FINANCE AND RISK ANALYTICS

PROBLEM STATEMENT - CODED

By: BENITA MERLIN.E

PGP-Data Science and Business Analytics.

BATCH: PGP DSBA. O. MAY24.A

Contents

PART A 7
1. CONTEXT 7
1.1 Objective 7
1.2 Problem Definition 7
2. DATA DESCRIPTION 8
2.1 Data Dictionary 8
2.2 Sample Dataset 10
2.3 Data Information 11
3. STATISTICAL ANALYSIS 13
4. EXPLORATORY DATA ANALYSIS 15
4.1 UNIVARIATE ANALYSIS 15
4.1.1 Count Plot for the Target Variable 16
4.1.2 Boxplots for all the Numerical Columns 17
4.1.3 Violin Plots for all the Numerical Columns 20
4.1.4 Hisplots for all the Numerical Columns 22
4.1.5 Cumulative Distribution of Financial Variables 24
4.2 BIVARIATE & MULTIVARIATE ANALYSIS 26
4.2.1 Financial Variables and Default Risk 26
4.2.2 Heatmap 28
4.2.3 Important Financial Indicators for Default Prediction 29
4.2.4 Debt Management Practices of Defaulters vs. Non-Defaulters 30
4.2.5 Financial Ratios & Risk Assessment 31
4.2.6 Patterns in Financial Data for Risk Assessment 32
5.DATA PRE-PROCESSING 33
5.1 Feature Engineering 33
5.1.1 Outlier Detection 33
5.1.2 Missing value Treatment 38
5.1.3 Duplicate value Treatment 40
5.1.4 Data Scaling 40
6. MODEL BUILDING 41
Model Evaluation Criterion 41
6.1 Evaluation Metrics for Default Prediction 41
6.2 Model Selection and Justification 42
6.3 Logistic Regression 43
6.3.1 Logistic Regression Model - Training Performance 45

6.3.2 Logistic Regression Model - Test Performance 47

6.4 Random Forest 49

6.4.1 Random Forest Model - Training Performance 49

6.4.2 Random Forest Model - Test Performance 50

7. MODEL PERFORMANCE IMPROVEMENT 51

7.1 Model Performance Improvement – Logistic Regression 51

7.1.1 Variance Inflation Factor (VIF) Analysis 51

7.1.2 Evaluation of Model Performance Using ROC Curve and AUC Score 57

7.1.3 Logistic Regression Performance - Training Set 58

7.1.4 Logistic Regression Performance - Test Set 60

7.2 Model Performance Improvement - Random Forest 61

7.2.1 Random Forest Hyperparameters Visualization 61

7.2.2 Random Forest Performance - Training Set 62

7.2.3 Random Forest Performance - Test Set 64

8.MODEL COMPARISON AND FINAL MODEL SELECTION 65

8.1 Training Phase Performance Comparison 66

8.2 Testing Phase Performance Comparison 67

8.3 Final Model Selection 67

8.4 Feature Importance of Tuned Random Forest 69

9. ACTIONABLE INSIGHTS & RECOMMENDATIONS 70

9.1 Model Performance Insights 70

9.2 Key Financial Indicators for Default Risk 70

9.3 Business Recommendations 70

PART-B 71

1. CONTEXT 71

1.1 Objective 71

1.2 Problem Definition 71

2. DATA DESCRIPTION 72

2.1 Data Dictionary 72

2.2 Sample Dataset 72

2.3 Data Information 73

3. STATISTICAL ANALYSIS 73

4. DATA PREPROCESSING 75

4.1 Feature Engineering 75

5. EXPLORATORY DATA ANALYSIS 76

5.1 UNIVARIATE ANALYSIS 76

5.1.1 Stock Price Distribution 76
5.1.2 Box Plot of Stock Prices 77
5.2 BIVARIATE & MULTIVARIATE ANALYSIS 78
5.2.1 Stock Price Correlation Matrix 78
6. STOCK PRICE ANALYSIS 80
6.1 Line Plot for Trend Visualization 80
6.1.1 Trend of ITC Limited Over Time 80
6.1.2 Trend of Bharti Airtel Over Time 81
6.1.3 Trend of Tata Motors Over Time 82
6.1.4 Trend of DLF Limited Over Time 83
6.1.5 Trend of Yes Bank Over Time 84
7. STOCK RETURNS CALCULATION AND ANALYSIS 86
7.1 Stock Return Calculation 86
7.2 Stock Return Volatility Analysis Over Time 87
7.3 Cumulative Stock Return Volatility Over Time 90
7.4 Stock Return Correlation Analysis 91
7.5 Stock Return Distribution Analysis 92
7.6 Rolling Volatility Analysis of Stocks 93
8. AVERAGE RETURNS 94
9. VOLATILITY 95
10. VISUALIZING RETURNS AND VOLATILITY 96
11. ACTIONABLE INSIGHTS & RECOMMENDATIONS 97

Table of Figures

PART -A

Figure A.1 Sample Dataset 10
Figure A.2 Data Information 11
Figure A. 3 Statistical Summary 14
Figure A. 4 Count Plot for the Target Variable 16
Figure A. 5 Boxplots for all the Numerical Columns 18
Figure A. 6 Violin Plots for all the Numerical Columns 20
Figure A. 7 Hisplots for all the Numerical Columns 22
Figure A. 8 Cumulative Distribution of Financial Variables 24
Figure A. 9 Boxplot for Financial Variables and Default Risk 27
Figure A. 10 Heatmap for numeric columns 28
Figure A. 11 Important Financial Indicators for Default Prediction 29
Figure A. 12 Boxplot of Debt Management Practices of Defaulters vs. Non-Defaulters 30
Figure A. 13 Financial Ratios Distribution for Defaulters & Non-Defaulters 31

Figure A. 14 Pair Plot of Financial Data for Risk Assessment 32
Figure A. 15 Number of Outliers in each column 36
Figure A. 16 Missing values in X-Train Dataset 38
Figure A. 17 Missing values in X-Test Dataset 39
Figure A. 18 X-Train Scaled Dataset 40
Figure A. 19 X-Test Scaled Dataset 41
Figure A. 20 Adding Constant Term 43
Figure A. 21 Logistic Regression Results 44
Figure A. 22 Logistic Regression Model - Training Performance 45
Figure A. 23 Model Performance Metrics for Train Data 46
Figure A. 24 Logistic Regression Model - Test Performance 47
Figure A. 25 Model Performance Metrics for Test Data 47
Figure A. 26 Random Forest Model - Training Performance 49
Figure A. 27 Model Performance Metrics for Train Data 50
Figure A. 28 Random Forest Model - Test Performance 50
Figure A. 29 Model Performance Metrics for Test Data 51
Figure A. 30 Variance Inflation Factor (VIF) Analysis 54
Figure A. 31 VIF after dropping column 55
Figure A. 32 Bar Plot of Logistic Regression Coefficients 56
Figure A. 33 ROC Curve and AUC Score 57
Figure A. 34 Logistic Regression Performance - Training Set 58
Figure A. 35 Performance Metrics - Logistic Regression - Training 59
Figure A. 36 Logistic Regression Model - Test Performance 60
Figure A. 37 Performance Metrics - Logistic Regression - Test 61
Figure A. 38 Random Forest Hyperparameters Visualization 61
Figure A. 39 Random Forest Performance - Training Set 62
Figure A. 40 Performance Metrics - Random Forest - Training 63
Figure A. 41 Random Forest Performance - Test Set 64
Figure A. 42 Performance Metrics - Logistic Regression - Test 65
Figure A. 43 Training Phase Performance Comparison 66
Figure A. 44 Testing Phase Performance Comparison 67
Figure A. 45 Feature Importance 69

PART - B

Figure B 1. Sample Dataset 72
Figure B 2. Data Information 73
Figure B 3. Statistical Summary 74
Figure B 4.Checking Missing Values 75
Figure B 5.Stock Price Distribution 76
Figure B 6.Box Plot of Stock Prices 77
Figure B 7.Stock Price Correlation Matrix 78
Figure B 8.Line Plot for Trend Visualization-ITC Limited 80
Figure B 9.Line Plot for Trend Visualization- Bharati Airtel 81
Figure B 10.Line Plot for Trend Visualization-Tata Motors 82
Figure B 11.Line Plot for Trend Visualization-DLF Limited 83
Figure B 12.Line Plot for Trend Visualization- yes Bank 84
Figure B 13.Stock Return Calculation 86
Figure B 14.Stock Return Volatility Analysis Over Time 87
Figure B 15.Boxplot for Stock Volatility Distribution 89
Figure B 16.Cumulative Stock Return Volatility Over Time 90
Figure B 17.Stock Return Correlation Analysis 91
Figure B 18.Violin Plot for Stock Return Distribution 92
Figure B 19. Rolling Volatility Analysis of Stocks 93
Figure B 20.Average Returns 94
Figure B 21.Volatility 95
Figure B 22. Scatter Plot. for Net Return vs. Volatility 96

PART A

1. CONTEXT

In the realm of modern finance, businesses encounter the perpetual challenge of managing debt obligations effectively to maintain a favorable credit standing and foster sustainable growth. Investors keenly scrutinize 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.

