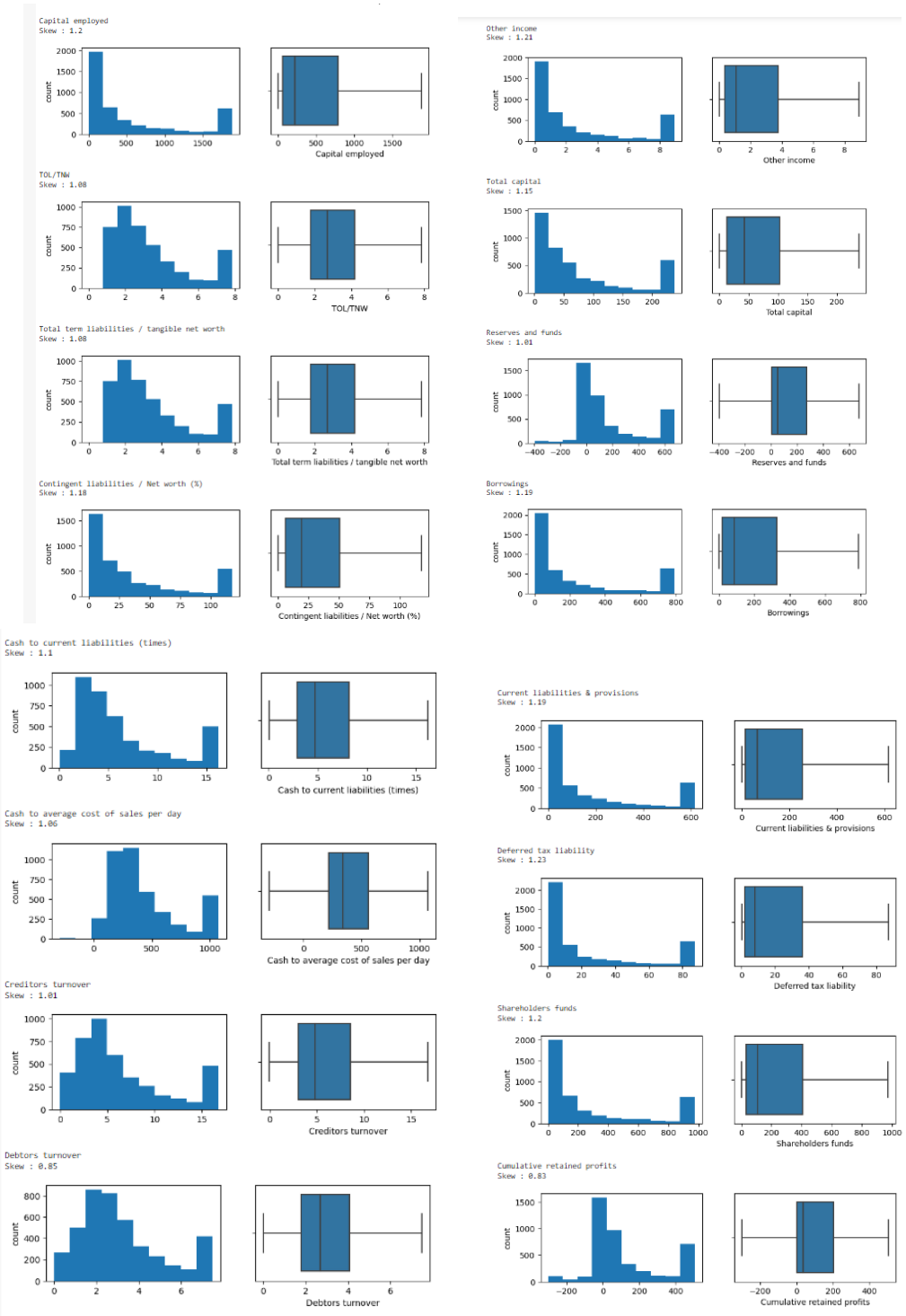


PBT as % of total income
Skew : 0.17



PAT as % of total income
Skew : 0.14




Capital employed
Skew : 1.2
Capital employed
TOL/TNW
Skew : 1.08
TOL/TNW
Total term liabilities / tangible net worth
Skew : 1.08
Total term liabilities / tangible net worth
Contingent liabilities / Net worth (%)
Skew : 1.18
Contingent liabilities / Net worth (%)
Other income
Skew : 1.21
Other income
Total capital
Skew : 1.15
Total capital
Reserves and funds
Skew : 1.01
Reserves and funds
Borrowings
Skew : 1.19
Borrowings
Cash to current liabilities (times)
Skew : 1.1
Cash to current liabilities (times)
Cash to average cost of sales per day
Skew : 1.06
Cash to average cost of sales per day
Creditors turnover
Skew : 1.01
Creditors turnover
Debtors turnover
Skew : 0.85
Debtors turnover
Current liabilities & provisions
Skew : 1.19
Current liabilities & provisions
Deferred tax liability
Skew : 1.23
Deferred tax liability
Shareholders funds
Skew : 1.2
Shareholders funds
Cumulative retained profits
Skew : 0.83
Cumulative retained profits
count

