FINANCE AND RISK ANALYTICS (FRA) PROJECT-CODED

BY — *Harsh Patel*

24th November 2024

Problem-A:- Financial Health Assessment Tool

In the realm of modern finance, businesses encounter the perpetual challenge of managing debt obligations effectively to maintain a favourable credit standing and foster sustainable growth. Investors keenly scrutinise companies capable of navigating financial complexities while ensuring stability and profitability. A pivotal instrument in this evaluation process is the balance sheet, which provides a comprehensive overview of a company's assets, liabilities, and shareholder equity, offering insights into its financial health and operational efficiency. In this context, leveraging available financial data, particularly from preceding fiscal periods, becomes imperative for informed decision-making and strategic planning.

Objective

A group of venture capitalists want to develop a Financial Health Assessment Tool. With the help of the tool, it endeavours to empower businesses and investors with a robust mechanism for evaluating the financial well-being and creditworthiness of companies. By harnessing machine learning techniques, they aim to analyse historical financial statements and extract pertinent insights to facilitate informed decision-making via the tool. Specifically, they foresee facilitating the following with the help of the tool:

1. Debt Management Analysis: Identify patterns and trends in debt management practices to assess the ability of businesses to fulfil financial obligations promptly and efficiently, and identify potential cases of default.
2. 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.

They have hired you as a data scientist and provided you with the financial metrics of different companies. The task is to analyse the data provided and develop a predictive model leveraging machine learning techniques to identify whether a given company will be tagged as a defaulter in terms of net worth next year. The predictive model will help the organisation anticipate potential challenges with the financial performance of the companies and enable proactive risk mitigation strategies.

Data Dictionary

The data consists of financial metrics from the balance sheets of different companies. The detailed data dictionary is given below:

- Net Worth Next Year: Net worth of the customer in the next year
- Total assets: Total assets of customer
- Net worth: Net worth of the customer of the present year
- Total income: Total income of the customer
- Change in stock: Difference between the current value of the stock and the value of stock in the last trading day
- Total expenses: Total expenses done by the customer
- Profit after tax: Profit after tax deduction
- PBDITA: Profit before depreciation, income tax, and amortisation
- PBT: Profit before tax deduction
- Cash profit: Total Cash profit
- PBDITA as % of total income: PBDITA / Total income
- PBT as % of total income: PBT / Total income
- PAT as % of total income: PAT / Total income
- Cash profit as % of total income: Cash Profit / Total income
- PAT as % of net worth: PAT / Net worth

- Sales: Sales done by the customer
- Income from financial services: Income from financial services
- Other income: Income from other sources
- Total capital: Total capital of the customer
- Reserves and funds: Total reserves and funds of the customer
- Borrowings: Total amount borrowed by the customer
- Current liabilities & provisions: current liabilities of the customer
- Deferred tax liability: Future income tax customer will pay because of the current transaction
- Shareholders funds: Amount of equity in a company which belongs to shareholders
- Cumulative retained profits: Total cumulative profit retained by customer
- Capital employed: Current asset minus current liabilities
- TOL/TNW: Total liabilities of the customer divided by Total net worth
- Total term liabilities / tangible net worth: Short + long term liabilities divided by tangible net worth
- Contingent liabilities / Net worth (%): Contingent liabilities / Net worth
- Contingent liabilities: Liabilities because of uncertain events
- Net fixed assets: The purchase price of all fixed assets
- Investments: Total invested amount
- Current assets: Assets that are expected to be converted to cash within a year
- Net working capital: Difference between the current liabilities and current assets
- Quick ratio (times): Total cash divided by current liabilities
- Current ratio (times): Current assets divided by current liabilities
- Debt to equity ratio (times): Total liabilities divided by its shareholder equity
- Cash to current liabilities (times): Total liquid cash divided by current liabilities

- Cash to average cost of sales per day: Total cash divided by the average cost of the sales
- Creditors turnover: Net credit purchase divided by average trade creditors
- Debtors turnover: Net credit sales divided by average accounts receivable
- Finished goods turnover: Annual sales divided by average inventory
- WIP turnover: The cost of goods sold for a period divided by the average inventory for that period
- Raw material turnover: Cost of goods sold is divided by the average inventory for the same period
- Shares outstanding: Number of issued shares minus the number of shares held in the company
- Equity face value: cost of the equity at the time of issuing
- EPS: Net income divided by the total number of outstanding share
- Adjusted EPS: Adjusted net earnings divided by the weighted average number of common shares outstanding on a diluted basis during the plan year
- Total liabilities: Sum of all types of liabilities
- PE on BSE: Company's current stock price divided by its earnings per share

1.1. Problem Definition and Exploratory Data Analysis:

First, we will look at the first and last five rows using function head and tail respectively, of the dataset from the csv file called comp_Fin_data.csv that we loaded using the read csv function. In table-1 and table-2 below shows the dataset.

	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	Adjusted EPS
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	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	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	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	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	-6.52

5 rows × 51 columns

Table-1 First Five rows of Dataset.

	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	Adjus
4251	4252	0.2	0.4	0.2	NaN	NaN	NaN	NaN	NaN	NaN	...	0.00	NaN	NaN	0.00	NaN	NaN	0.00	(
4252	4253	93.3	159.6	86.7	172.9	0.1	169.7	3.3	18.4	3.7	...	1.80	11.00	8.28	9.88	8162700.0	10.0	0.42	(
4253	4254	932.2	833.8	664.6	2314.7	32.1	2151.6	195.2	348.4	303.0	...	6.08	59.28	31.14	9.87	7479762.0	10.0	26.58	2(
4254	4255	64.6	95.0	48.5	110.5	4.6	113.5	1.6	9.7	2.6	...	3.71	78.99	11.51	14.95	NaN	NaN	0.00	(
4255	4256	0.0	384.6	111.3	345.8	11.3	341.7	15.4	57.6	20.7	...	4.71	53.37	8.33	3.74	960000.0	10.0	15.63	1!

5 rows × 51 columns

Table-2 Last Five rows of Dataset.

Now, we use the shape function of the dataset and we get that there are 4256 rows and 51 columns. Then, we used the info function and found out the data type of each column and used value counts functions on the categorical variables as shown in the below table. we will check for the duplicated rows are present or not using duplicate function and found out that in the dataset there are zero same or duplicated rows present.

```
<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)
```

Table-3 Information of Dataset

We used describe function and obtained the five important summaries namely count, mean, std, min, max as shown in Table-4 below.

	Num	Networth Next Year	Total assets	Net worth	Total income	Change in stock	Total expenses	Profit after tax	PBDITA	PBT	...	Debto turnov
count	4256.000000	4256.000000	4.256000e+03	4256.000000	4.025000e+03	3706.000000	4.091000e+03	4102.000000	4102.000000	4102.000000	...	3871.0000(
mean	2128.500000	1344.740883	3.573617e+03	1351.949601	4.688190e+03	43.702482	4.356301e+03	295.050585	605.940639	410.259044	...	17.9290
std	1228.745702	15936.743168	3.007444e+04	12961.311651	5.391895e+04	436.915048	5.139809e+04	3079.902071	5646.230633	4217.415307	...	90.1644
min	1.000000	-74265.600000	1.000000e-01	0.000000	0.000000e+00	-3029.400000	-1.000000e-01	-3908.300000	-440.700000	-3894.800000	...	0.0000(
25%	1064.750000	3.975000	9.130000e+01	31.475000	1.071000e+02	-1.800000	9.680000e+01	0.500000	6.925000	0.800000	...	3.8100(
50%	2128.500000	72.100000	3.155000e+02	104.800000	4.551000e+02	1.600000	4.268000e+02	9.000000	36.900000	12.600000	...	6.4700(
75%	3192.250000	330.825000	1.120800e+03	389.850000	1.485000e+03	18.400000	1.395700e+03	53.300000	158.700000	74.175000	...	11.8500(
max	4256.000000	805773.400000	1.176509e+06	613151.600000	2.442828e+06	14185.500000	2.366035e+06	119439.100000	208576.500000	145292.600000	...	3135.2000(

8 rows × 51 columns

Table-4 Description of dataset.

We can drop two variables Num and Adjusted EPS as Num variable is nothing but it acts as an index and Adjusted EPS to maintain consistency, focus on standard metrics, simplify the analysis, ensure data reliability, align with GAAP compliance, or meet specific analysis objectives. By focusing on regular EPS, we can often achieve a more straightforward and comparative evaluation.

Univariate Analysis:

Now, we create a new variable called Default in the dataset using Networth next year variable and obtain the below figure-1 and can be said that 78.76% are defaulters. In this dataset, there are significantly more instances where no default occurred (category '1') compared to instances where a default did occur (category '0'). This suggests that non-default cases are more common than default cases. A company will be classified as a non-defaulter if its net worth is positive in the following year. Conversely, if a company's net worth is negative, it will be classified as a defaulter.

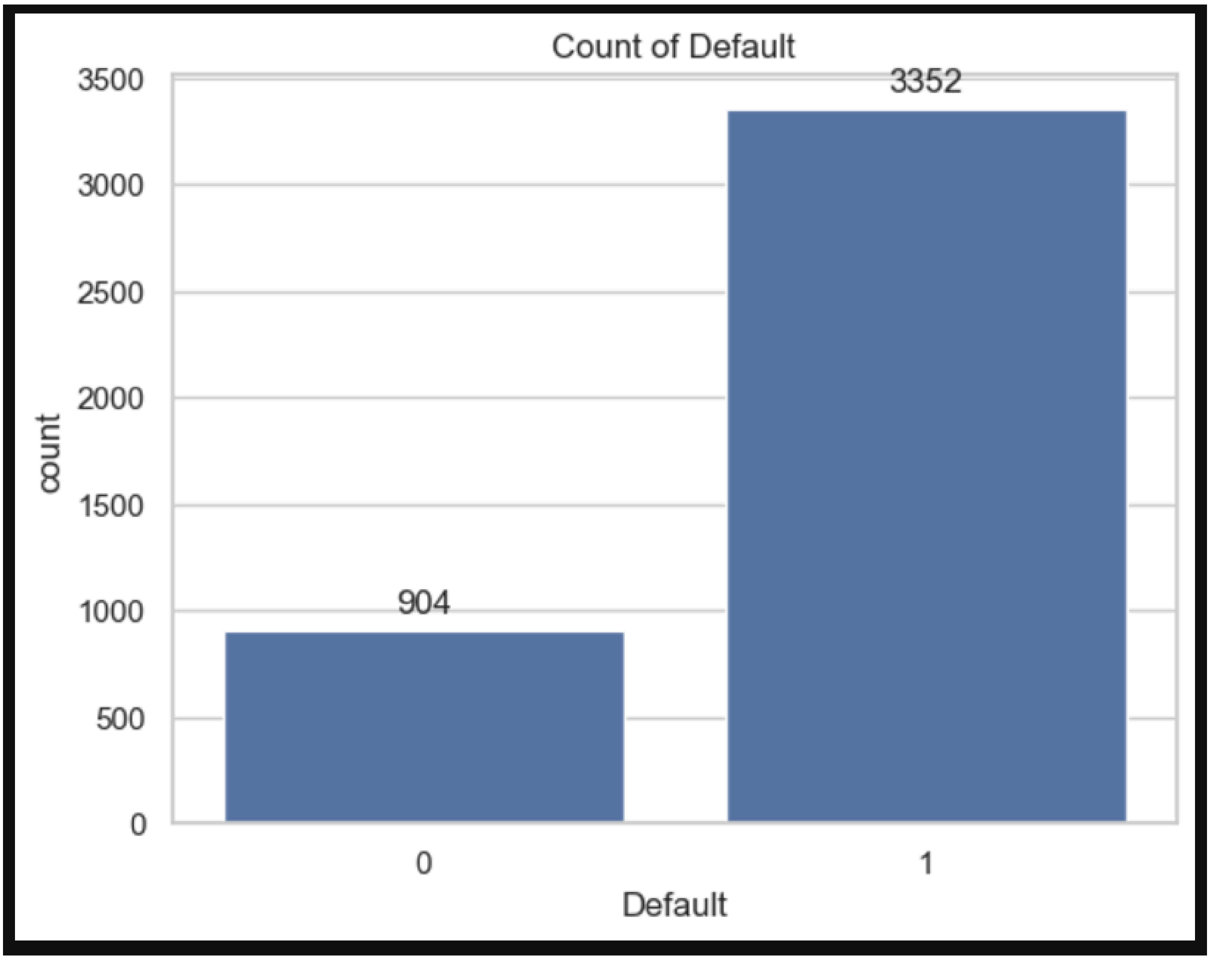
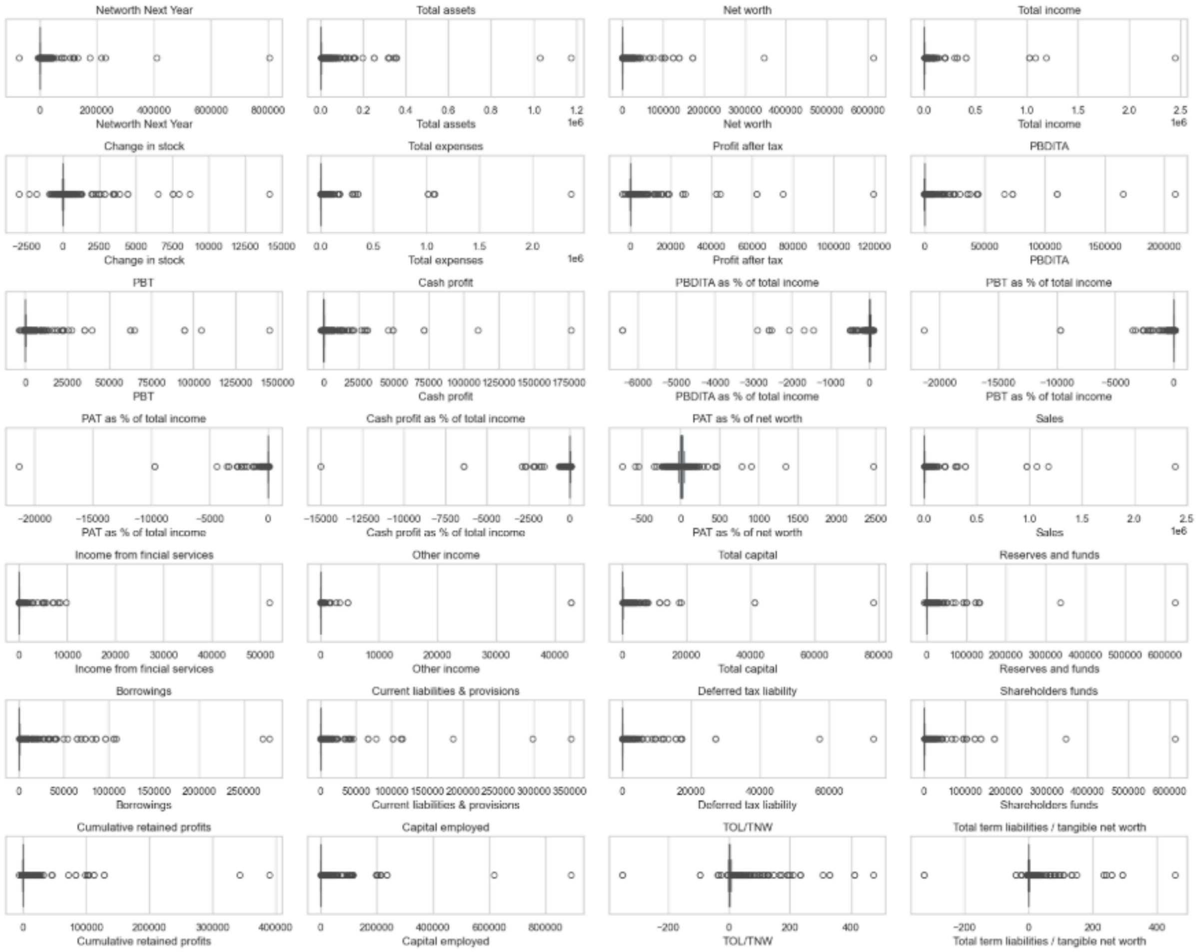


Fig-1 Count of Default.