1.1 Objective

A group of venture capitalists want to develop a Financial Health Assessment Tool. With the help of the tool, it endeavors 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 analyze 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 fulfill financial obligations promptly and efficiently, and identify potential cases of default.

2. Credit Risk Evaluation: Evaluate credit risk exposure by analyzing 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 analyze 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 organization anticipate potential challenges with the financial performance of the companies and enable proactive risk mitigation strategies.

1.2 Problem Definition

- Businesses need to manage debt effectively to ensure financial stability and maintain a strong credit profile.

- Investors analyze financial statements to assess a company’s capability to fulfill obligations and remain profitable.

- This project focuses on creating a Financial Health Assessment Tool to predict potential company defaults based on past financial data.

- Target Variable: Net Worth Next Year (if negative, the company is classified as a defaulter).

- Objective: Detect critical financial trends that signal a company's risk of default.

2. DATA DESCRIPTION

2.1 Data Dictionary

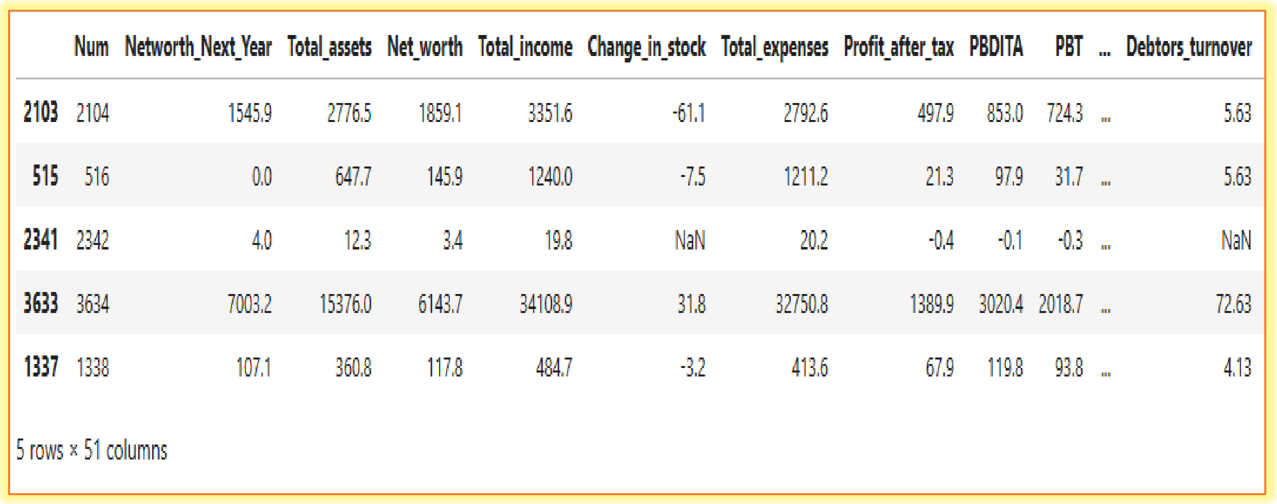
The data consists of financial metrics from the balance sheets of different companies. The detailed data dictionary is given below.

- Networth 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 amortization
- 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

Note: A company will not be tagged as a defaulter if its net worth next year is positive, or else, it'll be tagged as a defaulter.

2.2 Sample Dataset



	Num	Networth_Next_Year	Total_assets	Net_worth	Total_income	Change_in_stock	Total_expenses	Profit_after_tax	PBDITA	PBT	...	Debtors_turnover
2103	2104	1545.9	2776.5	1859.1	3351.6	-61.1	2792.6	497.9	853.0	724.3	...	5.63
515	516	0.0	647.7	145.9	1240.0	-7.5	1211.2	21.3	97.9	31.7	...	5.63
2341	2342	4.0	12.3	3.4	19.8	NaN	20.2	-0.4	-0.1	-0.3	...	NaN
3633	3634	7003.2	15376.0	6143.7	34108.9	31.8	32750.8	1389.9	3020.4	2018.7	...	72.63
1337	1338	107.1	360.8	117.8	484.7	-3.2	413.6	67.9	119.8	93.8	...	4.13

5 rows × 51 columns

Figure A.1 Sample Dataset

This is a sample dataset of financial data for businesses, showing key metrics like net worth, total assets, income, expenses, profitability, and turnover ratios. The dataset has 51 columns and contains missing values in some fields. The values vary significantly, indicating differences in business sizes and financial health.

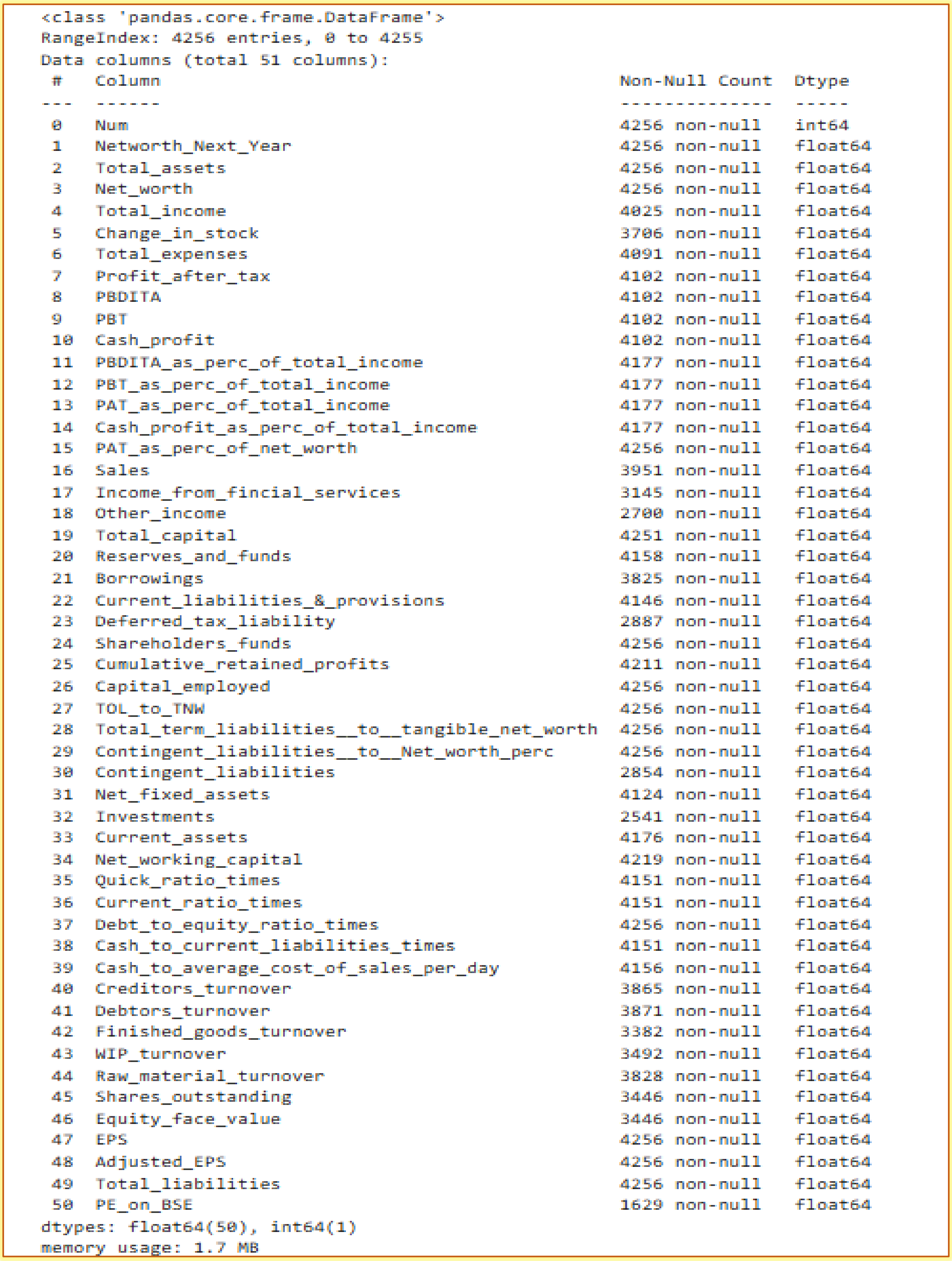
Shape:

The dataset contains 4256 Records and 51 Financial features.

Datatype:

The dataset consists of numerical values, with data types including both float and integer.