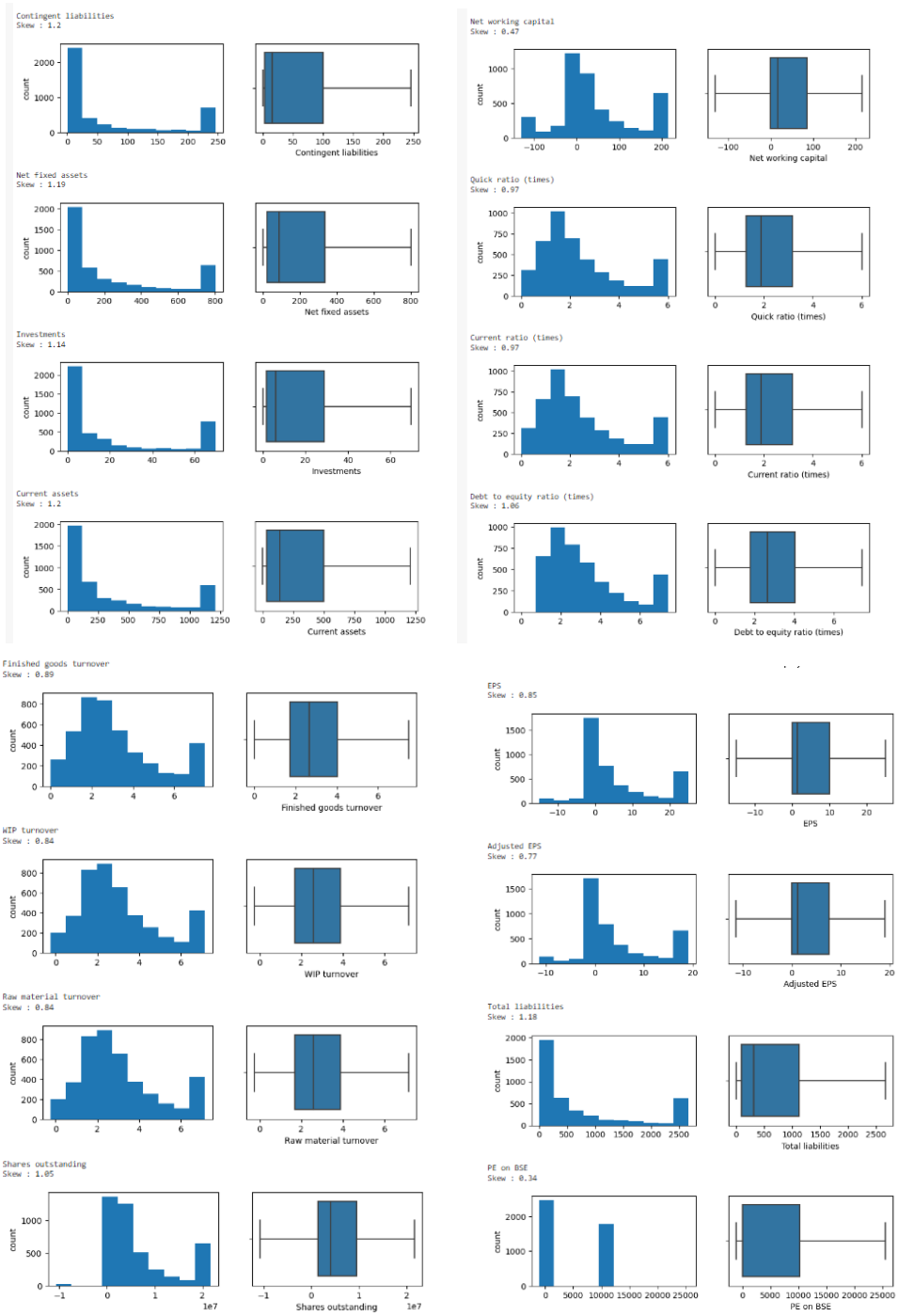


Fig.3 Univariate Analysis

Bivariate Analysis:

Let’s see correlation of columns with each other.