We can see the univariate analysis of all the variables in the form of boxplot and Histogram as seen in figure-2 and can see that there are outliers present in the variables that needed to be treated.


Networth Next Year
Total assets
Net worth
Total income
Change in stock
Total expenses
Profit after tax
PBDITA
PBT
Cash profit
PBDITA as % of total income
PBT as % of total income
PAT as % of total income
Cash profit as % of total income
PAT as % of net worth
Sales
Income from fincial services
Other income
Total capital
Reserves and funds
Borrowings
Current liabilities & provisions
Deferred tax liability
Shareholders funds
Cumulative retained profits
Capital employed
TOL/TNW
Total term liabilities / tangible net worth

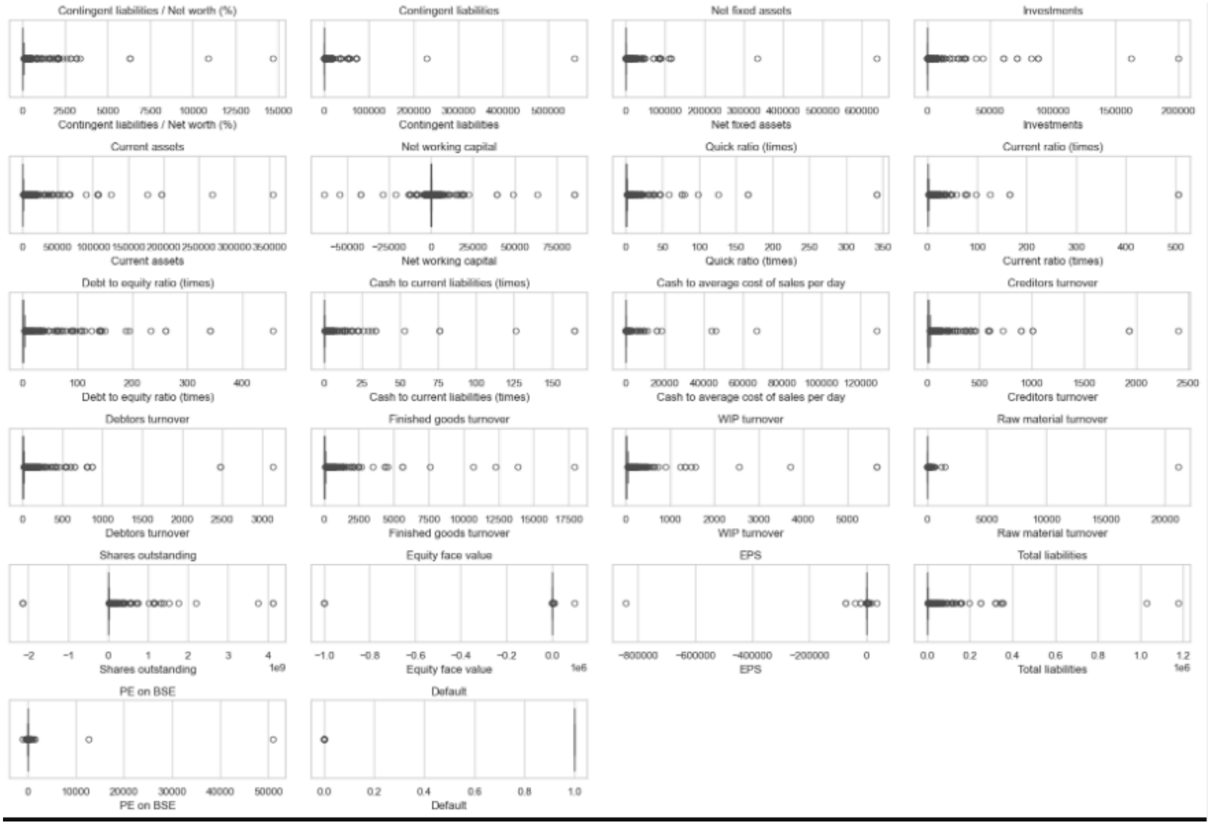


Fig-2 Boxplot before outlier treatment

Multivariate Analysis:

We will create a Correlation matrix as seen in table-5 and we can visualise it in heatmap as seen in figure-3.

	Networth Next Year	Total assets	Net worth	Total income	Change in stock	Total expenses	Profit after tax	PBDI
Networth Next Year	1.000000	0.877803	0.930135	0.710953	0.345199	0.690526	0.867992	0.8723
Total assets	0.877803	1.000000	0.959404	0.868607	0.470735	0.852863	0.907560	0.9433
Net worth	0.930135	0.959404	1.000000	0.783831	0.393760	0.761549	0.954399	0.9629
Total income	0.710953	0.868607	0.783831	1.000000	0.276395	0.999203	0.727438	0.7932
Change in stock	0.345199	0.470735	0.393760	0.276395	1.000000	0.273717	0.366994	0.3895

Table-5 Correlation Matrix

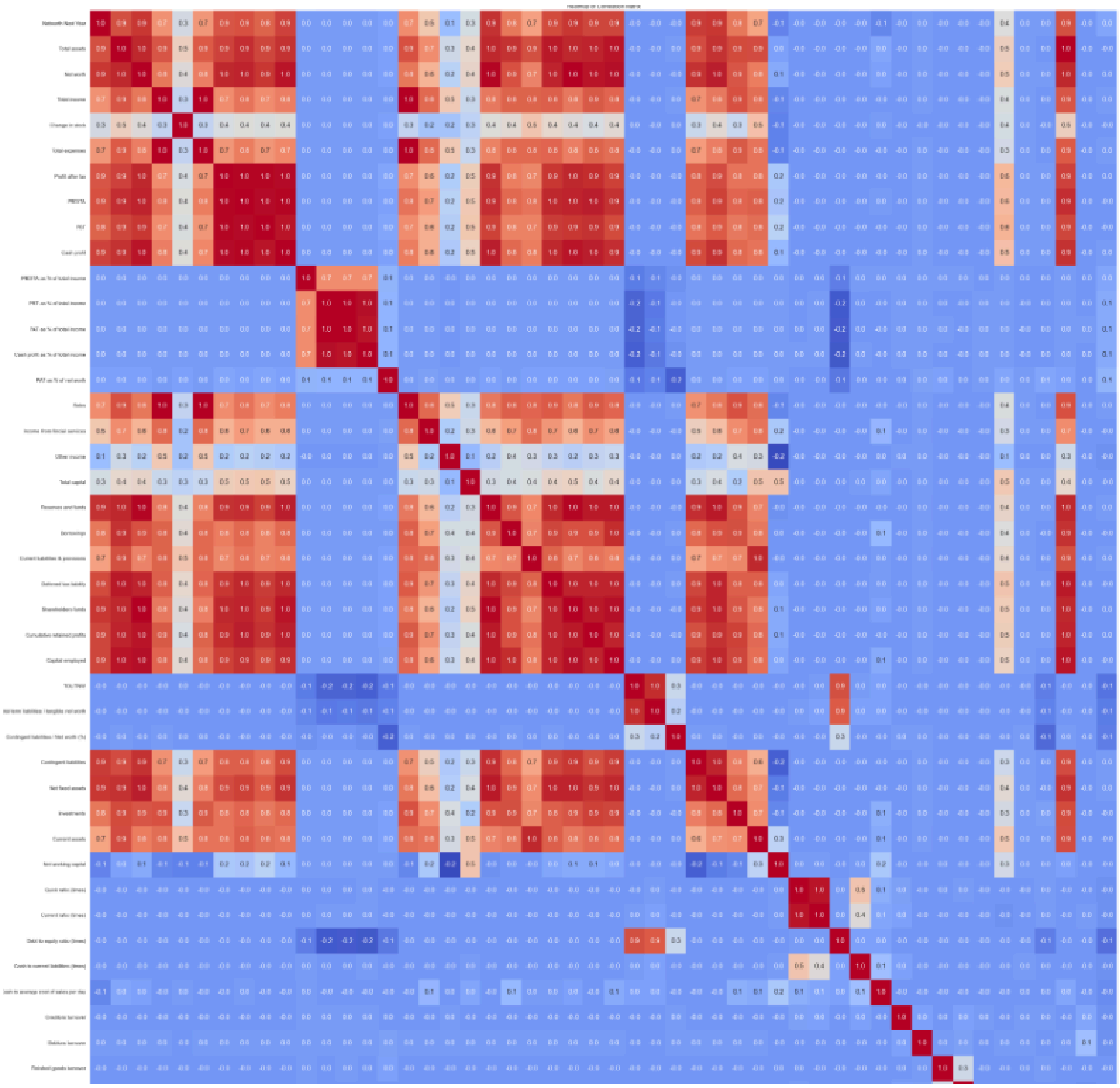


Fig-3 Heatmap of Correlation Matrix.

Observations:

1. Net Worth Next Year:

- Strong Positive Correlations: There is a significant positive correlation between Networth Next Year and Net Worth (0.930), Total Assets (0.878), and Profit After Tax (0.868). This suggests that these factors are strong predictors of a company's future net worth.
- Moderate Positive Correlations: There are moderate positive correlations with Total Income (0.711) and Total Expenses (0.691), indicating they also play a role, albeit to a lesser extent.

- Weak Positive Correlations: The correlation with Change in Stock (0.345) is weaker, suggesting it has a minimal impact on predicting future net worth.

2. Total Assets:

- Strongest Positive Correlations: Total Assets are highly positively correlated with Net Worth (0.959), PBDITA (0.943), and Cash Profit (0.940), highlighting these as key components of a company's asset base.
- Strong Positive Correlations: There are also strong positive correlations with Profit After Tax (0.908), PBT (0.895), and Total Income (0.869), indicating these financial metrics significantly contribute to Total Assets.

3. Net Worth:

- Strongest Positive Correlations: Net Worth shows very high positive correlations with PBDITA (0.963), Cash Profit (0.978), and Profit After Tax (0.954), marking them as major drivers of a company's equity.
- Strong Positive Correlations: Additionally, strong positive correlations with Total Assets (0.959), PBT (0.932), and Total Income (0.784) further emphasize their importance.

4. Total Income:

- Extremely High Positive Correlation: Total Income is almost perfectly correlated with Total Expenses (0.999), signifying that as income increases, expenses follow suit, likely due to proportional operating costs.
- High Positive Correlations: There are high positive correlations with PBDITA (0.793), Cash Profit (0.763), and Net Worth (0.784), indicating that Total Income significantly impacts these metrics.
- Moderate Positive Correlation: The correlation with Change in Stock (0.276) is moderate, indicating some level of impact but not as strong.

5. Change in Stock:

- Weak to Moderate Positive Correlations: Change in Stock shows weak positive correlations with Total Income (0.276), Total Assets (0.471), and Net Worth (0.394). There are moderate positive correlations with Total Expenses (0.274) and Profit After Tax (0.367), suggesting limited influence on overall financial performance.

6. Total Expenses:

- Extremely High Positive Correlation: Total Expenses have an almost perfect correlation with Total Income (0.999), reflecting the direct relationship between the two.
- High Positive Correlations: High correlations with PBDITA (0.769), Profit After Tax (0.700), and Cash Profit (0.737) indicate that expenses are closely tied to these profitability measures.

7. Profit After Tax:

- Strongest Positive Correlations: Profit After Tax shows very high positive correlations with Cash Profit (0.990), PBDITA (0.990), and PBT (0.995), highlighting its alignment with other profitability metrics.
- Strong Positive Correlations: Strong correlations with Net Worth (0.954), Total Assets (0.908), and Networth Next Year (0.868) emphasize its importance in overall financial health.

8. PBDITA:

- Strongest Positive Correlations: PBDITA has very high positive correlations with Profit After Tax (0.990), PBT (0.989), and Cash Profit (0.992), underlining its significance in evaluating company performance.
- Strong Positive Correlations: Additionally, strong correlations with Net Worth (0.963), Total Assets (0.943), and Total Income (0.793) further highlight its relevance.

9. PBT:

- Strongest Positive Correlations: PBT shows very high positive correlations with Profit After Tax (0.995), PBDITA (0.989), and Cash Profit (0.978), indicating these as key profitability indicators.
- Strong Positive Correlations: Strong correlations with Net Worth (0.932), Total Assets (0.895), and Networth Next Year (0.834) suggest PBT's significant role in future financial outcomes.

10. Cash Profit:

- Strongest Positive Correlations: Cash Profit is extremely highly correlated with PBDITA (0.992), Profit After Tax (0.990), and PBT (0.978), marking it as a crucial measure of liquidity.
- Strong Positive Correlations: Strong correlations with Net Worth (0.978), Total Assets (0.940), and Networth Next Year (0.907) show its influence on financial stability.