2.3 Data Information



```
<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_perc_of_total_income                  4177 non-null   float64
 12  PBT_as_perc_of_total_income                     4177 non-null   float64
 13  PAT_as_perc_of_total_income                     4177 non-null   float64
 14  Cash_profit_as_perc_of_total_income             4177 non-null   float64
 15  PAT_as_perc_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_to_TNW                                      4256 non-null   float64
 28  Total_term_liabilities__to__tangible_net_worth  4256 non-null   float64
 29  Contingent_liabilities__to__Net_worth_perc      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
```

Figure A.2 Data Information

Observation:

1. Dataset Overview

    - Contains 4,256 records and 51 financial features related to business performance, liabilities, profitability, and risk.
    - Missing values exist in key financial indicators like Profit After Tax (4102 non-null), Sales (3951), and PE_on_BSE (1629), requiring imputation or handling.

2. Key Financial Metrics

    - Total Assets, Net Worth, and Borrowings indicate company size and financial stability.
    - Income & Expenses: Variations in Total Income (4256), Total Expenses (4102), and Profitability Ratios suggest different financial strategies and efficiency levels.

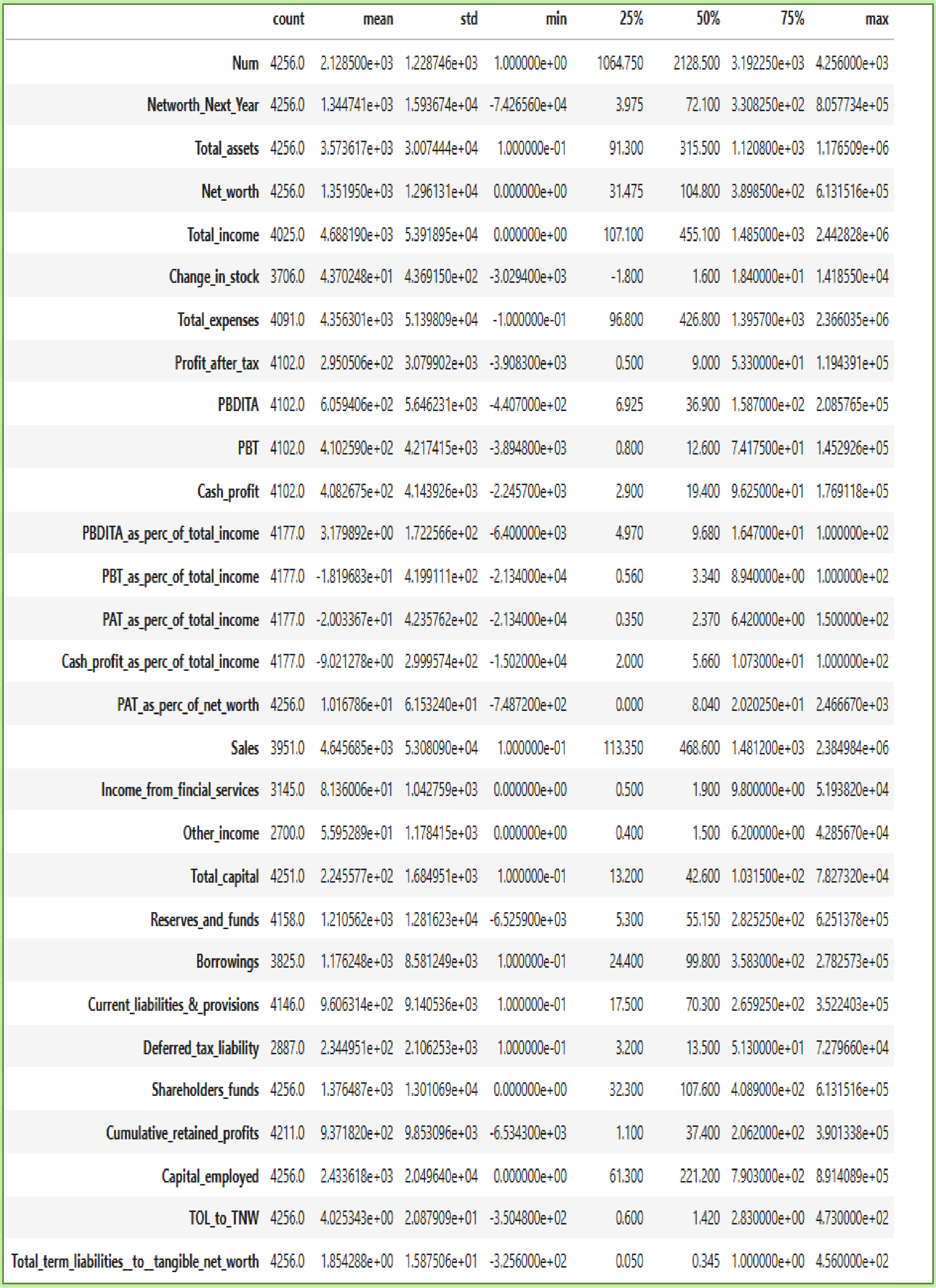
3. Risk & Leverage Factors

    - Debt-to-Equity Ratio, Total Liabilities, and Borrowings show leverage levels and potential default risks.
    - Liquidity Ratios (Quick Ratio, Current Ratio) highlight short-term financial health.

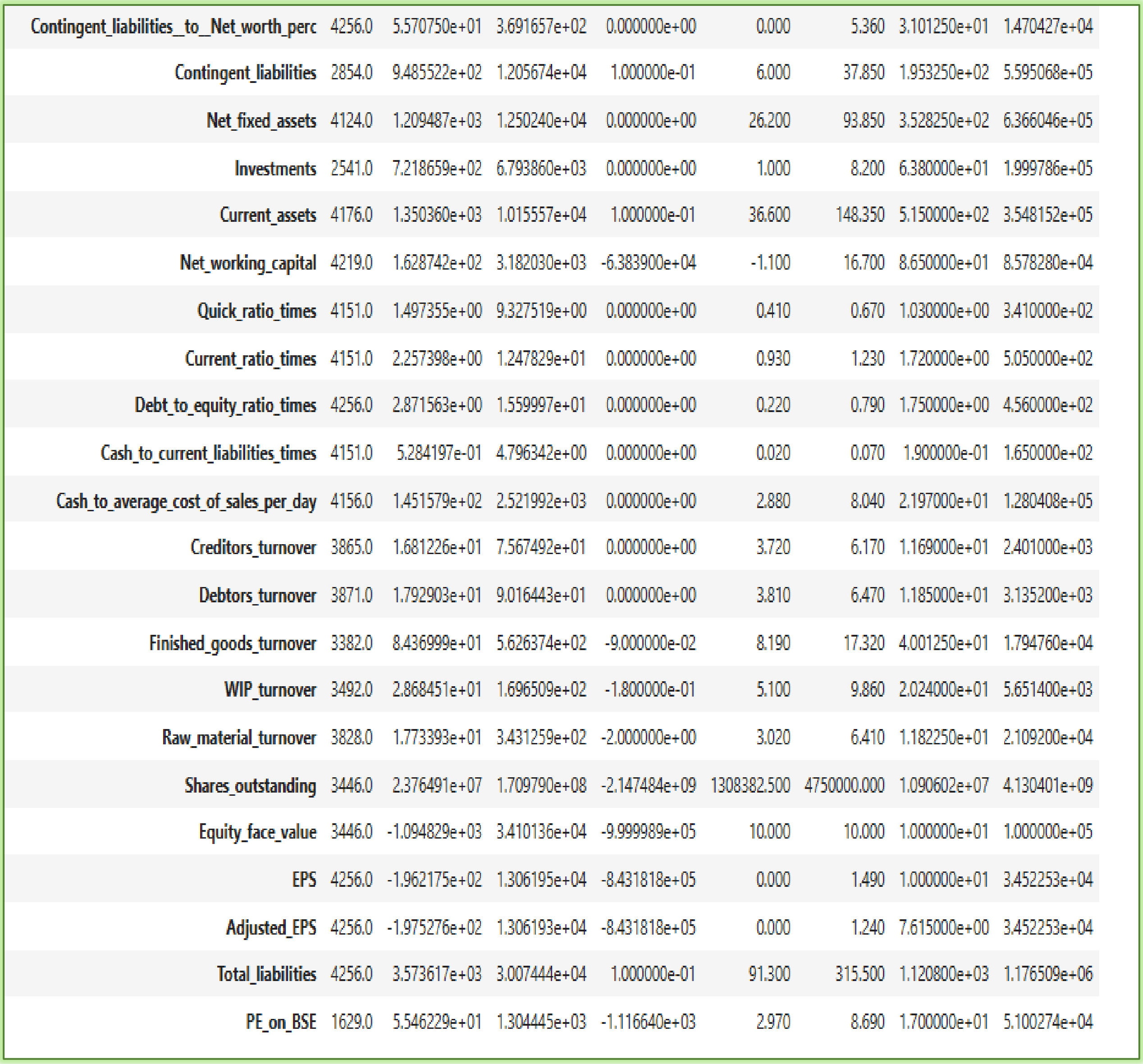
4. Potential Business Insights

    - Companies with missing sales or profitability data may indicate financial instability.
    - High leverage (borrowings + liabilities) vs profitability should be analyzed for default risk predictions.
    - Outliers in financial ratios (like PE_on_BSE) may indicate market inefficiencies or high-risk businesses.