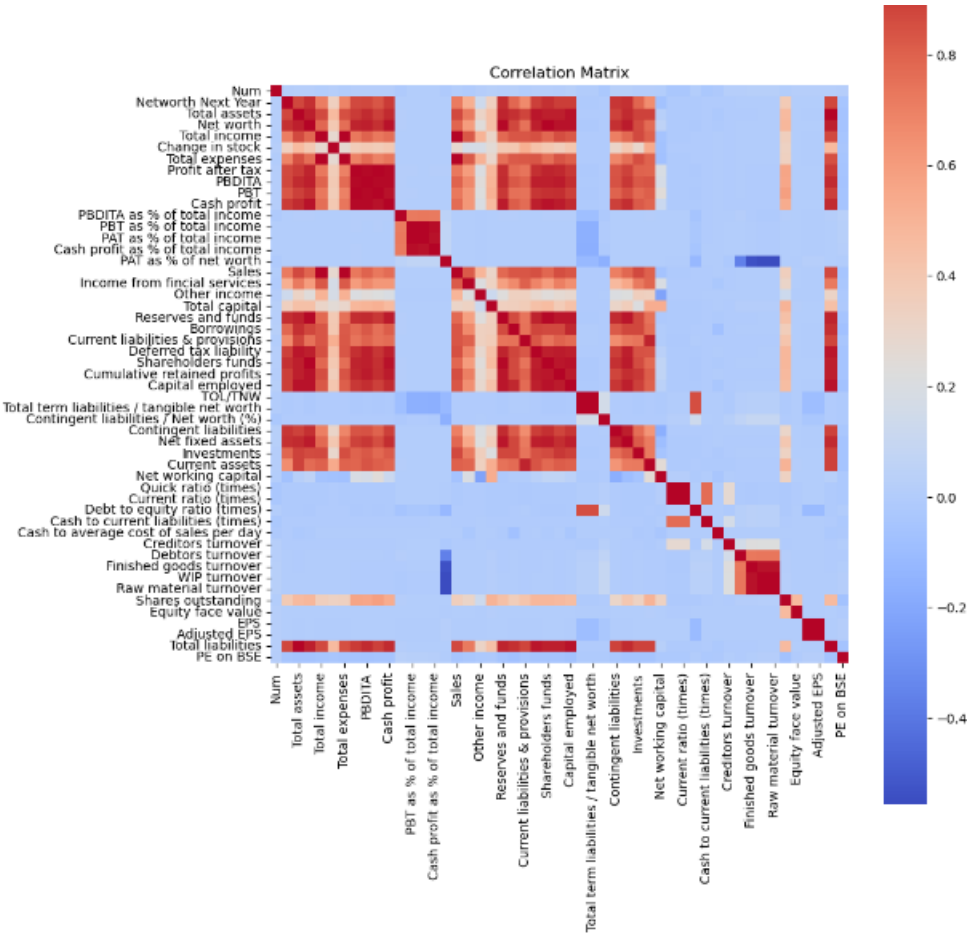


Fig.4 Bivariate Analysis

Key meaningful observations :

- **Profit after Tax (PAT) and PBDITA:** Strong correlation (0.99042), indicating that changes in operating income before depreciation, interest, and taxes are closely followed by changes in net income after taxes.
- **Total Income and Total Expenses:** Extremely high correlation (0.99920), suggesting that increases in income are almost entirely offset by increases in expenses.
- Total Income has a strong positive correlation with Total Expenses (0.99920) and Sales (0.99987).
- Profit after Tax (PAT) is highly correlated with PBDITA (0.99042) and PBT (0.99514).
- Debtors Turnover shows some notable correlations with other turnover metrics like Finished Goods Turnover (0.72799).

Encode the data:
There are no categorical variables to encode the data.

Train-test split:

- Let us create the X and y variable data with respect to 'Networth Next Year' column as the target variable. Now X having every data except the target variable and y having only the target variable.

- Train-test split is essential in machine learning to evaluate a model's performance on new data, prevent overfitting, and facilitate model comparison and parameter tuning. It provides insights into how well a model generalizes to real-world scenarios.

- Using sklearn to split the data into x_train and y_train in 70:30 ratio. This split ratio provides a balance between having enough data to train the model effectively and having enough data to evaluate the model's performance reliably.

Details of train and test data are as follows:

The number of rows (observations) in TRAIN set is 2851
The number of columns (variables) in TRAIN set is 49

The number of rows (observations) in TEST set is 1405
The number of columns (variables) in TEST set is 49

Scaling:

- Applied Standard Scaler approach to scale variables. Standardization involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one.
- Scaling features ensures that all features contribute equally to the model's learning process, preventing features with larger scales from dominating, thus improving convergence and numerical stability, leading to better model performance and interpretability.

Renaming of Columns:

Columns are renamed to remove special characters and spaces.

Target variable creation:
The target variable is default and should take the value 1 when net worth next year is negative & 0 when net worth next year is positive,
Count of non-default cases: 4022
Count of Default cases: 234

Target variable is renamed to "Default".
The dataset is highly imbalanced.

Model building:

Metrics of Choice:

When evaluating machine learning models, especially in classification tasks, choosing the right metrics is crucial for understanding the performance and making informed decisions. Two commonly used metrics are Accuracy and F1 Score. Each has its strengths and is suitable for different scenarios.

- Accuracy is a good measure when the classes in the dataset are roughly equal in number.
- The F1 Score is particularly useful when the dataset has imbalanced classes. It gives a better measure of the incorrectly classified cases than Accuracy.
- The F1 Score takes both false positives and false negatives into account. It is especially useful when the cost of false positives and false negatives is different.
- It provides a more balanced view of the model's performance on imbalanced datasets, avoiding the misleadingly high accuracy that can occur when the majority class dominates.