Key Observations:

- Primary Predictors: Net Worth, Total Assets, and Profit After Tax are the most influential factors positively impacting Networth Next Year.
- In Tandem Movement: Total Income and Total Expenses move almost perfectly together, suggesting a direct proportional relationship.
- Profitability Metrics: PBDITA, PBT, and Cash Profit exhibit strong inter-correlations and with Net Worth and Total Assets, indicating their critical role in financial performance.
- Limited Impact: Change in Stock has the weakest correlations among the metrics analysed, indicating it has a lesser impact on other financial metrics.

1.2 Data Pre-processing

From the above we can drop some of the variables like Equity face value and Cash to current Liabilities for having a high correlation.

We need to do this because below we can see the list of outliers present in each variable and indicate that treatment needs to be done properly.

	Column	No. of outliers
0	Networth Next Year	624
1	Total assets	585
2	Net worth	595
3	Total income	508
4	Change in stock	750
5	Total expenses	518
6	Profit after tax	712
7	PBDITA	584
8	PBT	704
9	Cash profit	627
10	PBDITA as % of total income	346
11	PBT as % of total income	546
12	PAT as % of total income	610
13	Cash profit as % of total income	426
14	PAT as % of net worth	427
15	Sales	500
16	Income from fincial services	517
17	Other income	389

Table-6 Outliner Detected

Now, we will perform outlier treatment using q1 and q3 to find IQR and using that to find out the upper and lower limit whiskers and finally bring all those outlier's points to these whiskers.

After taking care of outliers we will separate the target variable from the rest of the variables. And split the data set into test and train dataset.

Now, using its null function we see how many empty or null values are present in each variable for each dataset as seen in table-7 and table-8

Networth Next Year	0
Total assets	0
Net worth	0
Total income	177
Change in stock	424
Total expenses	125
Profit after tax	114
PBDITA	114
PBT	114
Cash profit	114
PBDITA as % of total income	65
PBT as % of total income	65
PAT as % of total income	65
Cash profit as % of total income	65
PAT as % of net worth	0
Sales	237
Income from fincial services	848
Other income	1180
Total capital	5
Reserves and funds	75
Borrowings	328
Current liabilities & provisions	88
Deferred tax liability	1043
Shareholders funds	0
Cumulative retained profits	34
Capital employed	0
TOL/TNW	0
Total term liabilities / tangible net worth	0
Contingent liabilities / Net worth (%)	0
Contingent liabilities	1061
Net fixed assets	100
Investments	1290

Table-7 Missing values in Train dataset

Networth Next Year	0
Total assets	0
Net worth	0
Total income	54
Change in stock	126
Total expenses	40
Profit after tax	40
PBDITA	40
PBT	40
Cash profit	40
PBDITA as % of total income	14
PBT as % of total income	14
PAT as % of total income	14
Cash profit as % of total income	14
PAT as % of net worth	0
Sales	68
Income from fincial services	263
Other income	376
Total capital	0
Reserves and funds	23
Borrowings	103
Current liabilities & provisions	22
Deferred tax liability	326
Shareholders funds	0
Cumulative retained profits	11
Capital employed	0
TOL/TNW	0
Total term liabilities / tangible net worth	0
Contingent liabilities / Net worth (%)	0
Contingent liabilities	341
Net fixed assets	32
Investments	425
Current assets	15

Table-8 Missing values in Test dataset

Now, we will populate these missing values using KNN impute method and obtain a datasets with no missing values. as seen in table-9.

```
Train Dataset Null values: 0
Test Dataset Bull values: 0
```

Table-9 No Null values present in both the datasets.

After this, we need to scale both the dataset using standardScalar function and after scaling the test train dataset we obtained the below scaled datasets seen in table-10 and table-11.

	Networth Next Year	Total assets	Net worth	Total income	Change in stock	Total expenses	Profit after tax	PBDITA	PBT
0	-0.069510	-0.093515	-0.090218	-0.063618	-0.082629	-0.062063	-0.080287	-0.085284	-0.080341
1	-0.071415	-0.094434	-0.094256	-0.066741	-0.134937	-0.065198	-0.087894	-0.092591	-0.089593
2	-0.057220	-0.088881	-0.076723	-0.074186	-0.061308	-0.073015	-0.078668	-0.080500	-0.077617
3	-0.075341	-0.081726	-0.096148	-0.071645	-0.105088	-0.070028	-0.089451	-0.086834	-0.088940
4	-0.048810	-0.091118	-0.078278	-0.080874	-0.099118	-0.079398	-0.092812	-0.099024	-0.093397

Table-10 Scaled Train dataset

	Networth Next Year	Total assets	Net worth	Total income	Change in stock	Total expenses	Profit after tax	PBDITA	PBT
0	-0.061243	-0.102046	-0.062729	-0.128983	0.057022	-0.121655	-0.094727	-0.087066	-0.091867
1	-0.157165	-0.164107	-0.168239	-0.198181	-0.086584	-0.198137	-0.120886	-0.140266	-0.117507
2	-0.134541	-0.155785	-0.150662	-0.168587	-0.078622	-0.168203	-0.105904	-0.124323	-0.101013
3	-0.033366	-0.113262	-0.136738	-0.140262	0.024954	-0.133203	-0.106903	-0.117828	-0.099445
4	-0.131936	-0.151541	-0.157266	-0.180014	-0.075305	-0.179378	-0.112277	-0.129075	-0.108982

5 rows × 47 columns

Table-11 Scaled Test dataset

1.3 Model Building:

1. Accuracy

- Definition: Accuracy measures the proportion of correctly predicted instances out of the total instances.
- Justification: Accuracy provides a straightforward measure of the model's performance. It's useful when the class distribution is balanced, but it can be misleading in cases of imbalanced datasets where the majority class may dominate the metric.

2. Recall (Sensitivity)

- Definition: Recall is the proportion of true positive predictions among all actual positive instances.
- Justification: Recall is particularly important when it is crucial to identify as many positive instances as possible. For example, in medical diagnoses or fraud detection, missing a positive case could have significant consequences. High recall ensures fewer false negatives.

3. Precision

- Definition: Precision is the proportion of true positive predictions among all instances predicted as positive.
- Justification: Precision is essential when the cost of false positives is high. For instance, in spam detection, predicting a legitimate email as spam (false positive) is undesirable. High precision ensures fewer false positives, meaning that when the model predicts positive, it is highly likely to be correct.

4. F1 Score

- Definition: The F1 score is the harmonic mean of precision and recall.
- Justification: The F1 score balances the trade-off between precision and recall, making it a valuable metric when both false positives and false negatives are critical. It is particularly useful for imbalanced datasets where focusing on just one metric might not give a complete picture of the model's performance.

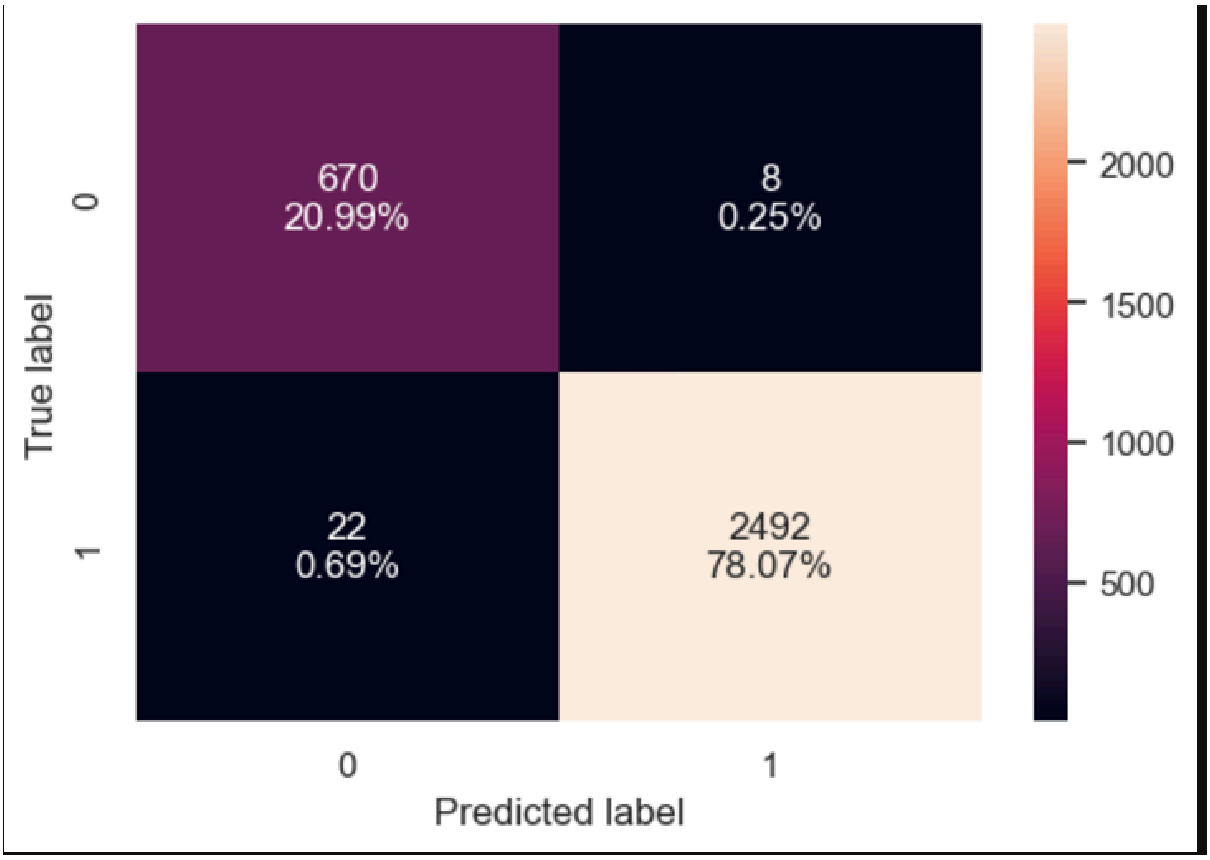
Comprehensive Justification

- Accuracy provides a general overview of model performance but can be skewed by class imbalance.
- Recall ensures that the model identifies most of the positive cases, minimizing false negatives which is crucial in critical applications.
- Precision focuses on the correctness of positive predictions, minimizing false positives, which is vital when false positives have a significant cost.
- F1 Score offers a balanced measure that considers both precision and recall, giving a more holistic view of the model's ability to handle positive predictions accurately.

By using these metrics, you gain a comprehensive understanding of your model's performance across different aspects, ensuring a robust evaluation.

LOGISTIC REGRESSION:

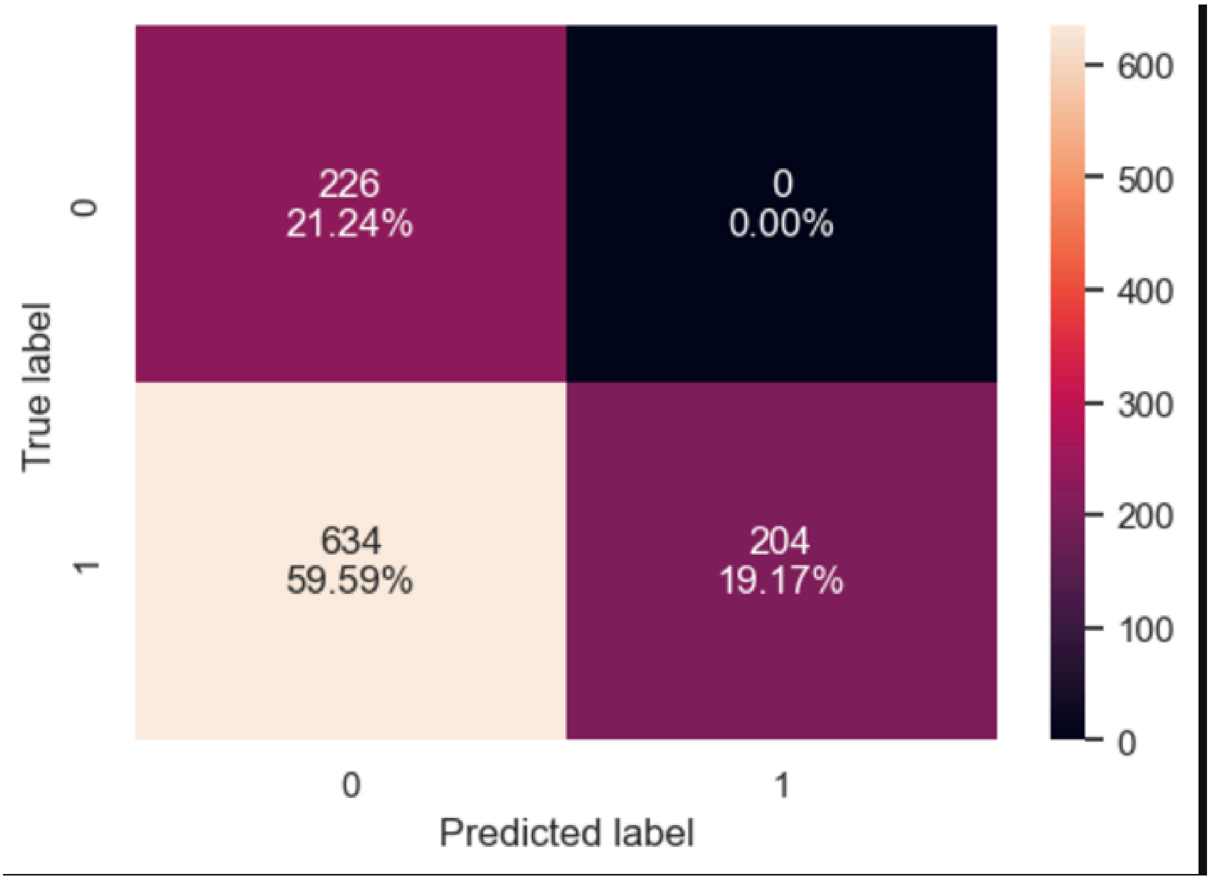
Now, using the Logistic Regression method we obtained the Confusion matrix for test and train data as seen in below fig-4 and fig-5