3. STATISTICAL ANALYSIS



	count	mean	std	min	25%	50%	75%	max
Num	4256.0	2.128500e+03	1.228746e+03	1.000000e+00	1064.750	2128.500	3.192250e+03	4.256000e+03
Networth_Next_Year	4256.0	1.344741e+03	1.593674e+04	-7.426560e+04	3.975	72.100	3.308250e+02	8.057734e+05
Total_assets	4256.0	3.573617e+03	3.007444e+04	1.000000e-01	91.300	315.500	1.120800e+03	1.176509e+06
Net_worth	4256.0	1.351950e+03	1.296131e+04	0.000000e+00	31.475	104.800	3.898500e+02	6.131516e+05
Total_income	4025.0	4.688190e+03	5.391895e+04	0.000000e+00	107.100	455.100	1.485000e+03	2.442828e+06
Change_in_stock	3706.0	4.370248e+01	4.369150e+02	-3.029400e+03	-1.800	1.600	1.840000e+01	1.418550e+04
Total_expenses	4091.0	4.356301e+03	5.139809e+04	-1.000000e-01	96.800	426.800	1.395700e+03	2.366035e+06
Profit_after_tax	4102.0	2.950506e+02	3.079902e+03	-3.908300e+03	0.500	9.000	5.330000e+01	1.194391e+05
PBDITA	4102.0	6.059406e+02	5.646231e+03	-4.407000e+02	6.925	36.900	1.587000e+02	2.085765e+05
PBT	4102.0	4.102590e+02	4.217415e+03	-3.894800e+03	0.800	12.600	7.417500e+01	1.452926e+05
Cash_profit	4102.0	4.082675e+02	4.143926e+03	-2.245700e+03	2.900	19.400	9.625000e+01	1.769118e+05
PBDITA_as_perc_of_total_income	4177.0	3.179892e+00	1.722566e+02	-6.400000e+03	4.970	9.680	1.647000e+01	1.000000e+02
PBT_as_perc_of_total_income	4177.0	-1.819683e+01	4.199111e+02	-2.134000e+04	0.560	3.340	8.940000e+00	1.000000e+02
PAT_as_perc_of_total_income	4177.0	-2.003367e+01	4.235762e+02	-2.134000e+04	0.350	2.370	6.420000e+00	1.500000e+02
Cash_profit_as_perc_of_total_income	4177.0	-9.021278e+00	2.999574e+02	-1.502000e+04	2.000	5.660	1.073000e+01	1.000000e+02
PAT_as_perc_of_net_worth	4256.0	1.016786e+01	6.153240e+01	-7.487200e+02	0.000	8.040	2.020250e+01	2.466670e+03
Sales	3951.0	4.645685e+03	5.308090e+04	1.000000e-01	113.350	468.600	1.481200e+03	2.384984e+06
Income_from_fincial_services	3145.0	8.136006e+01	1.042759e+03	0.000000e+00	0.500	1.900	9.800000e+00	5.193820e+04
Other_income	2700.0	5.595289e+01	1.178415e+03	0.000000e+00	0.400	1.500	6.200000e+00	4.285670e+04
Total_capital	4251.0	2.245577e+02	1.684951e+03	1.000000e-01	13.200	42.600	1.031500e+02	7.827320e+04
Reserves_and_funds	4158.0	1.210562e+03	1.281623e+04	-6.525900e+03	5.300	55.150	2.825250e+02	6.251378e+05
Borrowings	3825.0	1.176248e+03	8.581249e+03	1.000000e-01	24.400	99.800	3.583000e+02	2.782573e+05
Current_liabilities_&_provisions	4146.0	9.606314e+02	9.140536e+03	1.000000e-01	17.500	70.300	2.659250e+02	3.522403e+05
Deferred_tax_liability	2887.0	2.344951e+02	2.106253e+03	1.000000e-01	3.200	13.500	5.130000e+01	7.279660e+04
Shareholders_funds	4256.0	1.376487e+03	1.301069e+04	0.000000e+00	32.300	107.600	4.089000e+02	6.131516e+05
Cumulative_retained_profits	4211.0	9.371820e+02	9.853096e+03	-6.534300e+03	1.100	37.400	2.062000e+02	3.901338e+05
Capital_employed	4256.0	2.433618e+03	2.049640e+04	0.000000e+00	61.300	221.200	7.903000e+02	8.914089e+05
TOL_to_TNW	4256.0	4.025343e+00	2.087909e+01	-3.504800e+02	0.600	1.420	2.830000e+00	4.730000e+02
Total_term_liabilities__to_tangible_net_worth	4256.0	1.854288e+00	1.587506e+01	-3.256000e+02	0.050	0.345	1.000000e+00	4.560000e+02



Contingent_liabilities_to_Net_worth_perc	4256.0	5.570750e+01	3.691657e+02	0.000000e+00	0.000	5.360	3.101250e+01	1.470427e+04
Contingent_liabilities	2854.0	9.485522e+02	1.205674e+04	1.000000e-01	6.000	37.850	1.953250e+02	5.595068e+05
Net_fixed_assets	4124.0	1.209487e+03	1.250240e+04	0.000000e+00	26.200	93.850	3.528250e+02	6.366046e+05
Investments	2541.0	7.218659e+02	6.793860e+03	0.000000e+00	1.000	8.200	6.380000e+01	1.999786e+05
Current_assets	4176.0	1.350360e+03	1.015557e+04	1.000000e-01	36.600	148.350	5.150000e+02	3.548152e+05
Net_working_capital	4219.0	1.628742e+02	3.182030e+03	-6.383900e+04	-1.100	16.700	8.650000e+01	8.578280e+04
Quick_ratio_times	4151.0	1.497355e+00	9.327519e+00	0.000000e+00	0.410	0.670	1.030000e+00	3.410000e+02
Current_ratio_times	4151.0	2.257398e+00	1.247829e+01	0.000000e+00	0.930	1.230	1.720000e+00	5.050000e+02
Debt_to_equity_ratio_times	4256.0	2.871563e+00	1.559997e+01	0.000000e+00	0.220	0.790	1.750000e+00	4.560000e+02
Cash_to_current_liabilities_times	4151.0	5.284197e-01	4.796342e+00	0.000000e+00	0.020	0.070	1.900000e-01	1.650000e+02
Cash_to_average_cost_of_sales_per_day	4156.0	1.451579e+02	2.521992e+03	0.000000e+00	2.880	8.040	2.197000e+01	1.280408e+05
Creditors_turnover	3865.0	1.681226e+01	7.567492e+01	0.000000e+00	3.720	6.170	1.169000e+01	2.401000e+03
Debtors_turnover	3871.0	1.792903e+01	9.016443e+01	0.000000e+00	3.810	6.470	1.185000e+01	3.135200e+03
Finished_goods_turnover	3382.0	8.436999e+01	5.626374e+02	-9.000000e-02	8.190	17.320	4.001250e+01	1.794760e+04
WIP_turnover	3492.0	2.868451e+01	1.696509e+02	-1.800000e-01	5.100	9.860	2.024000e+01	5.651400e+03
Raw_material_turnover	3828.0	1.773393e+01	3.431259e+02	-2.000000e+00	3.020	6.410	1.182250e+01	2.109200e+04
Shares_outstanding	3446.0	2.376491e+07	1.709790e+08	-2.147484e+09	1308382.500	4750000.000	1.090602e+07	4.130401e+09
Equity_face_value	3446.0	-1.094829e+03	3.410136e+04	-9.999989e+05	10.000	10.000	1.000000e+01	1.000000e+05
EPS	4256.0	-1.962175e+02	1.306195e+04	-8.431818e+05	0.000	1.490	1.000000e+01	3.452253e+04
Adjusted_EPS	4256.0	-1.975276e+02	1.306193e+04	-8.431818e+05	0.000	1.240	7.615000e+00	3.452253e+04
Total_liabilities	4256.0	3.573617e+03	3.007444e+04	1.000000e-01	91.300	315.500	1.120800e+03	1.176509e+06
PE_on_BSE	1629.0	5.546229e+01	1.304445e+03	-1.116640e+03	2.970	8.690	1.700000e+01	5.100274e+04

Figure A. 3 Statistical Summary

Observation:

1. Financial Stability & Growth

    - Net Worth varies widely (Median: 104.8, Mean: 1351.95, Max: 613,151.6), indicating a mix of small and large businesses with significant financial disparity.
    - The projected Net Worth for the next year (Median: 72.1, Mean: 1344.74) shows high volatility, reflecting uncertain growth patterns.