Dealing with multicollinearity using VIF:

	variables	VIF
0	Total assets	inf
47	Total liabilities	inf
42	Raw material turnover	inf
41	WIP turnover	inf
34	Current ratio (times)	inf
33	Quick ratio (times)	inf
26	Total term liabilities / tangible net worth	inf
25	TOL/TNW	inf
2	Total income	295.512410
14	Sales	271.146471
4	Total expenses	190.932831
22	Shareholders funds	153.262016
1	Net worth	135.501818
24	Capital employed	82.394357
7	PBT	59.432840
5	Profit after tax	57.433628
44	Equity face value	39.550491
6	PBDITA	30.855003
10	PBT as % of total income	27.960840
11	PAT as % of total income	26.767162
35	Debt to equity ratio (times)	23.696929
8	Cash profit	23.085734
31	Current assets	21.253714
20	Current liabilities & provisions	17.681516
18	Reserves and funds	13.166704
39	Debtors turnover	12.967013
45	EPS	10.903689
29	Net fixed assets	10.689036
46	Adjusted EPS	10.063205
40	Finished goods turnover	9.626388
23	Cumulative retained profits	8.856594
19	Borrowings	8.043612
12	Cash profit as % of total income	7.770418
9	PBDITA as % of total income	6.640338
38	Creditors turnover	6.346895
17	Total capital	6.326848
21	Deferred tax liability	5.718330
43	Shares outstanding	5.501061
36	Cash to current liabilities (times)	4.929700
28	Contingent liabilities	4.414757
37	Cash to average cost of sales per day	3.679949
13	PAT as % of net worth	3.452657
15	Income from fincial services	2.631759
30	Investments	2.223129
27	Contingent liabilities / Net worth (%)	2.096023
16	Other income	2.062320
32	Net working capital	1.896121
3	Change in stock	1.265438
48	PE on BSE	1.209974

Table 6– VIF

Here, we see that the value of VIF is high for many variables. Here, we may drop variables with VIF more than 5 (very high correlation) & build our model.

Logistic Regression:

Considering only the variabes with VIF less than equal to 5 and Fitting the logistic regression model.

Equation :

f_1 = 'Default ~ Net_fixed_assets + Contingent_liabilities + Cash_profit_as_percent_of_total_income + Cash_to_average_cost_of_sales_per_day + Debtors_turnover + Total_term_liabilities_to_tangible_net_worth + Borrowings + Deferred_tax_liability + PAT_as_percent_of_net_worth + Cumulative_retained_profits + PE_on_BSE + Shares_outstanding + Quick_ratio + Contingent_liabilities_to_net_worth_percent + Income_from_financial_services + Investments + Other_income + Adjusted_EPS + Net_working_capital + Change_in_stock'

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	2851
Model:	Logit	**Df Residuals:**	2830
Method:	MLE	**Df Model:**	20
Date:	Sat, 27 Jul 2024	**Pseudo R-squ.:**	0.4375
Time:	21:20:56	**Log-Likelihood:**	-345.05
converged:	True	**LL-Null:**	-613.44
Covariance Type:	nonrobust	**LLR p-value:**	5.692e-101

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.9536	0.294	-16.850	0.000	-5.530	-4.377
Net_fixed_assets	-0.5386	0.341	-1.581	0.114	-1.206	0.129
Contingent_liabilities	-0.5131	0.214	-2.400	0.016	-0.932	-0.094
Cash_profit_as_percent_of_total_income	-0.3214	0.131	-2.455	0.014	-0.578	-0.065
Cash_to_average_cost_of_sales_per_day	0.2662	0.098	2.720	0.007	0.074	0.458
Debtors_turnover	0.3117	0.101	3.102	0.002	0.115	0.509
Total_term_liabilities_to_tangible_net_worth	0.6742	0.103	6.522	0.000	0.472	0.877
Borrowings	-0.0417	0.323	-0.129	0.897	-0.675	0.591
Deferred_tax_liability	0.0387	0.290	0.133	0.894	-0.530	0.607
PAT_as_percent_of_net_worth	-0.4723	0.120	-3.939	0.000	-0.707	-0.237
Cumulative_retained_profits	-0.7421	0.292	-2.538	0.011	-1.315	-0.169
PE_on_BSE	-0.0824	0.110	-0.750	0.453	-0.298	0.133
Shares_outstanding	0.0999	0.194	0.514	0.607	-0.281	0.481
Quick_ratio	0.0491	0.111	0.442	0.658	-0.169	0.267
Contingent_liabilities_to_net_worth_percent	0.3800	0.109	3.474	0.001	0.166	0.594
Income_from_financial_services	0.4050	0.191	2.116	0.034	0.030	0.780
Investments	0.0736	0.171	0.429	0.668	-0.262	0.410
Other_income	-0.2366	0.183	-1.294	0.196	-0.595	0.122
Adjusted_EPS	-0.8041	0.212	-3.792	0.000	-1.220	-0.388
Net_working_capital	-0.2376	0.192	-1.235	0.217	-0.614	0.139
Change_in_stock	0.2803	0.151	1.860	0.063	-0.015	0.576

Table 7–Logistic Eq1 summary

Pseudo R-squared: 0.43751108611789713

We can see that few variables are insignificant & may not be useful to discriminate cases of default. We will try & remove variables whose p value is greater than 0.05 & rebuild our model.