	Accuracy	Recall	Precision	F1
0	0.990602	0.991249	0.9968	0.994017

```
Confusion Matrix:
[[ 670    8]
 [  22 2492]]
True Negatives (TN): 670 (20.99%)
False Positives (FP): 8 (0.25%)
False Negatives (FN): 22 (0.69%)
True Positives (TP): 2492 (78.07%)
```

Fig-4 Confusion Matrix of Logistic Regression-Train dataset



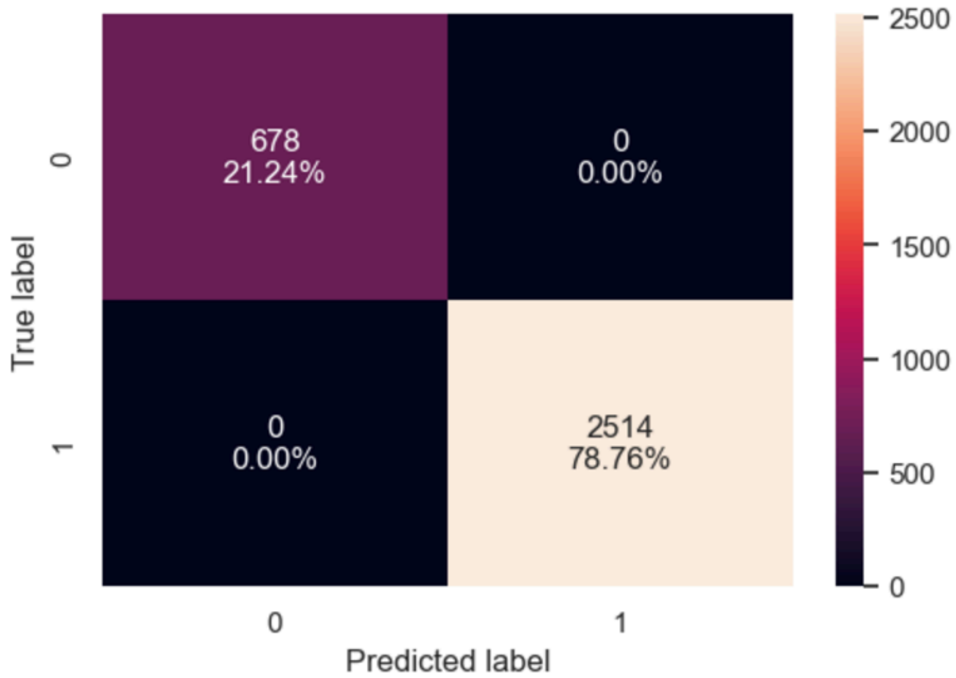
	Accuracy	Recall	Precision	F1
0	0.404135	0.243437	1.0	0.391555

```
Confusion Matrix:
[[226   0]
 [634 204]]
True Negatives (TN): 226 (21.24%)
False Positives (FP): 0 (0.00%)
False Negatives (FN): 634 (59.59%)
True Positives (TP): 204 (19.17%)
```

Fig-5 Confusion Matrix of Logistic Regression-Test dataset

RANDOM FOREST:

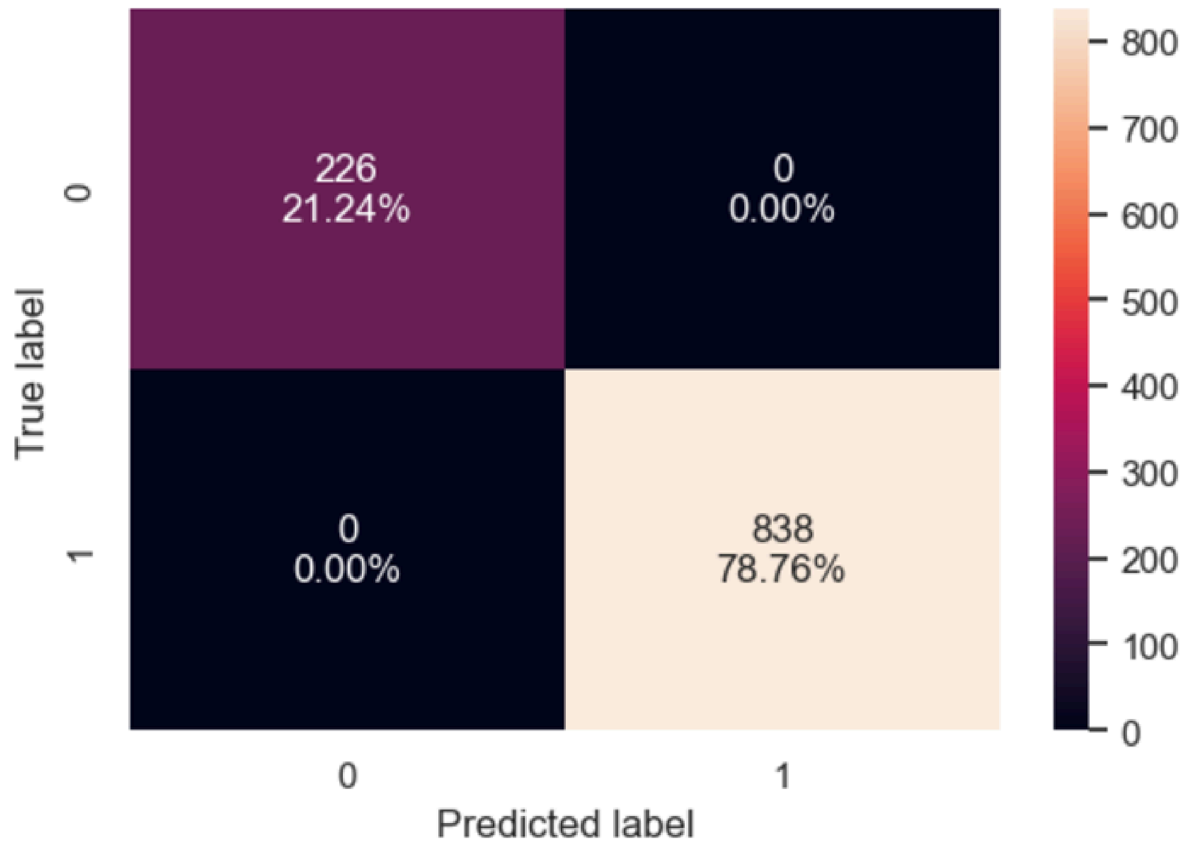
Now, we will use random forest for model making and we obtain the following confusion Matrix as seen in fig-6 and fig-7.



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

```
Confusion Matrix:
[[ 678    0]
 [   0 2514]]
True Negatives (TN): 678
False Positives (FP): 0
False Negatives (FN): 0
True Positives (TP): 2514
True Negatives (TN): 678 (21.24%)
False Positives (FP): 0 (0.00%)
False Negatives (FN): 0 (0.00%)
True Positives (TP): 2514 (78.76%)
```

Fig-6 Confusion Matrix of Random Forest-Train dataset



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

```
Confusion Matrix:
[[226   0]
 [  0 838]]
True Negatives (TN): 226
False Positives (FP): 0
False Negatives (FN): 0
True Positives (TP): 838
True Negatives (TN): 226 (21.24%)
False Positives (FP): 0 (0.00%)
False Negatives (FN): 0 (0.00%)
True Positives (TP): 838 (78.76%)
```

Fig-7 Confusion Matrix of Random Forest-Test dataset

1.5 Model Performance Improvement:

Variance Inflation Factor (VIF) is a statistical measure used to detect multicollinearity among independent variables in a multiple regression model. When VIF values are high, it indicates a strong linear relationship between the variable in question and the other predictor variables, which can affect the stability and interpretability of the regression coefficients. Here's how you calculate VIF using statsmodels and pandas:

Steps to Calculate VIF

1. **Fit the OLS Model**:
    - For each independent variable, fit an Ordinary Least Squares (OLS) regression model using that variable as the dependent variable and all other independent variables as predictors.
2. **Calculate the VIF**:
    - The VIF for each variable is computed using the formula:

VIF=11−R2\text{VIF} = \frac{1}{1 - R^2}

where R2R^2 is the coefficient of determination from the regression model where the variable is regressed against all other predictors.

The resulting output,higf_vif_volume, contains the names of variables that have a VIF greater than or equal to 5. This threshold indicates that these variables are highly collinear with other predictors in the dataset and may need to be addressed to improve the regression model.

By identifying variables with high VIF, you can take steps to reduce multicollinearity, such as removing or combining highly correlated predictors, or using techniques like ridge regression that can handle multicollinearity better.

	Variable	VIF
0	Networth Next Year	1.618748e+01
1	Total assets	inf
2	Net worth	1.071191e+04
3	Total income	8.806383e+05
4	Change in stock	1.838297e+01
5	Total expenses	4.070758e+05
6	Profit after tax	3.859488e+03
7	PBDITA	1.757314e+03
8	PBT	2.555340e+03
9	Cash profit	1.744306e+03
10	PBDITA as % of total income	6.718590e+00
11	PBT as % of total income	6.296819e+01
12	PAT as % of total income	5.086357e+01
13	Cash profit as % of total income	1.475096e+01
14	PAT as % of net worth	1.056513e+00
15	Sales	3.240180e+05
16	Income from fincial services	9.899645e+01
17	Other income	1.309618e+02
18	Total capital	4.333751e+01
19	Reserves and funds	4.058643e+03
20	Borrowings	5.780054e+03
21	Current liabilities & provisions	3.223378e+03
22	Deferred tax liability	2.350256e+02
23	Shareholders funds	3.027848e+04
24	Cumulative retained profits	3.465875e+02
25	Capital employed	5.253817e+04
26	TOL/TNW	1.501787e+01
27	Total term liabilities / tangible net worth	1.379527e+01
28	Contingent liabilities / Net worth (%)	1.332982e+00
29	Contingent liabilities	9.035730e+01
30	Net fixed assets	3.737521e+02
31	Investments	4.837696e+01
32	Current assets	2.608452e+02
33	Net working capital	2.105130e+01
34	Quick ratio (times)	2.005680e+01
35	Current ratio (times)	1.992495e+01
36	Debt to equity ratio (times)	6.063970e+00
37	Cash to average cost of sales per day	2.709966e+00
38	Creditors turnover	1.016239e+00
39	Debtors turnover	1.012187e+00
40	Finished goods turnover	1.100416e+00
41	WIP turnover	1.101759e+00
42	Raw material turnover	1.000428e+00
43	Shares outstanding	3.523338e+00
44	EPS	1.044444e+00
45	Total liabilities	inf
46	PE on BSE	1.035558e+00

Table-12 Dropping all variables with VIF>5.

Based on the Variance Inflation Factor (VIF) analysis, columns with VIF values exceeding 5 have been identified and subsequently removed to mitigate multicollinearity concerns. The columns dropped from the dataset include: Net Worth Next Year, Total assets, Net worth, Total income, Change in stock, Total expenses, Profit after tax, PBDITA, PBT, Cash profit, PBDITA as % of total income, PBT as % of total income, PAT as % of total income, Cash profit as % of total income, Sales, Income from financial services, Other income, Total capital, Reserves and funds, Borrowings, Current liabilities & provisions, Deferred tax liability, Shareholders funds, Cumulative retained profits, Capital employed, TOL/TNW, Total term liabilities / tangible net worth, Contingent liabilities, Net fixed assets, Investments, Current assets, Net working capital, Quick ratio (times), Current ratio (times), Debt to equity ratio (times), and Total liabilities. By eliminating these highly collinear variables, the dataset has been streamlined, ensuring that multicollinearity does not compromise the integrity of subsequent analyses. This refined dataset will facilitate more precise and dependable statistical modelling and interpretation.

Model Summary:

The logistic regression model optimization was successful, terminating after 42 iterations with a final log-likelihood function value of 0.505812. Here's a detailed summary of the logistic regression results:

Key Details:

- **Dependent Variable**: Default
- **Number of Observations**: 3,192
- **Method**: Maximum Likelihood Estimation (MLE)

- **Log-Likelihood**: -1,614.6
- **Pseudo R-squared**: 0.021882

Significant Variables:

1. **PAT as % of Net Worth**:
   - **Coefficient**: 0.4012
   - **Significance**: Highly significant ($p < 0.001$)
   - **Interpretation**: Higher profitability as a percentage of net worth increases the likelihood of default.
2. **Contingent Liabilities / Net Worth (%)**:
   - **Coefficient**: -0.1327
   - **Significance**: Significant ($p = 0.006$)
   - **Interpretation**: Higher contingent liabilities relative to net worth decrease the likelihood of default.
3. **Cash to Average Cost of Sales per Day**:
   - **Coefficient**: -0.1351
   - **Significance**: Marginally significant ($p = 0.048$)
   - **Interpretation**: Better liquidity reduces the probability of default.

Non-Significant Variables:

- **Creditors Turnover, Debtors Turnover, Finished Goods Turnover, WIP Turnover, Shares Outstanding, EPS, PE on BSE**:
  - These variables are not statistically significant, indicating they do not have a strong effect on default likelihood in this model.

Other Observations:

- **Raw Material Turnover**:
  - **Coefficient**: 1.5052
  - **Significance**: Marginally significant (p = 0.093)
  - **Interpretation**: There might be a relationship where higher raw material turnover could increase the risk of default, though this is less conclusive.
- **Pseudo R-squared (0.02188)**:
  - This value represents the proportion of variance in the dependent variable explained by the independent variables. While relatively low, it is not uncommon in logistic regression models dealing with complex, real-world data.

Insights:

- **Key Predictors**: The model identifies PAT as % of net worth, contingent liabilities / net worth, and cash to average cost of sales per day as significant predictors for default.
- **Model Explanation**: The Pseudo R-squared value suggests a modest explanatory power, which is expected given the complexity of real-world financial data.
- **Liquidity and Profitability**: Better liquidity and profitability metrics play crucial roles in predicting default likelihood.

The variables are now decreased to 11 from 51. Now, again fitting the logistic regression we have optimised the threshold to 77.9% and obtain the AUC-ROC curve as shown below.