2. Income & Profitability Trends

    - Total Income is highly skewed (Mean: 4688.19, Median: 455.1), with a few businesses generating exceptionally high revenue.
    - Total Expenses (Mean: 4356.3, Median: 426.8) highlight substantial operational costs across businesses.

- Profit After Tax varies significantly (Mean: 295.05, Median: 9.0), with some businesses facing heavy losses while others record substantial profits.
- Pre-Tax Profitability (% of Total Income) averages at -18.2%, but the median (3.34%) suggests that many businesses remain profitable despite some high losses.

3. Asset & Liability Management

- Total Assets (Mean: 3,573.61, Median: 315.5) indicate that most businesses have moderate asset bases, with a few holding exceptionally large assets.
- Current Liabilities & Provisions (Mean: 960.63, Median: 70.3) suggest that some companies bear substantial financial obligations.
- The Debt-to-Equity Ratio (Mean: 2.87, Median: 0.79) implies that while most businesses maintain reasonable leverage, a few rely heavily on debt financing.

4. Liquidity & Working Capital

- Current Ratio (Mean: 2.26, Median: 1.23) suggests that businesses generally have more assets than liabilities, ensuring short-term stability.
- Quick Ratio (Mean: 1.5, Median: 0.67) indicates that some businesses maintain strong liquidity, while others may face challenges covering short-term obligations.
- Net Working Capital (Mean: 162.87, Median: 16.7) highlights variability in financial health, with some businesses experiencing liquidity shortages.

5. Operational Efficiency & Turnover

- The Debtor Turnover Ratio (Mean: 17.93, Median: 6.47) shows variation in how quickly businesses collect payments, with some being highly efficient.
- The Creditor Turnover Ratio (Mean: 16.81, Median: 6.17) suggests differences in how businesses manage supplier payments.
- Finished Goods Turnover (Mean: 84.37, Median: 17.32) reflects that while some businesses move inventory quickly, others may struggle with stock turnover.

4. EXPLORATORY DATA ANALYSIS

4.1 UNIVARIATE ANALYSIS

Creating a binary target variable using 'Networth_Next_Year'

The new column Default categorizes the data into two groups:

0: Indicates that the Networth_Next_Year is positive (which may indicate financial stability or a non-default situation).

1: Indicates that the Networth_Next_Year is zero or negative (which might be a sign of potential financial distress or default risk).

4.1.1 Count Plot for the Target Variable

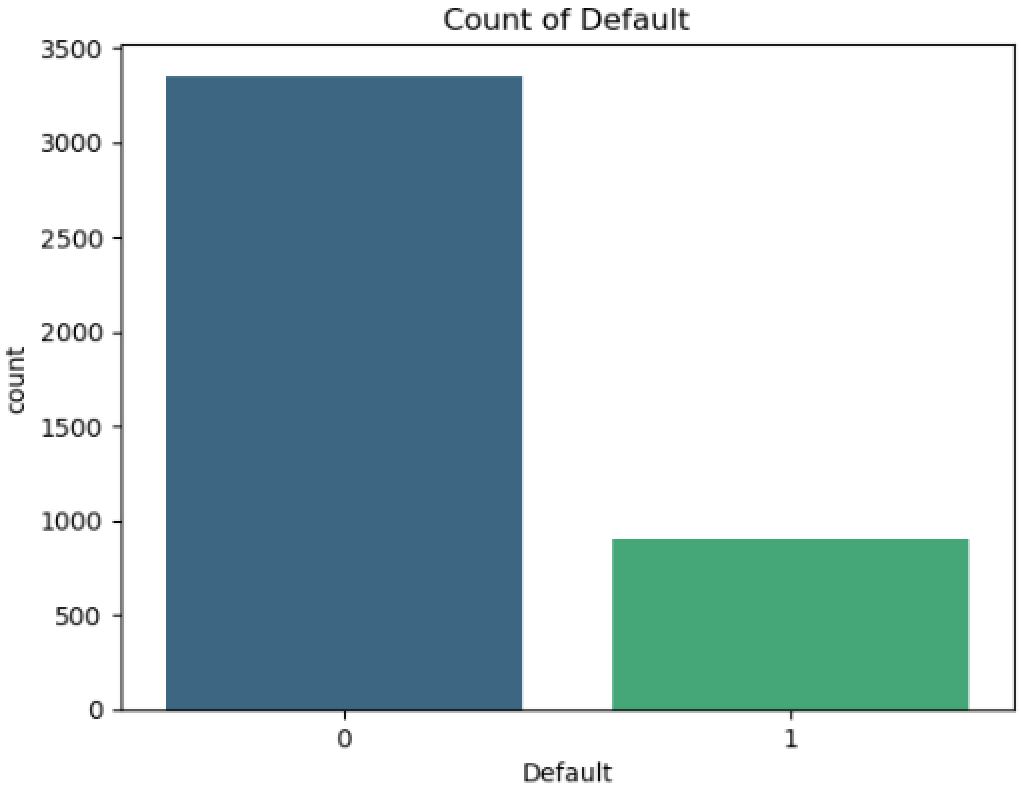
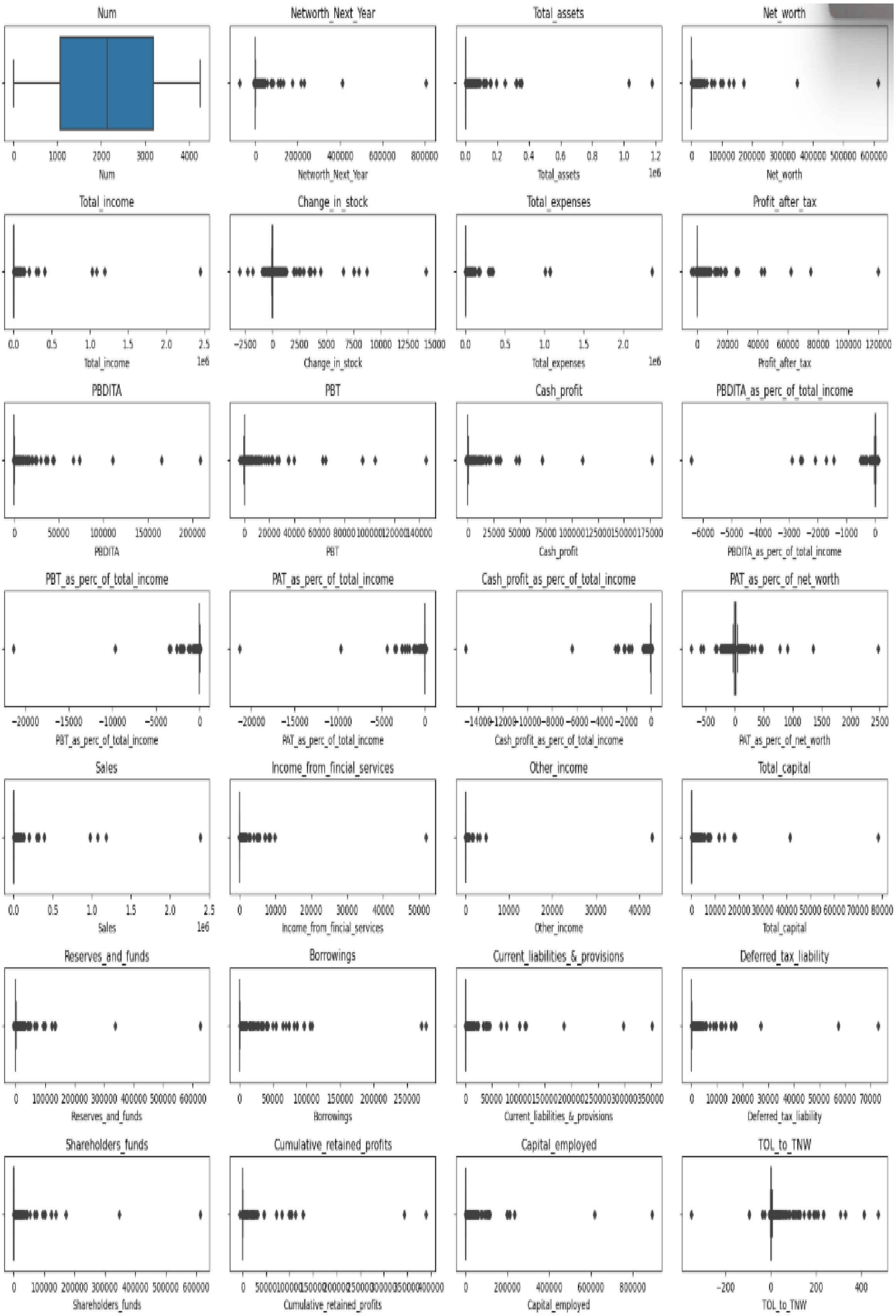


Figure A. 4 Count Plot for the Target Variable

Observation from the Default Count Plot

- The dataset is highly imbalanced, with significantly more non-default cases (0) than default cases (1).
- The majority of companies do not default, while a smaller proportion falls into the default category.
- This imbalance can affect predictive modeling, requiring techniques such as oversampling (SMOTE), undersampling, or class weighting to improve model performance.
- Given the disparity, recall (sensitivity) should be a key focus in evaluating model performance to ensure the model correctly identifies defaulters.