Equation 2:
Updated formula with insignificant variables removed
f_2 = 'Default ~ Contingent _liabilities + Cash_profit_as_percent_of_total_income + Debtors_turnover + Total_term_liabilities_to_tangible_net_worth + PAT_as_percent_of_net_worth + Contingent_liabilities_to_net_worth_percent + Income_from_financial_services + Investments + Adjusted_EPS'

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	2851
Model:	Logit	**Df Residuals:**	2841
Method:	MLE	**Df Model:**	9
Date:	Sat, 27 Jul 2024	**Pseudo R-squ.:**	0.4112
Time:	21:21:03	**Log-Likelihood:**	-361.18
converged:	True	**LL-Null:**	-613.44
Covariance Type:	nonrobust	**LLR p-value:**	6.221e-103

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.3745	0.192	-22.781	0.000	-4.751	-3.998
Contingent_liabilities	-0.6817	0.170	-4.010	0.000	-1.015	-0.348
Cash_profit_as_percent_of_total_income	-0.3845	0.142	-2.698	0.007	-0.664	-0.105
Debtors_turnover	0.2005	0.089	2.250	0.024	0.026	0.375
Total_term_liabilities_to_tangible_net_worth	0.6759	0.094	7.158	0.000	0.491	0.861
PAT_as_percent_of_net_worth	-0.5755	0.121	-4.759	0.000	-0.812	-0.338
Contingent_liabilities_to_net_worth_percent	0.4985	0.101	4.916	0.000	0.300	0.697
Income_from_financial_services	0.1485	0.168	0.884	0.377	-0.181	0.478
Investments	0.0675	0.155	0.437	0.662	-0.235	0.370
Adjusted_EPS	-0.7210	0.179	-4.028	0.000	-1.072	-0.370

Table 8–Logistic Eq 2 Summary

Pseudo R-squared: 0.41121509200526085

Results on train data:

Let us now check the confusion matrix and the classification report for train data

	precision	recall	f1-score	support
0	0.965	0.986	0.975	2692
1	0.618	0.396	0.483	159
accuracy			0.953	2851
macro avg	0.791	0.691	0.729	2851
weighted avg	0.946	0.953	0.948	2851

Accuracy: 0.953
F1 Score: 0.483

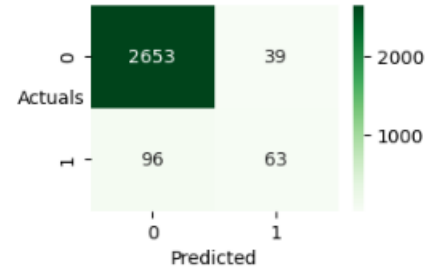


Fig.5 Logistic regression Eq2 - train

Results on test data:

Let us now check the confusion matrix and the classification report for test data followed by the AUC and the AUC-ROC curve.

	precision	recall	f1-score	support
0	0.957	0.989	0.973	1330
1	0.533	0.213	0.305	75
accuracy			0.948	1405
macro avg	0.745	0.601	0.639	1405
weighted avg	0.934	0.948	0.937	1405

Accuracy: 0.948
F1 Score: 0.305

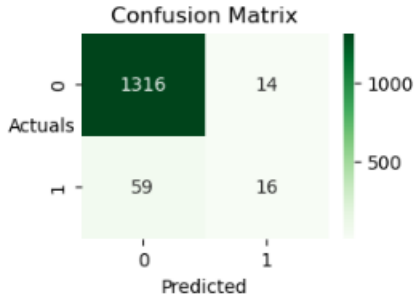


Fig.6 Logistic regression Eq2 - test

Identify optimal threshold for Logistic Regression using ROC curve :

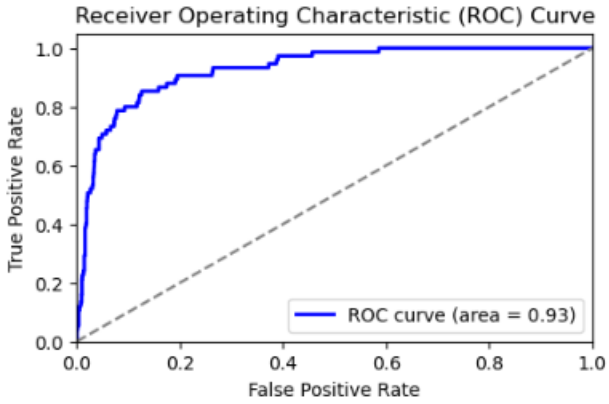


Fig.7 Logistic regression ROC curve

From your ROC curve, the point that appears to be farthest from the diagonal line is around a False Positive Rate (FPR) of 0.2 and a True Positive Rate (TPR) of 0.8. This is usually the point where the sum of True Positive Rate (TPR) and True Negative Rate (1 - False Positive Rate, FPR) is maximized.

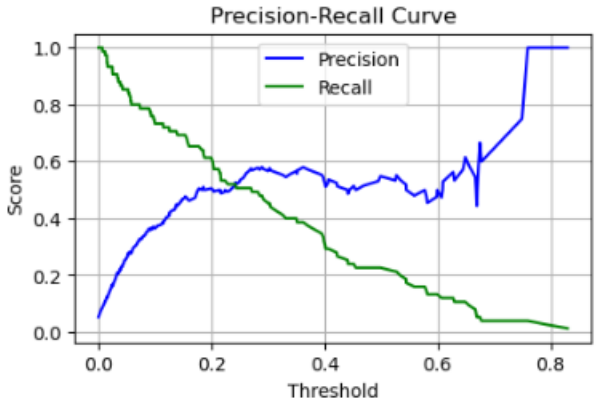


Fig.8 Logistic regression Precision - Recall curve

We can see that at 0.2 threshold we can see both Precision and recall are crossing each other. So, let's select 0.2 as threshold.