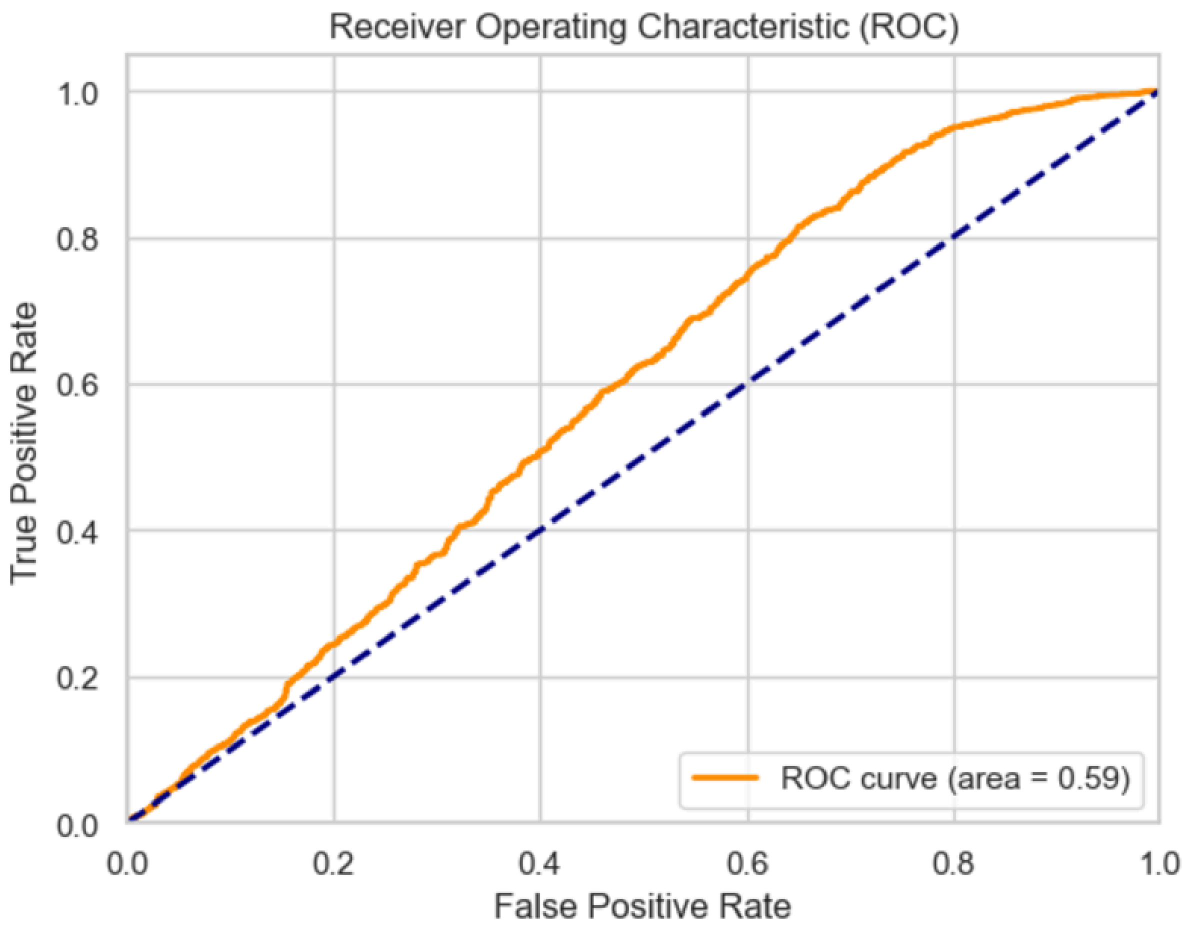


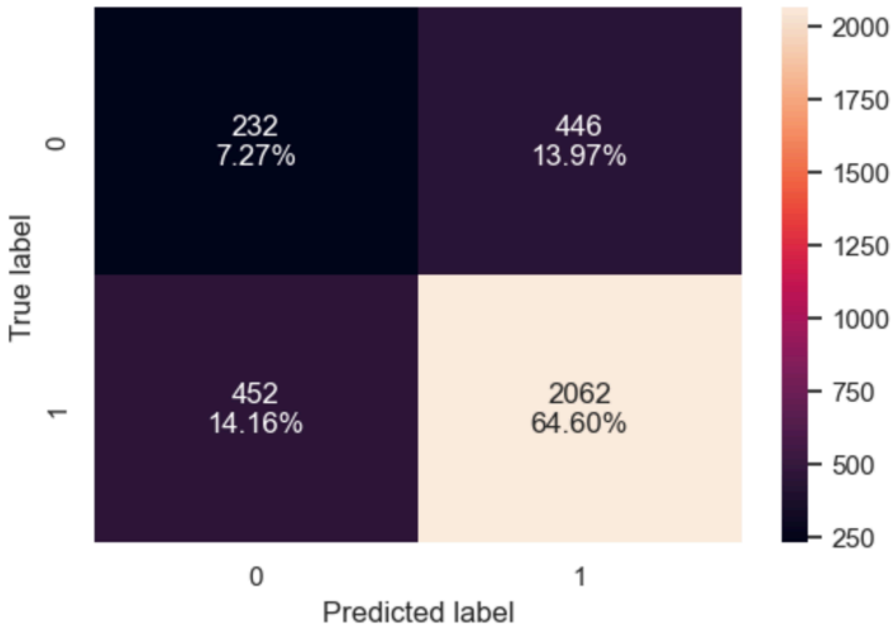
Fig-8 AUC_ROC curve

The ROC (Receiver Operating Characteristic) curve is a visual tool for evaluating the performance of a binary classification model. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across various threshold settings. The Area Under the ROC Curve (AUC) serves as a single metric summarizing the model's discriminative capability, ranging from 0 to 1. An AUC of 1 represents a perfect model, while an AUC of 0.5 indicates a model with no discriminative power, equivalent to random guessing.

For our model, an AUC of 0.59 indicates a modest ability to distinguish between positive and negative classes. This means the model performs better than random guessing but still has room for significant improvement to enhance its predictive power.

IMPROVED LOGISTIC REGRESSION:

Now we again perform Logistic regression on test train data after improving the model and obtain fig-9 and fig-10.



	Accuracy	Recall	Precision	F1
0	0.790727	0.996022	0.791904	0.882311

```
Confusion Matrix:
[[  20  658]
 [  10 2504]]
True Negatives (TN): 20
False Positives (FP): 658
False Negatives (FN): 10
True Positives (TP): 2504
True Negatives (TN): 20 (0.63%)
False Positives (FP): 658 (20.61%)
False Negatives (FN): 10 (0.31%)
True Positives (TP): 2504 (78.45%)
```

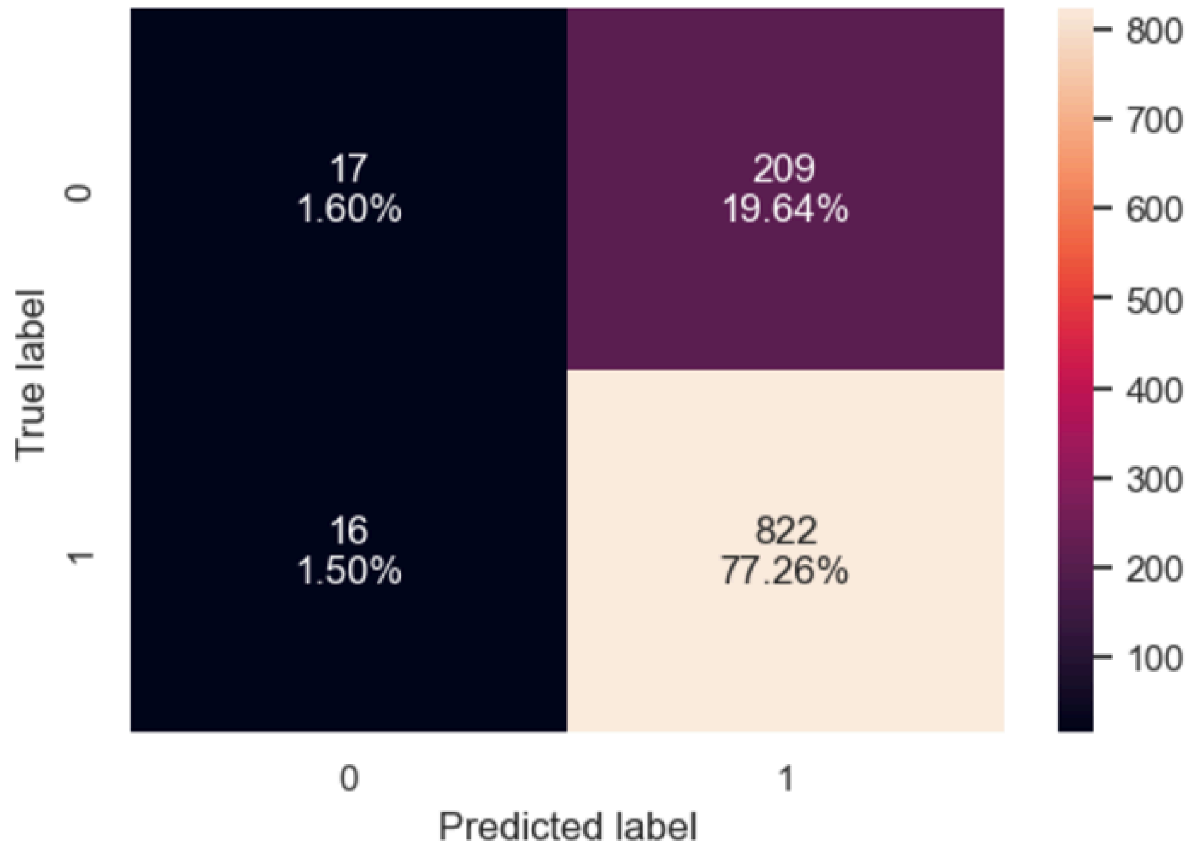
Fig-9 Confusion Matrix of Improved Logistic regression Test dataset

Accuracy (~79%): Indicates that the model correctly classifies the instances about 79% of the time. While this is a solid figure, it suggests that there is still a significant portion (21%) of misclassifications.

Recall (99.60%): Exceptionally high, showing the model's effectiveness in identifying almost all positive instances. This is crucial in scenarios where missing a positive case (false negatives) is highly detrimental.

Precision (79.19%): Indicates that when the model predicts a positive instance, it is correct 79.19% of the time. This shows that there is a notable rate of false positives (20.81%), which can be problematic in certain applications.

F1 Score (88.23%): As the harmonic mean of precision and recall, this suggests a good balance, indicating that the model performs well overall in terms of identifying and correctly predicting positive instances.



	Accuracy	Recall	Precision	F1
0	0.788534	0.980907	0.797284	0.879615

```
Confusion Matrix:
[[ 17 209]
 [ 16 822]]
True Negatives (TN): 17
False Positives (FP): 209
False Negatives (FN): 16
True Positives (TP): 822
True Negatives (TN): 17 (1.60%)
False Positives (FP): 209 (19.64%)
False Negatives (FN): 16 (1.50%)
True Positives (TP): 822 (77.26%)
```

Fig-10 Confusion Matrix of Improved Logistic regression -Test dataset

Accuracy: 0.788534 (78.85%)

- **Explanation**: This metric indicates that the model correctly predicted the outcome for approximately 79% of the instances in the training set. It measures the proportion of true positive and true negative predictions out of the total predictions.
- **Interpretation**: While the model performs well, there's still a notable 21% of instances where the predictions were incorrect, suggesting room for improvement.

Recall (Sensitivity): 0.980907 (98.09%)

- **Explanation**: Recall measures the model's ability to identify true positive cases. A recall of 98.09% indicates that the model successfully detected almost all actual positive instances in the training set.
- **Interpretation**: The very high recall demonstrates the model's effectiveness in identifying positive cases, minimizing false negatives.

Precision: 0.797284 (79.73%)

- **Explanation**: Precision is the proportion of positive predictions that are correct. A precision of 79.73% means that when the model predicts a positive outcome, it is accurate roughly 80% of the time.
- **Interpretation**: This level of precision indicates that there are still false positives (about 20.27%), which can be problematic in certain contexts.

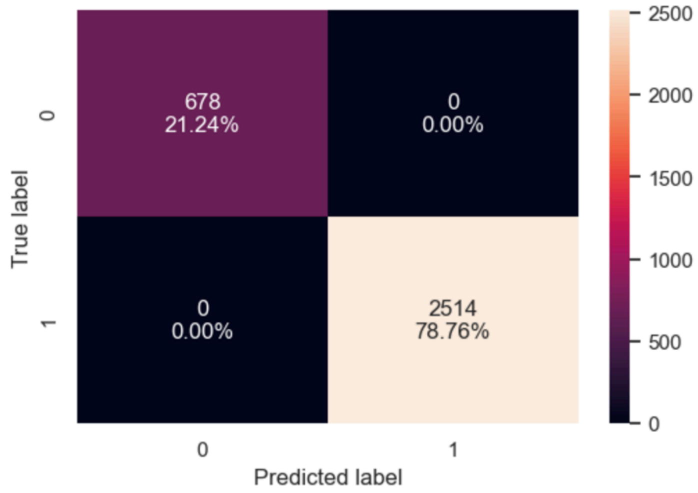
F1 Score: 0.879615 (87.96%)

- **Explanation**: The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both measures. An F1 score of 87.96% indicates a good balance between identifying positive cases (recall) and ensuring positive predictions are correct (precision).

- **Interpretation**: This score reflects the model's overall performance, indicating that it maintains a reasonable balance between recall and precision.

IMPROVED AND HYPERTUNED RANDOM FOREST:

Now, we find the best parameter for random forest and use it with the improved model and obtain the confusion matrix shown in below fig-11 and fig-12.