4.1.2 Boxplots for all the Numerical Columns



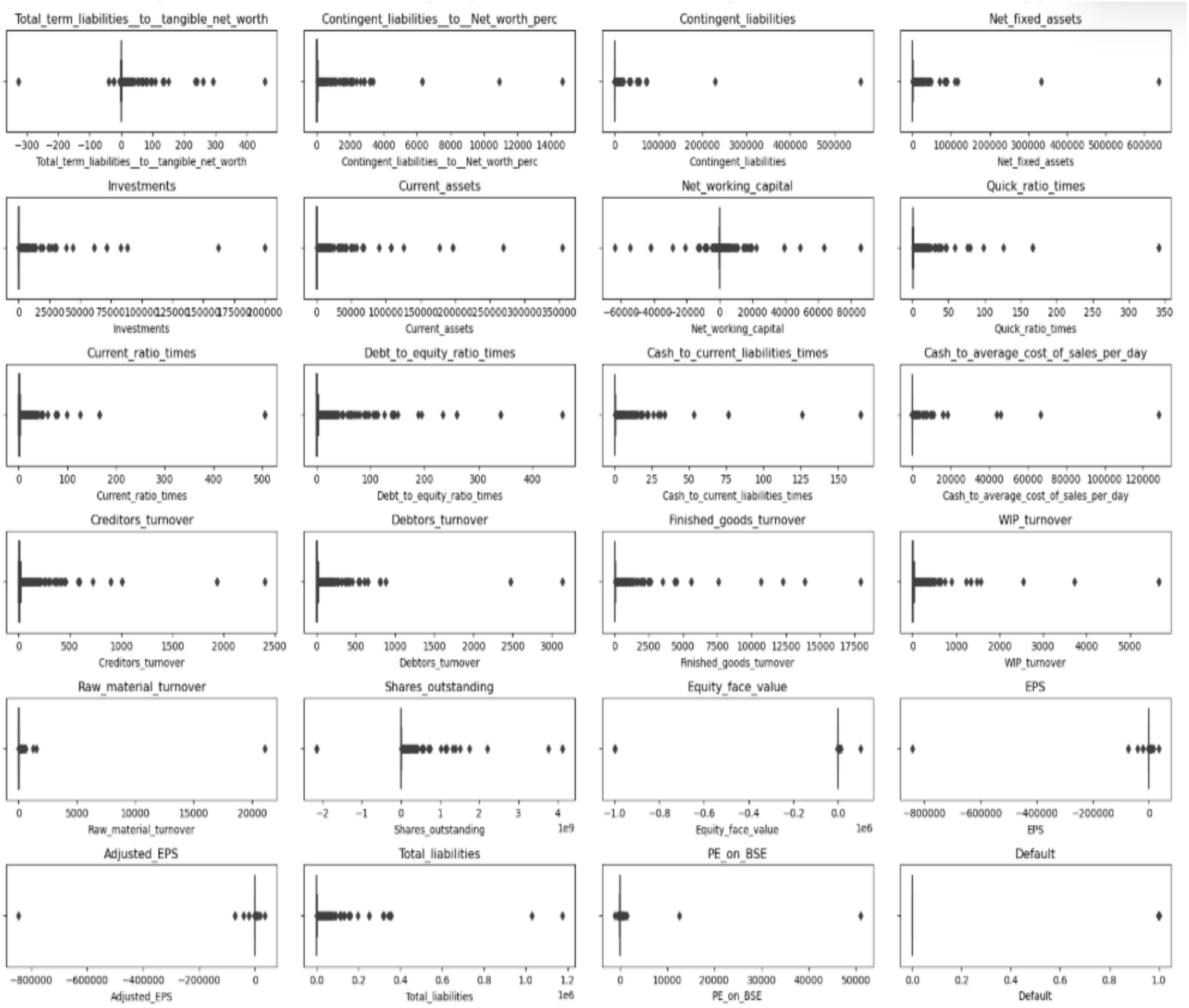


Figure A. 5 Boxplots for all the Numerical Columns

Observation

1. Outliers and Extreme Values
   - Several financial variables exhibit extreme values that could influence analysis outcomes.
   - Key metrics such as Net Worth, Total Income, Total Assets, Borrowings, and Current Liabilities contain significant outliers.
   - A deeper investigation is required to determine if these values are valid business trends or potential data anomalies.
2. Skewness in Financial Metrics
   - Many financial variables show right-skewed distributions, indicating a few businesses hold disproportionately high financial Figure A.s.
   - This skewness may impact statistical modeling, necessitating techniques like log transformation or normalization for better data representation.
3. Net Worth and Default Correlation

- A noticeable difference in Net Worth between defaulting and non-defaulting businesses suggests that companies with lower net worth may have a higher probability of defaulting.
- Identifying a critical net worth threshold could help in risk assessment and early intervention strategies.

4. Debt Burden and Default Risk
   - Higher borrowings and liabilities may be strong indicators of financial stress.
   - If defaulting companies consistently show elevated debt levels, it emphasizes the need for proactive debt management strategies.
5. Profitability and Default Trends
   - Businesses with lower or negative profitability indicators such as Profit After Tax (PAT), EBIT, and ROCE appear more susceptible to default.
   - Tracking declining profit trends can serve as a predictive measure for early detection of financial distress, aiding in risk mitigation efforts.

4.1.3 Violin Plots for all the Numerical Columns

Violin plots provide a richer view of data by combining density estimation with a box plot, making it easier to detect skewness, clustering, and outliers.

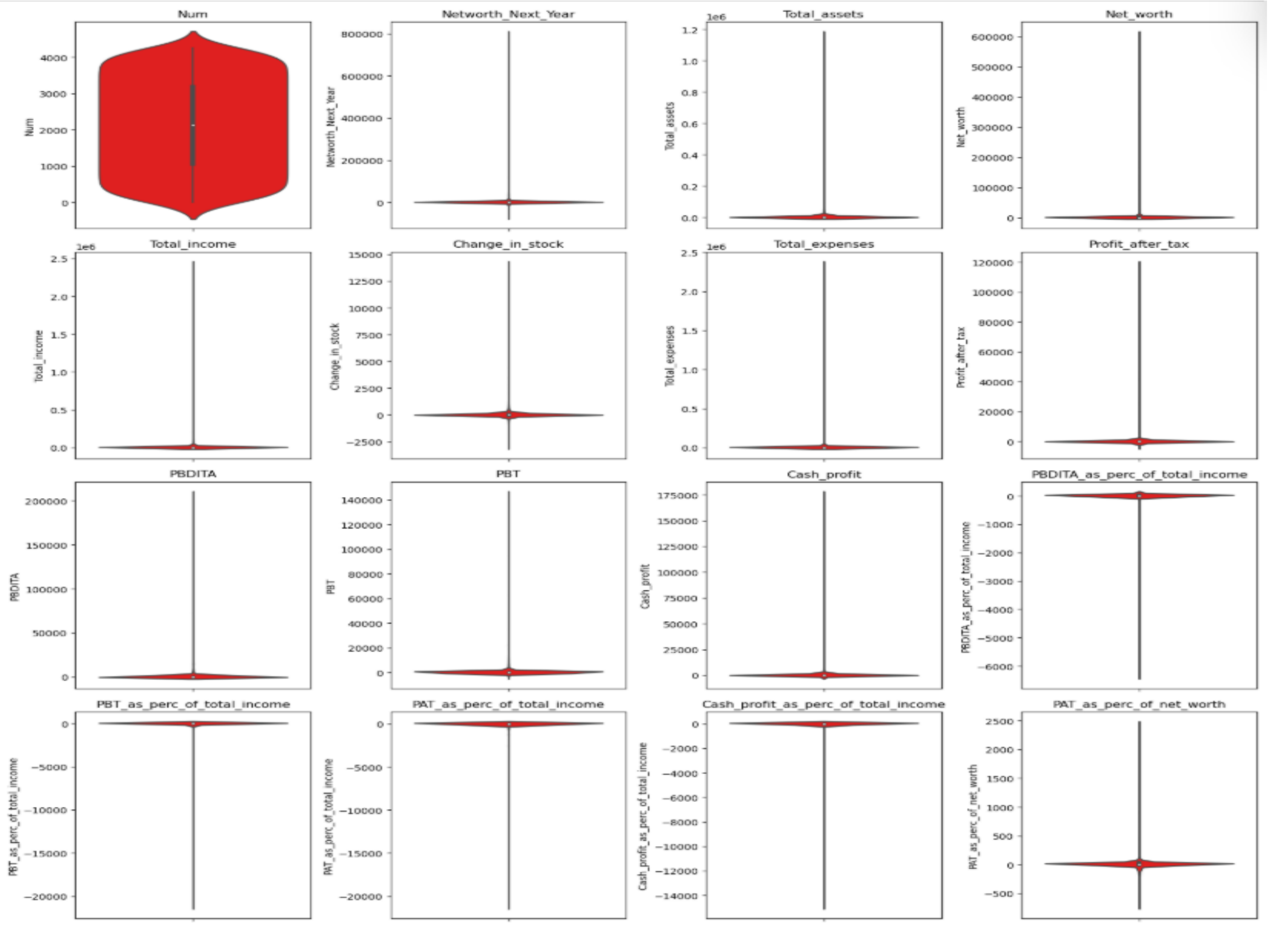


Figure A. 6 Violin Plots for all the Numerical Columns

Observations from Violin Plots: Distribution & Outliers Analysis

1. Highly Skewed Distributions

   - Most variables, including Net Worth, Total Assets, Total Income, and Profit After Tax (PAT), show extreme right skewness.