Results on test data:

	precision	recall	f1-score	support
0	0.977	0.966	0.972	1330
1	0.500	0.600	0.545	75
accuracy			0.947	1405
macro avg	0.739	0.783	0.759	1405
weighted avg	0.952	0.947	0.949	1405

Accuracy: 0.947
F1 Score: 0.545

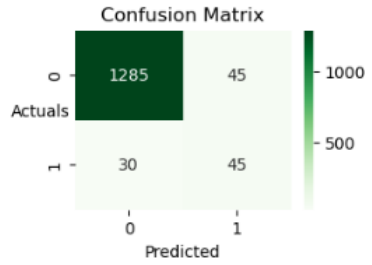


Fig.9 Logistic regression results

Model performance check across different metrics

- The accuracy remains almost the same between the two thresholds, showing that the overall prediction capability does not drastically change.
- The F1 score shows a notable increase when the threshold is lowered to 0.2. This indicates that the model is better at balancing precision and recall at this threshold, making it more effective for situations where detecting the positive class is crucial.

Random Forest Classifier:

Results on test data:

	precision	recall	f1-score	support
0	0.960	0.986	0.973	1330
1	0.525	0.280	0.365	75
accuracy			0.948	1405
macro avg	0.743	0.633	0.669	1405
weighted avg	0.937	0.948	0.940	1405

Accuracy: 0.9480427046263346
F1 Score: 0.365

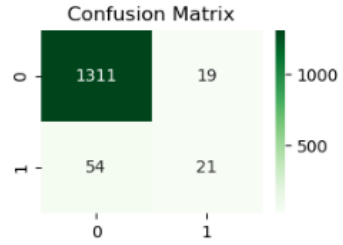


Fig.9 Random Forest results

Handling Imbalance:

- SMOTE (Synthetic Minority Over-sampling Technique) is a popular technique used to address the problem of class imbalance in machine learning datasets.
- When dealing with imbalanced datasets, where one class is significantly underrepresented compared to the other, models can become biased towards the majority class. SMOTE helps mitigate this issue by generating synthetic samples for the minority class.
- Balanced datasets lead to better performance of machine learning models, especially in terms of metrics like precision, recall, and F1 score.
- Random Forest is a robust ensemble learning method that is often less sensitive to class imbalance compared to simpler models. However, in cases of severe class imbalance, even Random Forest can benefit from techniques like SMOTE

No of training samples after Smote: 2851

Results on test data:

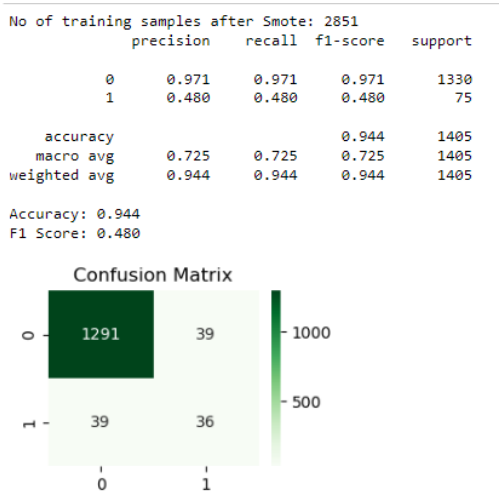
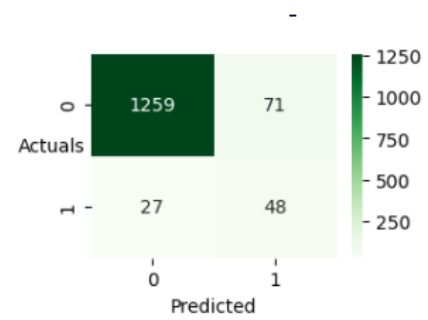


Fig.10 Random Forest with Smote results

Smote did not help much as performance is not improved.

Hyperparameter Tuning for Random Forest :

param_grid = {
'min_samples_split' : [10,25,50],
'min_samples_leaf':[5,15,30],
'max_depth':[5,10],
'n_estimators' : [50,100,200]
}
Grid search is applied on Random forest Parameters.



	precision	recall	f1-score	support
0	0.979	0.947	0.963	1330
1	0.403	0.640	0.495	75
accuracy			0.930	1405
macro avg	0.691	0.793	0.729	1405
weighted avg	0.948	0.930	0.938	1405

Accuracy: 0.930
F1 Score: 0.495

Fig.11 Random Forest with Grid Search results

Identify optimal threshold for Random Forest using Precision – recall curve :