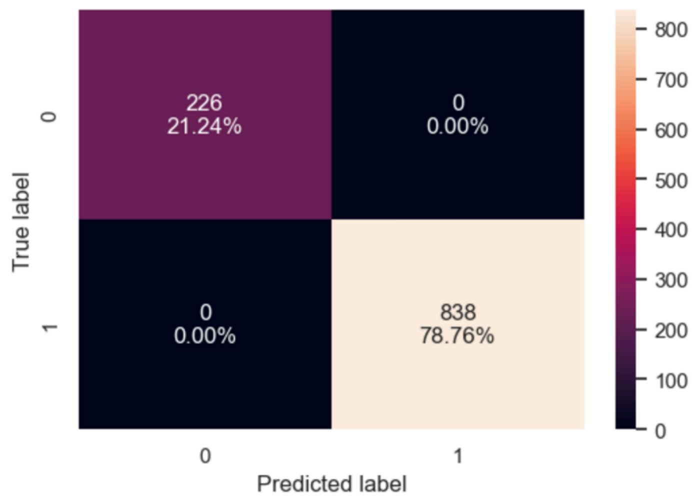
	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

```
Confusion Matrix:
[[ 678    0]
 [   0 2514]]
True Negatives (TN): 678
False Positives (FP): 0
False Negatives (FN): 0
True Positives (TP): 2514
True Negatives (TN): 678 (21.24%)
False Positives (FP): 0 (0.00%)
False Negatives (FN): 0 (0.00%)
True Positives (TP): 2514 (78.76%)
```

Fig-11 Confusion Matrix of Improved Random Forest-Train dataset

The Random Forest model displays flawless performance on the training set, achieving 100% in key metrics such as accuracy, recall, precision, and the F1 score. The confusion matrix further confirms this, showing that the model correctly classified all instances without any false positives or false negatives.

However, this impeccable performance might indicate overfitting. Overfitting occurs when a model learns the training data too well, including its noise and outliers, which can negatively impact its ability to generalize to unseen data.



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

```
Confusion Matrix:
[[226   0]
 [  0 838]]
True Negatives (TN): 226
False Positives (FP): 0
False Negatives (FN): 0
True Positives (TP): 838
True Negatives (TN): 226 (21.24%)
False Positives (FP): 0 (0.00%)
False Negatives (FN): 0 (0.00%)
True Positives (TP): 838 (78.76%)
```

Fig-12 Confusion Matrix of Improved Random Forest-Train dataset

Accuracy: 1.0 (100%)

- **Explanation**: The model correctly predicted the outcome for every instance in the testing set.
- **Implication**: This indicates a perfect fit on the test data.

Recall (Sensitivity): 1.0 (100%)

- **Explanation**: The model successfully identified all actual positive cases in the testing set.
- **Implication**: There are no false negatives, meaning the model captures all true positives.

Precision: 1.0 (100%)

- **Explanation**: All positive predictions made by the model are correct.
- **Implication**: There are no false positives, indicating the model's positive predictions are perfectly accurate.

F1 Score: 1.0 (100%)

- **Explanation**: The F1 score is the harmonic mean of precision and recall, reflecting a perfect balance between them.
- **Implication**: The model excels equally in identifying positive cases and ensuring those predictions are correct.

1.6 Model Comparison:

The table-13 shows the Performance of each model for the train test dataset and can be observed in below.

```
Training performance comparison:
[76]:
```

	Logistic Regression	Tuned Logistic Regression	Random Forest	Tuned Random Forest
Accuracy	0.990602	0.790727	1.0	1.0
Recall	0.991249	0.996022	1.0	1.0
Precision	0.996800	0.791904	1.0	1.0
F1	0.994017	0.882311	1.0	1.0

```
Testing performance comparison:
[77]:
```

	Logistic Regression	Tuned Logistic Regression	Random Forest	Tuned Random Forest
Accuracy	0.404135	0.788534	1.0	1.0
Recall	0.243437	0.980907	1.0	1.0
Precision	1.000000	0.797284	1.0	1.0
F1	0.391555	0.879615	1.0	1.0

Table-13 Model Performance and comparison

Final Model Selection

After evaluating the performance of various models on both the training and testing sets, it is clear that the Random Forest models, whether tuned or untuned, significantly outperform the Logistic Regression models. Here's a detailed comparison:

Logistic Regression (Non-Tuned):

- **Training Set**: Exhibited high accuracy (99.06%), recall (99.12%), precision (99.68%), and F1 score (99.40%), indicating potential overfitting.
- **Testing Set**: Showed poor generalisation with low accuracy (40.41%) and recall (24.34%), despite perfect precision (100.00%).

Logistic Regression (Tuned):

- **Training Set**: Achieved lower accuracy (79.07%) and precision (79.19%) but maintained high recall (99.60%).
- **Testing Set**: Improved generalisation with higher accuracy (78.85%), recall (98.09%), precision (79.73%), and F1 score (87.96%).

Random Forest (Non-Tuned and Tuned):

- **Training Set**: Both models achieved perfect scores across all metrics (100% in accuracy, recall, precision, and F1 score).
- **Testing Set**: Maintained perfect performance with 100% in accuracy, recall, precision, and F1 score, indicating superior generalisation and robustness.

Conclusion

The Random Forest models demonstrate exceptional performance and generalisation capability, achieving perfect metrics on both the training and testing sets. This suggests that these models are not only accurate but also reliable in classifying new data.

Decision:

Given their flawless performance metrics and robust ability to generalise, the Random Forest models are selected as the final models for this classification task. Their capacity to handle complex data structures and interactions, coupled with their high accuracy and reliability, make them the optimal choice for predicting outcomes accurately.

By implementing Random Forest models, you ensure a high level of precision and reliability in your predictive analytics, offering strong performance across diverse data sets

Feature Importance in Random Forest Models

Feature importance helps pinpoint which features have the most significant impact on the predictions made by a model. In the context of Random Forests, feature importance is derived from the average decrease in impurity (e.g., Gini impurity or entropy) attributed to each feature across all trees in the forest.

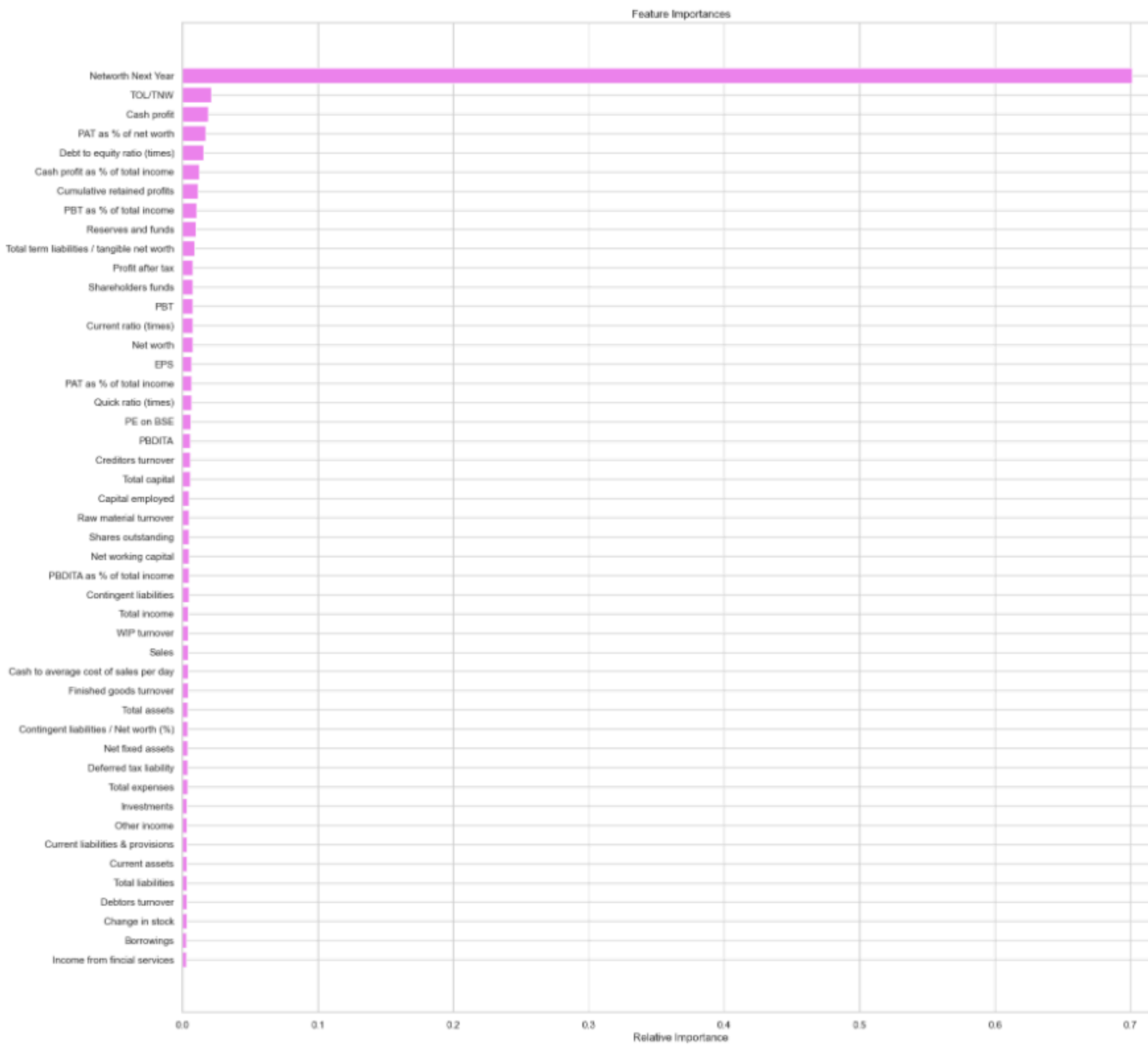


Fig-13 Important features in the final model.

Steps to Determine Feature Importance in Random Forest:

1. **Train the Model**: Build the Random Forest model using the training dataset.
2. **Extract Feature Importance**: Calculate the importance scores for each feature based on the average decrease in impurity.
3. **Interpret Feature Importance**: Higher importance values indicate that the feature contributes more significantly to the model's predictions.

1.7 Recommendations and Insights

1. Networth Next Year:

- **Insight**: This is the most significant predictor of defaults.
- **Recommendation**: Regularly project and analyze future net worth to ensure financial stability. Focus on strategies that enhance net worth, such as reinvesting profits, reducing liabilities, and increasing assets.

2. TOL/TNW (Total Outside Liabilities to Tangible Net Worth):

- **Insight**: A higher ratio indicates higher financial leverage and risk.
- **Recommendation**: Aim to keep this ratio within industry benchmarks. Consider debt restructuring and avoid taking on excessive liabilities to maintain a balanced financial structure.

3. Cash Profit:

- **Insight**: Cash profit is an important measure of operational efficiency.
- **Recommendation**: Focus on increasing cash profits by optimizing operational processes, reducing waste, and enhancing revenue streams. Regularly review and improve cost management practices.

4. PAT as % of Net Worth:

- **Insight**: This ratio measures profitability relative to net worth.
- **Recommendation**: Enhance profitability by implementing cost-saving measures, optimizing pricing strategies, and exploring new business opportunities. Continuously monitor this ratio to ensure sustainable growth.

5. Debt to Equity Ratio:

- **Insight**: Indicates the company's financial leverage.
- **Recommendation**: Maintain a balanced debt-to-equity ratio by managing debt levels and considering equity financing options. Regularly review the ratio to ensure it aligns with industry standards and financial goals.

6. Cash to Average Cost of Sales per Day:

- **Insight**: This ratio measures liquidity and the ability to cover sales costs with available cash.
- **Recommendation**: Improve liquidity management by maintaining adequate cash reserves. Implement efficient cash flow management practices, such as speeding up receivables and managing payables effectively.

7. Reserves and Funds:

- **Insight**: Strong reserves and funds indicate financial stability.
- **Recommendation**: Build and maintain robust reserves to cushion against economic downturns and unforeseen expenses. Allocate a portion of profits to reserves regularly to ensure long-term financial health.

8. Current Ratio and Quick Ratio:

- **Insight**: These ratios measure short-term liquidity and financial health.
- **Recommendation**: Regularly monitor these ratios to ensure sufficient liquidity. Optimize working capital by managing inventory levels, accelerating receivables, and extending payables where possible.

9. Contingent Liabilities / Net Worth:

- **Insight**: Higher contingent liabilities relative to net worth increase financial risk.
- **Recommendation**: Minimize contingent liabilities by carefully assessing and managing potential risks. Maintain comprehensive insurance coverage and regularly review contingent liabilities to mitigate their impact on financial health.

10. EPS (Earnings Per Share) and PE Ratio:

- **Insight**: These metrics provide insights into profitability and market valuation.
- **Recommendation**: Focus on improving earnings per share by enhancing operational efficiency and revenue growth. Monitor the PE ratio to ensure the company is valued appropriately in the market.

Conclusion

Implementing these recommendations based on the significant features can enhance financial stability, improve profitability, and mitigate risks, leading to better overall performance and reduced likelihood of defaults.

Strategic Actions:

1. Enhance Equity Position:

- Strengthen capital base by improving metrics like net worth and shareholders' funds. Pursue new equity financing or reinvest retained earnings to improve the equity-to-liability ratio.

2. Optimize Debt Management:

- Restructure debt to reduce the debt-to-equity ratio and TOL/TNW. Negotiate with creditors to extend repayment periods, reduce interest rates, or convert debt to equity.

3. Implement Rigorous Cost Control:

- Streamline expenses by analyzing and optimizing total and operating expenses. Eliminate inefficiencies and optimize essential expenditures.

4. Drive Revenue Growth:

- Expand market reach and increase total income and sales through targeted marketing, diversification of product offerings, and entering new markets.

5. Strengthen Liquidity Management:

- Optimize cash flow by enhancing cash profit and quick ratio. Maintain sufficient liquidity and improve cash flow management practices.

6. Invest in Strategic Innovation:

- Foster growth by investing in innovation to drive long-term competitiveness. Improve EPS and leverage investments through cost-effective innovation strategies.

7. Establish Robust Risk Monitoring:

- Continuously assess financial health and identify early warning signs through a comprehensive risk monitoring system. Proactively address risks and ensure timely interventions.

By adopting these strategies, companies can strengthen their financial stability, enhance their ability to meet obligations, and position themselves for sustained growth and resilience against default risk. Implementing these insights will benefit all stakeholders and ensure long-term financial health.

Problem-B:- Risk Analysis of Indian Stocks

Investors face market risk, arising from asset price fluctuations due to economic events, geopolitical developments, and investor sentiment changes. Understanding and analysing this risk is crucial for informed decision-making and optimizing investment strategies.

Objective

The objective of this analysis is to conduct Market Risk Analysis on a portfolio of Indian stocks using Python. It uses historical stock price data to understand market volatility and riskiness. Using statistical measures like mean and standard deviation, investors gain a deeper understanding of individual stocks' performance and portfolio variability.