   - A small number of companies have significantly higher values, leading to elongated upper tails in the distributions.

2. Presence of Outliers

   - Several financial indicators, such as Total Income, Total Assets, and Profitability Ratios, exhibit a narrow central region with extreme values extending far beyond the bulk of the data.

   - This suggests the presence of companies with disproportionately high financial Figure A.s compared to the majority.

3. Profitability and Liquidity Metrics Show Variation

    - PBDITA, PBT, and Cash Profit distributions reveal that most companies operate on thin margins, with many data points clustered near zero.
    - A subset of companies has high profitability, as indicated by the long tails in these plots.

4. Negative Values Indicating Financial Struggles
    - Some variables, such as PAT % of Net Worth, Cash Profit %, and PBT % of Total Income, show data points extending into negative regions, indicating financial distress for some firms.
    - This suggests liquidity issues, high debt burdens, or operational inefficiencies leading to losses.
5. Narrow Spread in Key Ratios
    - The profitability ratios (PBDITA %, PBT %, PAT %) exhibit highly compressed distributions, with most values concentrated near the center.
    - This implies that while a few companies achieve substantial profitability, most firms operate with relatively modest margins.

Conclusion

- The violin plots highlight severe skewness and extreme outliers in key financial metrics, emphasizing the need for log transformation or normalization for better model performance.
- A significant number of firms experience financial strain, as seen in the negative profitability indicators.
- The presence of extreme values in Net Worth, Total Income, and Total Assets suggests that a few large companies dominate the dataset.
- Further segmentation analysis by industry or company size could help understand patterns in financial health and risk factors.

4.1.4 Hisplots for all the Numerical Columns

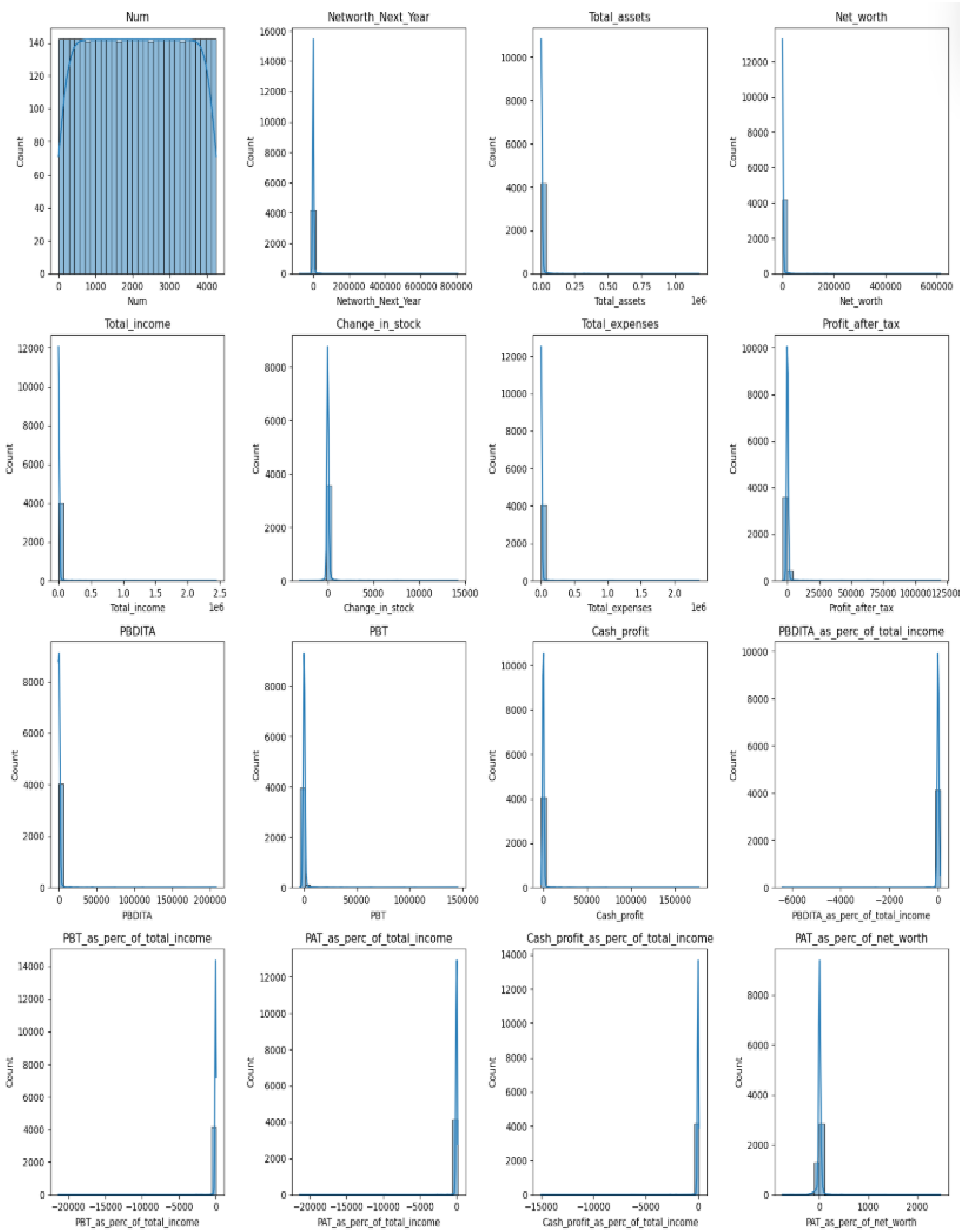


Figure A. 7 Hisplots for all the Numerical Columns

Observations from the Distribution Plots

1. Highly Skewed Distributions
   - Most financial variables, including Net Worth, Total Assets, Total Income, and Profit After Tax (PAT), exhibit right-skewed distributions.
   - A small number of firms have significantly high values, leading to long tails in the distributions.
2. Presence of Outliers
   - Several variables, including Net Worth Next Year, Total Income, Total Expenses, and PBDITA, show extreme outliers.
   - These outliers may represent financially strong businesses or potential data anomalies that require further investigation.
3. Profitability Metrics Concentrated Near Zero
   - PBDITA, PBT, and Cash Profit have distributions highly concentrated near zero, indicating that many firms operate on thin profit margins.
   - Some firms show negative values, suggesting financial distress or losses.
4. Negative Profitability Ratios
   - PBDITA %, PBT %, PAT % of Net Worth, and Cash Profit % of Total Income show a significant portion of firms with negative values.
   - This indicates that many companies are experiencing financial strain, potentially affecting long-term sustainability.
5. Liquidity Challenges
   - The Cash Profit distribution shows skewness, indicating that while some firms have strong liquidity, others may struggle with cash flow issues.
   - This is crucial for assessing a company's ability to manage short-term obligations and avoid default.

Conclusion

- The data exhibits high skewness and outliers in multiple financial variables, indicating the need for normalization techniques such as log transformation.
- The presence of negative profit and profitability ratios suggests financial vulnerability for many firms.
- Further analysis is needed to understand if specific industries or business types are more prone to financial distress.