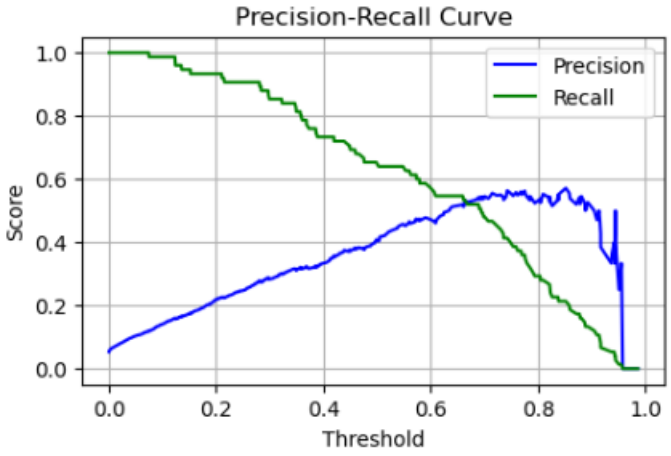
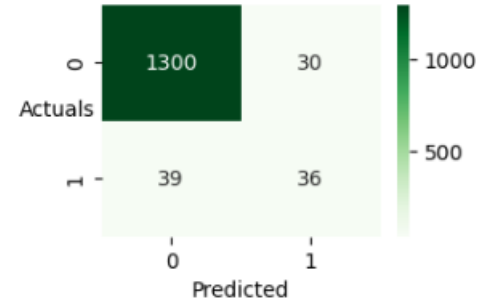


Fig.12 Random Forest with Grid Search Precision-Recall curve

We can see that at 0.3 threshold we can see both Precision and recall are crossing each other. So lets select 0.7 as threshold.

Results on test data:



	precision	recall	f1-score	support
0	0.971	0.977	0.974	1330
1	0.545	0.480	0.511	75
accuracy			0.951	1405
macro avg	0.758	0.729	0.742	1405
weighted avg	0.948	0.951	0.949	1405

Accuracy: 0.951
F1 Score: 0.511

Fig.12 Random Forest with Grid Search results

Model performance check across different metrics

- The accuracy remains almost the same between the two thresholds, showing that the overall prediction capability does not drastically change.
- The F1 score shows a notable increase when the threshold is lowered to 0.7. This indicates that the model is better at balancing precision and recall at this threshold, making it more effective for situations where detecting the positive class is crucial.

Compare the performance of the models:

Model	Threshold	Accuracy	F1_Score
Logistic regression Equation 2	0.5	0.948	0.305
Logistic regression Equation 2	0.2	0.947	0.545
Random Forest	0.5	0.948	0.346
Random Forest	0.3	0.947	0.483
Random Forest with smote	0.5	0.94	0.48
Grid search Random forest model	0.5	0.93	0.495
Grid search Random forest model	0.7	0.951	0.511

Table 9– Model Evaluation

- **Grid Search Random Forest Model (Threshold 0.7)** perform best and selected as **Final model**, balancing accuracy and F1 score effectively.
- **Logistic Regression Equation 2 (Threshold 0.2)** has a high F1 score, indicating good performance in handling imbalanced data, even though it has slightly lower accuracy.
- **Random Forest with SMOTE** improves F1 score significantly compared to the standard Random Forest model, indicating better handling of imbalanced data.
- **Grid Search Random Forest Model (Threshold 0.5)** has the highest accuracy, but its F1 score is lower than the models with a 0.3 threshold.
- **Logistic Regression Equation 2** with 0.5 threshold performs the worst, with the lowest F1 score despite high accuracy.