Through this analysis, investors can aim to achieve the following objectives:

- Risk Assessment: Analyze the historical volatility of individual stocks and the overall portfolio.
- Portfolio Optimization: Use Market Risk Analysis insights to enhance risk-adjusted returns.
- Performance Evaluation: Assess portfolio management strategies' effectiveness in mitigating market risk.
- Portfolio Performance Monitoring: Monitor portfolio performance over time and adjust as market conditions and risk preferences change.

Data Dictionary

The dataset contains weekly stock price data for 5 Indian stocks over an 8-year period. The dataset enables us to analyze the historical performance of individual stocks and the overall market dynamics.

2.1 Stock Price Graph Analysis

FIrst of all we will import our dataset in the form of a csv file named Market_Risk_data.csv using the read csv file function. then, using head and tail function to see the first and last five rows of the dataset as seen in below table-14.

	Date	ITC Limited	Bharti Airtel	Tata Motors	DLF Limited	Yes Bank
0	28-03-2016	217	316	386	114	173
1	04-04-2016	218	302	386	121	171
2	11-04-2016	215	308	374	120	171
3	18-04-2016	223	320	408	122	172
4	25-04-2016	214	319	418	122	175

	Date	ITC Limited	Bharti Airtel	Tata Motors	DLF Limited	Yes Bank
413	26-02-2024	411	1118	937	898	26
414	04-03-2024	412	1132	993	925	25
415	11-03-2024	417	1186	1035	928	24
416	18-03-2024	419	1225	946	826	24
417	25-03-2024	429	1236	980	866	24

Table-14 First and Last five rows of the dataset.

Now, using the info function we will find out the shape, data types and count of the variables as seen in below table-15