Check the most important features in the final model and draw inferences

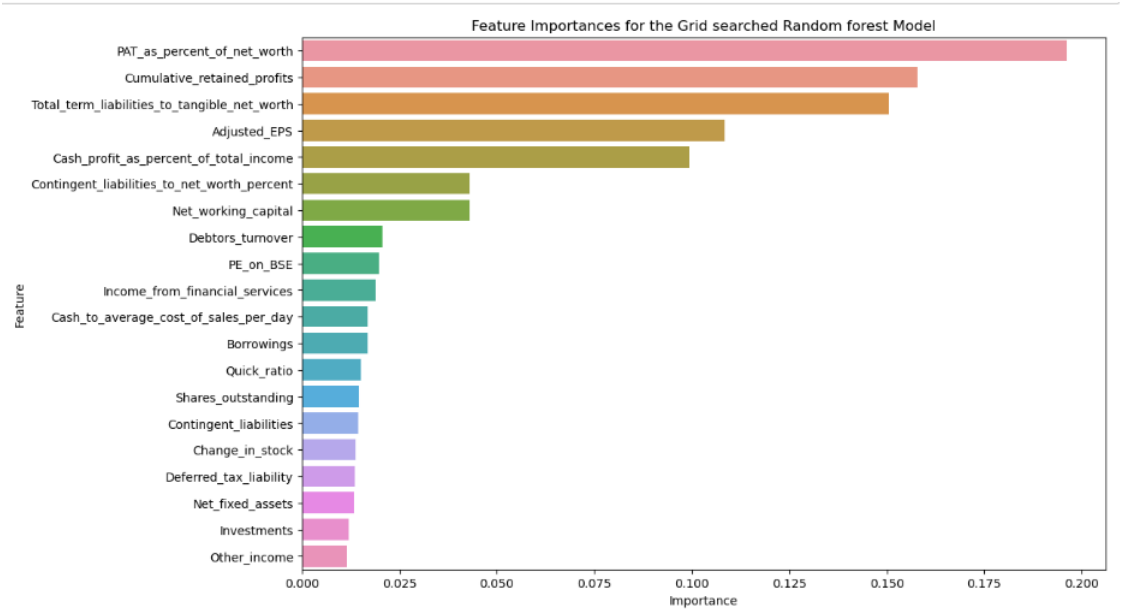


Fig.13 Feature importances

1. **Top Features**:
    - The most important feature is **PAT as percent of net worth**, followed by **Cumulative retained profits**, and **Total term liabilities to tangible net worth**.
2. **Profitability and Financial Stability**:
    - Features related to profitability and financial stability, such as **PAT as percent of net worth**, **Cumulative retained profits**, and **Cash profit as percent of total income**, are among the top features. This suggests that profitability metrics and retained earnings are crucial in determining the outcome, which could relate to financial health or creditworthiness.

3. **Liabilities and Leverage**:
    - **Total term liabilities to tangible net worth** and **Contingent liabilities to net worth percent** are also highly important features. These ratios are key indicators of leverage and financial risk, implying that higher liabilities relative to net worth may significantly impact the model's predictions.
4. **Earnings and Working Capital**:
    - **Adjusted EPS (Earnings Per Share)** and **Net working capital** are also important, emphasizing the role of earnings performance and liquidity in the model's predictive capabilities. This highlights the importance of a company's ability to generate earnings and maintain sufficient working capital for operations.
5. **Industry and Market Indicators**:
    - **PE on BSE (Price-to-Earnings ratio on Bombay Stock Exchange)** is included, indicating that market valuation metrics also play a role in the model. This could reflect investor sentiment and market conditions impacting the company's financial status.
6. **Operational Efficiency**:
    - **Debtors' turnover** and **Cash to average cost of sales per day** are operational efficiency metrics that are considered important. Efficient management of receivables and cash flow operations are likely significant factors in the model.
7. **Diverse Influences**:
    - The model considers a wide range of features, including **Income from financial services**, **Quick ratio**, **Borrowings**, and **Shares outstanding**, among others. This diversity suggests a comprehensive approach, capturing various aspects of financial performance and risk.

Actionable Insights & Recommendations

1. **Enhance Profitability**:
    a. **Increase PAT (Profit After Tax) as Percent of Net Worth**: Focus on strategies that boost net profits, such as cost optimization, improving operational efficiencies, and increasing revenue through market expansion or new product lines.
    b. **Retain More Profits**: Strengthen policies to retain earnings and reinvest them into the business. This can enhance cumulative retained profits and provide a buffer for future investments or downturns.
2. **Manage Liabilities Effectively**:
    a. **Reduce Term Liabilities**: Work on reducing long-term debts and liabilities to improve the total term liabilities to tangible net worth ratio. This can be achieved through better debt management and refinancing high-interest debt.

   b. **Mitigate Contingent Liabilities**: Implement robust risk management practices to reduce contingent liabilities. This could include better contract management, insurance, and hedging strategies to mitigate potential financial risks.
3. **Improve Earnings and Liquidity**:
   a. **Increase Earnings Per Share (EPS)**: Focus on increasing net income while managing the number of shares outstanding. This could involve profit growth strategies and share buyback programs.
   b. **Optimize Working Capital**: Enhance the efficiency of working capital management by reducing inventory levels, speeding up receivables, and extending payables where feasible. Efficient working capital management can significantly improve liquidity.
4. **Boost Operational Efficiency**:
   a. **Improve Debtors Turnover**: Implement stronger credit control measures to ensure faster collection of receivables. This can include stricter credit policies, better customer credit assessment, and efficient collection processes.
   b. **Enhance Cash Management**: Focus on improving the cash-to-average cost of sales per day ratio. This can be achieved through better cash flow forecasting, maintaining adequate cash reserves, and optimizing the timing of cash outflows.
5. **Market and Investment Strategies**:
   a. **Monitor Market Valuations**: Keep an eye on market valuations, such as the Price-to-Earnings (PE) ratio. Strategic decisions around market positioning, investor relations, and communication can help in maintaining a favorable market perception.
   b. **Invest in Growth Areas**: Allocate investments in high-potential areas that align with market trends and company strengths. This includes investing in technology, innovation, and expanding into new markets or product lines.
6. **Leverage Financial Services and Ratios**:
   a. **Diversify Income Streams**: Increase income from financial services if applicable to your business model. Diversifying income streams can reduce dependence on core operations and spread risk.
   b. **Monitor Quick Ratio**: Regularly monitor and aim to improve the quick ratio to ensure that the company can meet its short-term obligations without relying on the sale of inventory.

Conclusion:

By focusing on these strategic areas, companies can enhance their financial stability, operational efficiency, and overall market performance. These decisions should be informed by continuous monitoring and analysis of key financial metrics to adapt to changing business environments and maintain a competitive edge.