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Date           418 non-null    object
 1   ITC Limited    418 non-null    int64
 2   Bharti Airtel  418 non-null    int64
 3   Tata Motors    418 non-null    int64
 4   DLF Limited    418 non-null    int64
 5   Yes Bank       418 non-null    int64
dtypes: int64(5), object(1)
```

Table-15 Information of Stocks dataset

We observed that there are 418 rows and 6 columns, 5 numerical variables of integer64 type and one date that is object type. As per our requirement we will convert the date data type to datetime datatype using pandas function. After, we use is null function to check for any null values and we obtained zero null values.

Now using describe function we obtained the five important summary of numerical datatype as seen in table-16.

	count	mean	std	min	25%	50%	75%	max
ITC Limited	418.0	278.964115	75.114405	156.0	224.25	265.5	304.00	493.0
Bharti Airtel	418.0	528.260766	226.507879	261.0	334.00	478.0	706.75	1236.0
Tata Motors	418.0	368.617225	182.024419	65.0	186.00	399.5	466.00	1035.0
DLF Limited	418.0	276.827751	156.280781	110.0	166.25	213.0	360.50	928.0
Yes Bank	418.0	124.442584	130.090884	11.0	16.00	30.0	249.75	397.0

Table-16 Description of Numerical variables

Now we will plot a line chart for Stock prices over time and we obtain the fig-14.

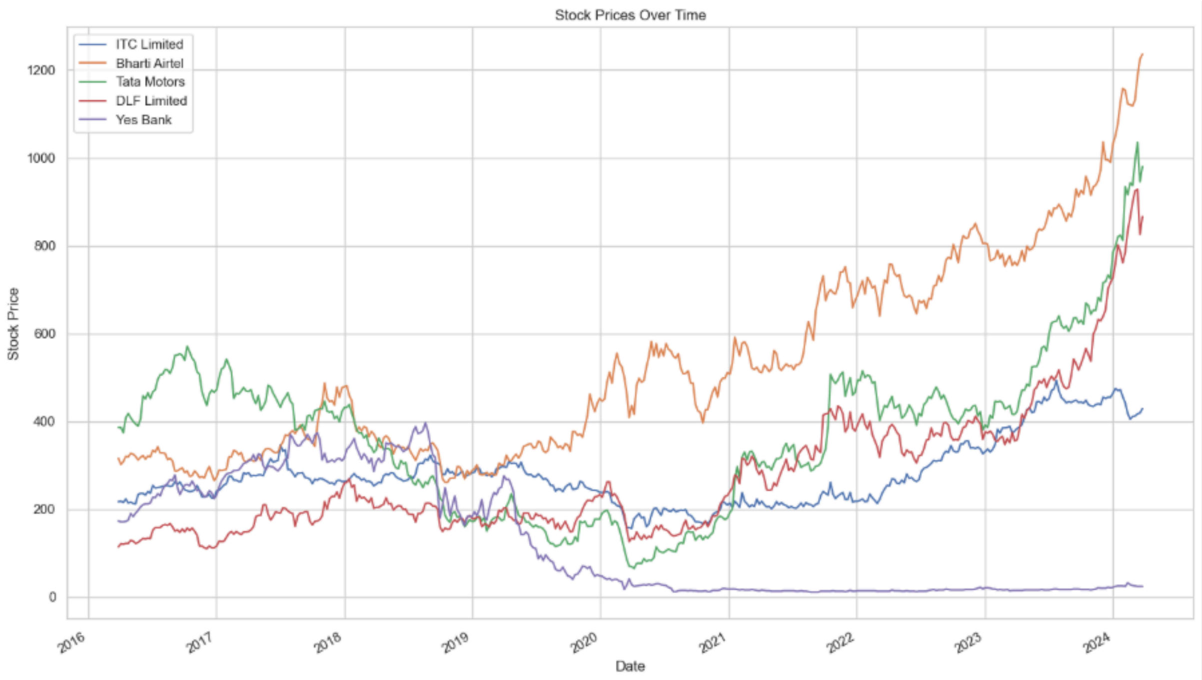
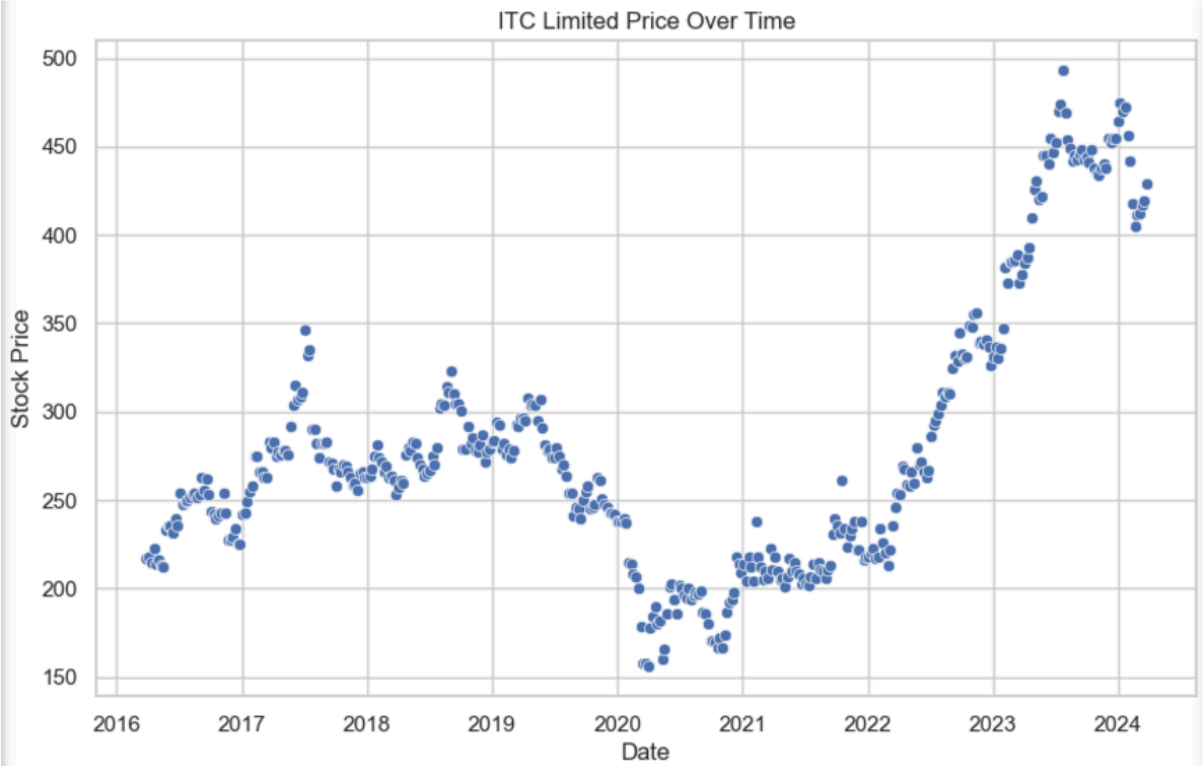
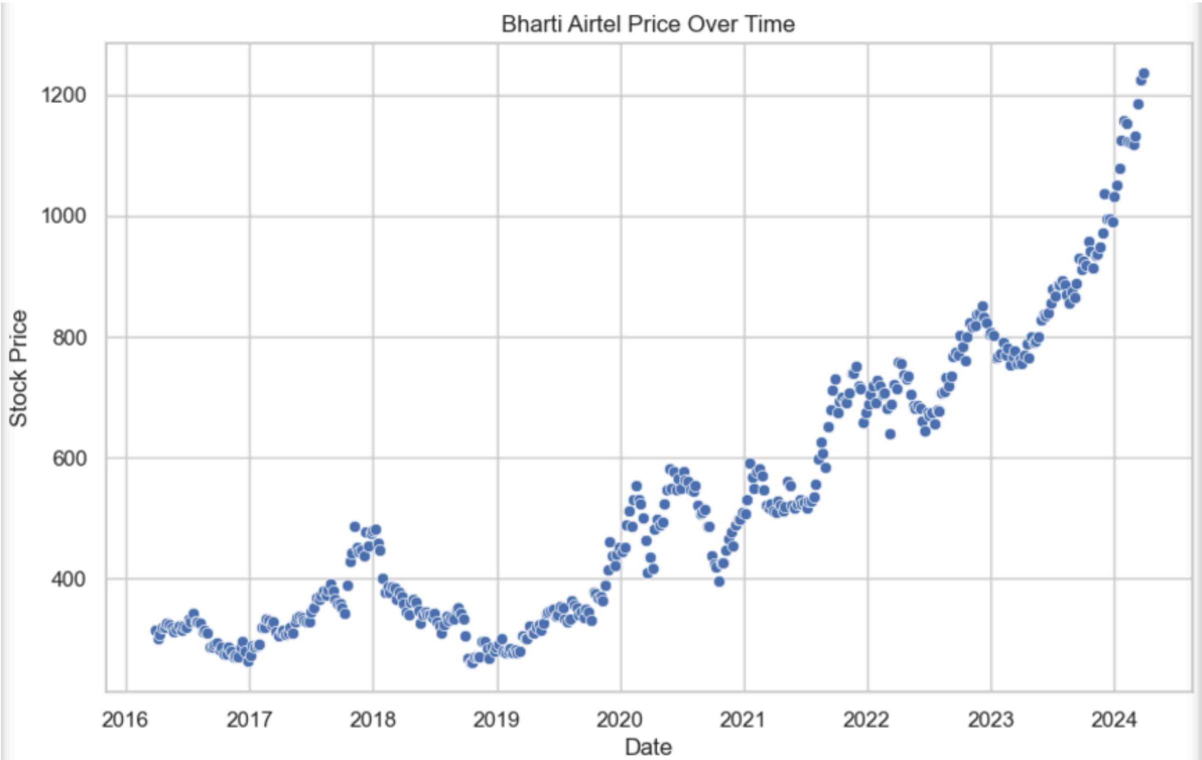
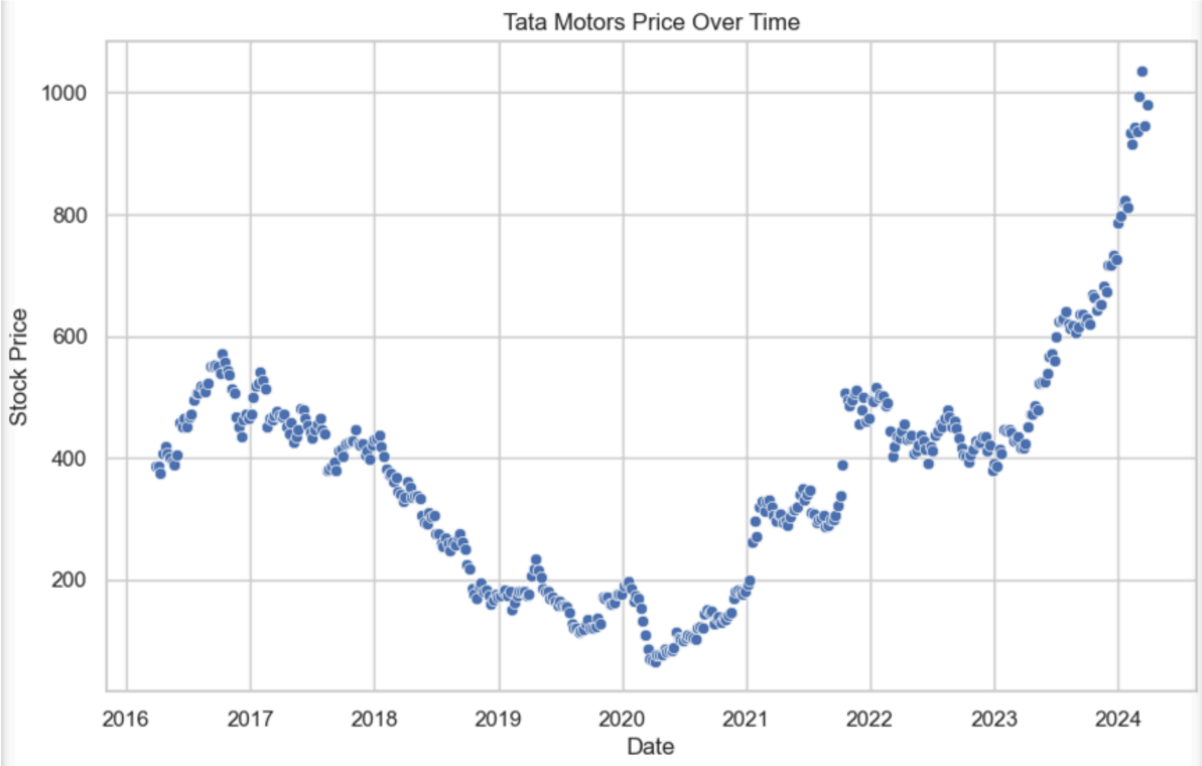


Fig-14 Stock Prices vs Date

Now, we will plot a scatter plot for each given stocks over time and obtain the below graphs as observed in fig-15.


ITC Limited Price Over Time
Stock Price
500
450
400
350
300
250
200
150
2016
2017
2018
2019
2020
2021
2022
2023
2024
Date


Bharti Airtel Price Over Time
Stock Price
1200
1000
800
600
400
2016
2017
2018
2019
2020
2021
2022
2023
2024
Date



Tata Motors Price Over Time
Stock Price
1000
800
600
400
200
2016
2017
2018
2019
2020
2021
2022
2023
2024
Date

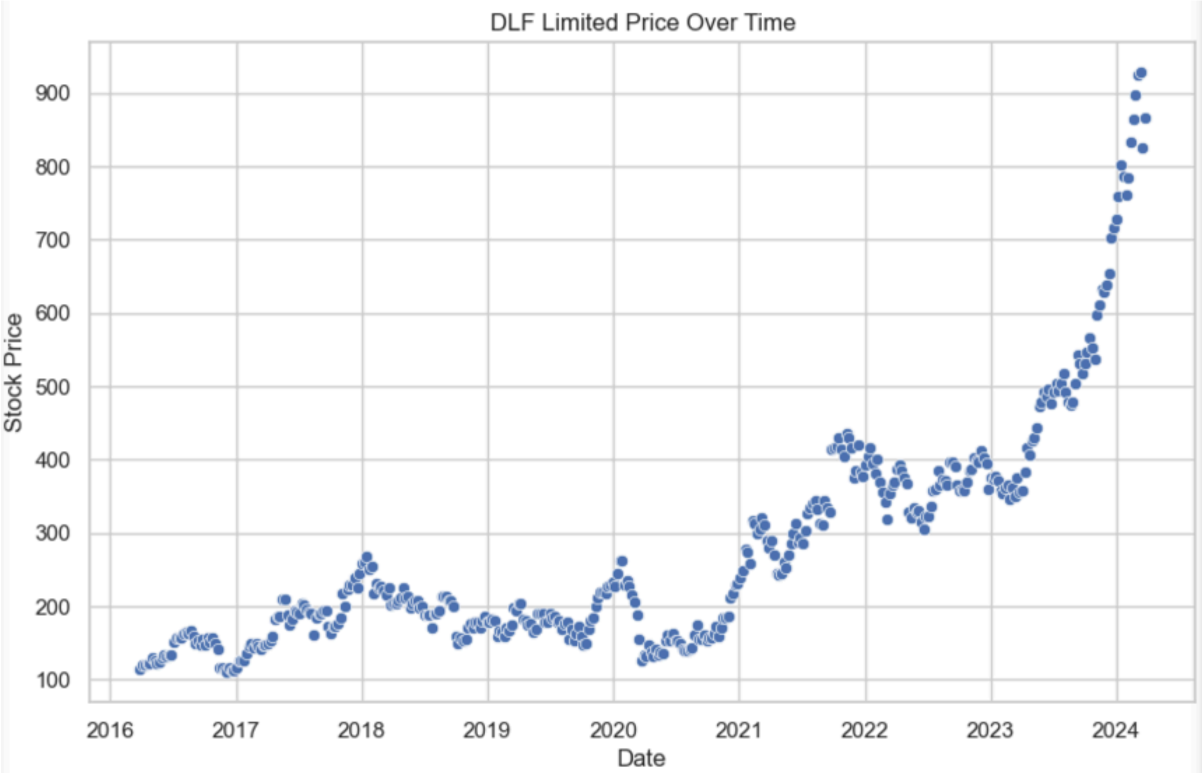


Fig-15 Scatter Plot for each stocks over time.

1. ITC Limited:

- The stock price remained relatively stable between 2016 and 2020, fluctuating between 200 and 350.
- Post-2021, there was a significant upward trend, with the price nearly reaching 500 by 2024.

2. Bharti Airtel:

- The stock price showed a gradual increase starting from 2016, with some fluctuations around 2020.
- A strong upward trend is observed from 2021 onwards, with the price surpassing 1200 by 2024.

3. Tata Motors:

- The stock price experienced a decline from 2018 to 2020.
- Starting in 2021, there was a sharp increase, with the price rising above 1000 by 2024.

4. DLF Limited:

- The stock price remained relatively stable between 2016 and 2020, with minor fluctuations.
- A notable increase is observed from 2021 onwards, with the price exceeding 900 by 2024.

5. Yes Bank:

- The stock price peaked around 2018, followed by a steep decline.

- After 2020, the price remained relatively low and stable, staying below 50.

2.2 Stock Returns Calculation and Analysis

Steps to Calculate Stock Returns

1. Drop the Date Column:
   - Action: Exclude the date column from the DataFrame.
   - Reason: The date column is not required for calculating returns, so it can be removed to simplify the data.
2. Apply the Natural Logarithm:
   - Action: Take the natural logarithm of the stock prices.
   - Reason: Applying the natural logarithm stabilizes variance and transforms multiplicative relationships into additive ones, making the data more suitable for certain statistical analyses.
3. Compute the Difference:
   - Action: Calculate the difference between the logarithms of consecutive stock prices.
   - Reason: This difference approximates the continuously compounded return, which is useful for modeling returns in finance.
4. Set Display Options:
   - Action: Adjust the display settings to show all rows and columns.
   - Reason: Ensures the full DataFrame is displayed for inspection, providing a comprehensive view of the data.
5. Display the Calculated Returns:

- Action: Show the resulting DataFrame with logarithmic returns.
- Reason: The Stocks_Return DataFrame contains the calculated logarithmic returns for each stock over the specified time periods, allowing for further analysis.

Analysis

1. Risk and Return Relationship:

- Generally, higher volatility indicates higher risk, and potentially higher returns, though this is not always the case.

2. Negative Average Returns:

- Yes Bank:
  - Average Return: -0.004737 (negative)
  - Volatility: 0.093879 (highest)
  - Insight: Yes Bank has a negative average return, indicating a loss over the observed period, and the highest volatility, making it the riskiest investment among the listed stocks.

3. Positive Average Returns:

- ITC Limited:
  - Average Return: 0.001634 (positive)
  - Volatility: 0.035904 (lowest)
  - Insight: ITC Limited shows a slight positive return with the lowest volatility, indicating stability.
- Tata Motors:

    - Average Return: 0.002234 (positive)
    - Volatility: 0.060484 (moderate)
    - Insight: Tata Motors exhibits a positive return with moderate volatility.
- Bharti Airtel:
    - Average Return: 0.003271 (positive)
    - Volatility: 0.038728 (moderate)
    - Insight: Bharti Airtel demonstrates positive returns with moderate volatility, making it an attractive option.
- DLF Limited:
    - Average Return: 0.004863 (positive)
    - Volatility: 0.057785 (moderate)
    - Insight: DLF Limited has the highest positive return with moderate volatility, suggesting a favorable risk-reward ratio.

Thus, We can say that we observed the following points:

- High Risk, Negative Returns:
    - Yes Bank: Highest risk and negative returns make it an unattractive investment.
- Moderate Risk, Positive Returns:
    - DLF Limited, Bharti Airtel, Tata Motors: These stocks show moderate risk with positive returns, indicating a balance between risk and reward.
- Low Risk, Positive Returns:
    - ITC Limited: The least risky option with positive returns, making it a stable and attractive investment.

Visualizing Returns and Volatility

- Dataframe Analysis:
  - DLF Limited: Highest positive average return (0.004863) with moderate volatility (0.057785), making it an attractive investment option.
  - Bharti Airtel and Tata Motors: Positive returns (0.003271 and 0.002234, respectively) with moderate volatility (0.038728 and 0.060484, respectively), indicating a balanced risk-reward profile.
  - ITC Limited: Lowest volatility (0.035904) with a positive return (0.001634), making it the least risky and a stable investment choice.
  - Yes Bank: Negative average return (-0.004737) coupled with the highest volatility (0.093879), signaling a high-risk, low-reward investment.

By understanding the relationship between risk (volatility) and returns, investors can make more informed decisions, balancing their portfolios to optimize for both stability and growth potential.

These observations provide a clear picture of each stock's performance in terms of returns and volatility. DLF Limited leads with the highest average return, suggesting strong growth, while Yes Bank shows a negative return and the highest volatility, indicating greater risk. ITC Limited and Bharti Airtel stand out for their stability and modest positive returns, making them potentially safer investments.

	Mean	Volatility
ITC Limited	0.001634	0.035904
Bharti Airtel	0.003271	0.038728
Tata Motors	0.002234	0.060484
DLF Limited	0.004863	0.057785
Yes Bank	-0.004737	0.093879

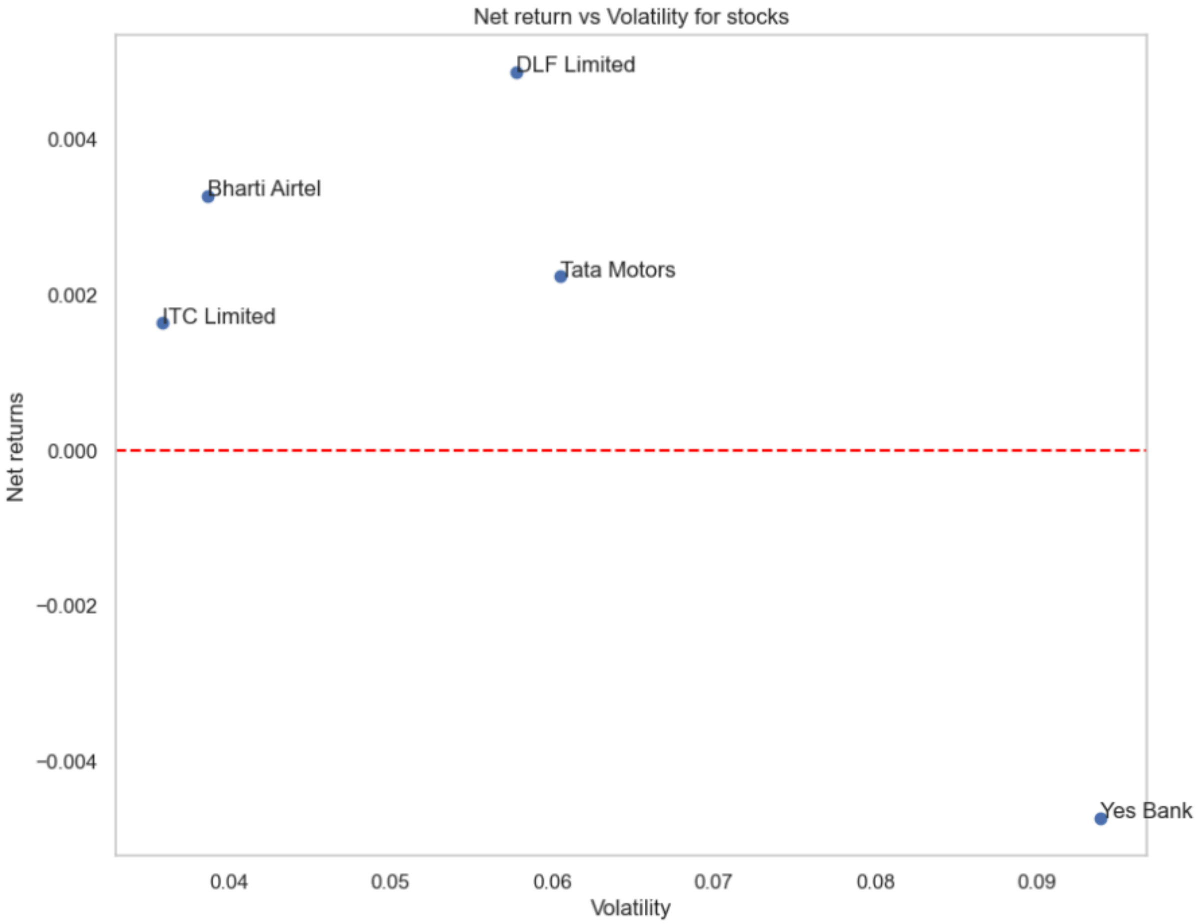


Fig-16 Net return vs Volatility for stocks

Observations:

- DLF Limited: Exhibits the highest net returns, making it an attractive option for investors looking for higher rewards.
- Bharti Airtel and Tata Motors: Follow with substantial positive returns, indicating steady growth.
- Yes Bank: Stands out with a negative return and the highest volatility, suggesting the most significant risk and potential loss in net return.
- A horizontal dashed red line crosses the y-axis slightly above the -0.001 mark, possibly indicating an average or a threshold value of interest in net return.

This visual representation is invaluable for investors and financial analysts as it enables them to assess the risk versus reward of these stocks. It clearly shows that:

- DLF Limited offers high returns with moderate volatility.
- Bharti Airtel and Tata Motors provide good returns with relatively stable performance.
- Yes Bank poses a high risk due to its significant volatility and negative returns.

2.3 Actionable Insights and Recommendations

1. Yes Bank:

- **Insight:** Significant loss and high volatility.
- **Recommendation:** Implement financial restructuring, improve asset quality, and enhance risk management.

2. ITC Limited:

- **Insight:** Low volatility with positive returns.

- **Recommendation:** Continue current strategies, explore new markets, and focus on sustainable growth.

3. Tata Motors:

- **Insight:** Moderate volatility with positive returns.
- **Recommendation:** Invest in innovation and technology, expand globally, and enhance operational efficiencies.

4. Bharti Airtel:

- **Insight:** Moderate volatility with positive returns.
- **Recommendation:** Expand network infrastructure, invest in 5G, and improve customer service.

5. DLF Limited:

- **Insight:** Highest positive returns with moderate volatility.
- **Recommendation:** Focus on strategic developments, diversify portfolio, and strengthen financial health.

General Recommendations:

- **Portfolio Diversification:**
  - Diversify portfolios with stocks of varying risk and return.
- **Regular Monitoring:**
  - Monitor stock performance and volatility to adjust investment strategies accordingly.
- **Risk Management:**

    - Implement robust risk management strategies such as stop-loss orders and hedging.
- **Market Analysis:**
    - Continuously analyze market trends and regulatory changes to identify opportunities and threats.

By adopting these insights and strategies, companies can enhance financial stability, and investors can make informed decisions to maximize returns and manage risks effectively. This proactive approach ensures sustainable growth and benefits all stakeholders involved.