4.1.5 Cumulative Distribution of Financial Variables

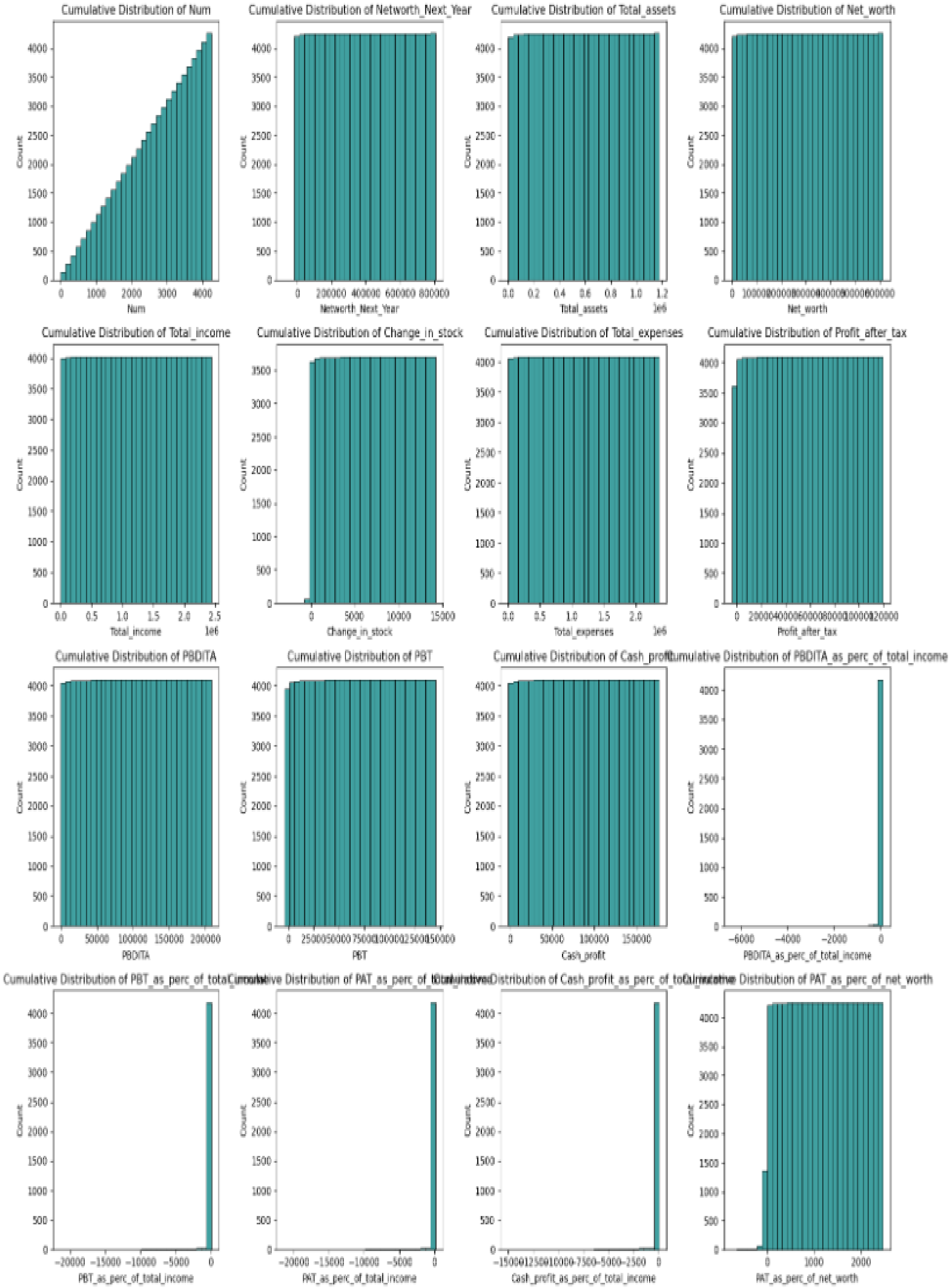


Figure A. 8 Cumulative Distribution of Financial Variables

Observations on Cumulative Distribution of Financial Variables

The provided cumulative distribution plots highlight the overall data distribution trends, helping identify skewness, outliers, and potential insights about company financial performance.

Key Observations:

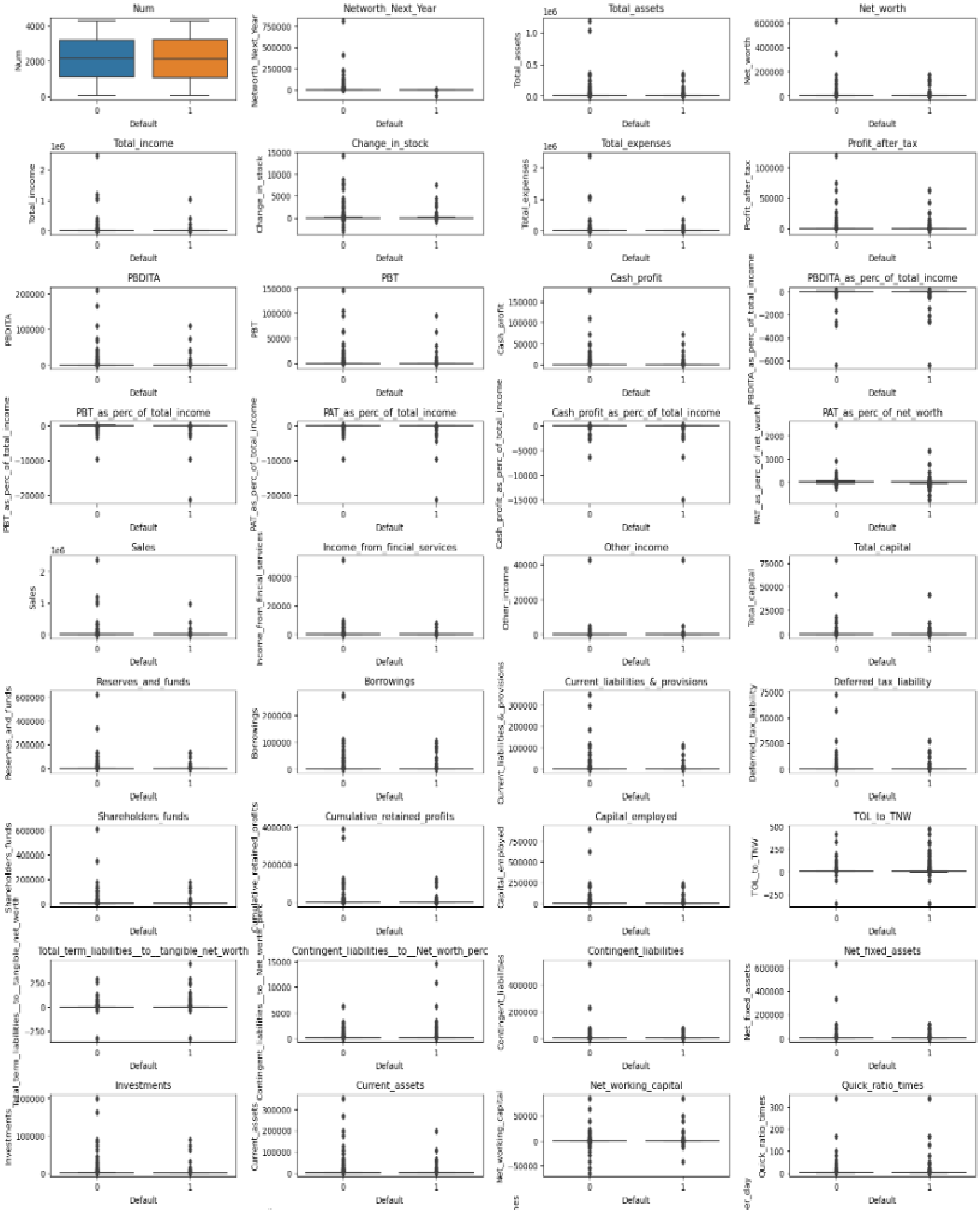
1. Right-Skewed Distribution in Financial Metrics:
   - Variables like Net Worth, Total Assets, Total Income, Profit After Tax (PAT), PBDITA, and Cash Profit exhibit strong right skewness.
   - A few companies have significantly high values, leading to extreme outliers.
   - Most firms fall within a low to moderate range, with only a few reaching extremely high values.
2. Cumulative Distribution of Net Worth and Total Assets:
   - A large proportion of firms have low values, while only a few exhibit extremely high values.
   - This suggests financial disparity among companies, with only a minority having substantial assets or net worth.
3. Profitability Ratios and Cash Flow Trends:
   - PBDITA as % of Total Income, PBT as % of Total Income, and PAT as % of Net Worth show concentration around zero or negative values, indicating that many companies operate with thin margins or losses.
   - Cash Profit and related percentages reveal liquidity concerns, as some firms have significantly negative values, potentially indicating financial distress or cash flow challenges.
4. Presence of Extreme Negative Values:
   - Some firms exhibit high negative values in profitability ratios, implying losses or financial instability.
   - These companies may be struggling to sustain profitability, requiring further investigation into financial risk factors.

Conclusion:

- The cumulative distributions suggest a highly skewed financial landscape, with a few companies dominating total financial Figure A.s while the majority remain in the lower range.
- High negative values in profitability and cash flow metrics indicate potential financial distress, requiring further analysis to assess business risks.
- Outliers and extreme values affect financial trends, and appropriate transformations (e.g., log scaling) may be necessary for better analysis.

4.2 BIVARIATE & MULTIVARIATE ANALYSIS

4.2.1 Financial Variables and Default Risk



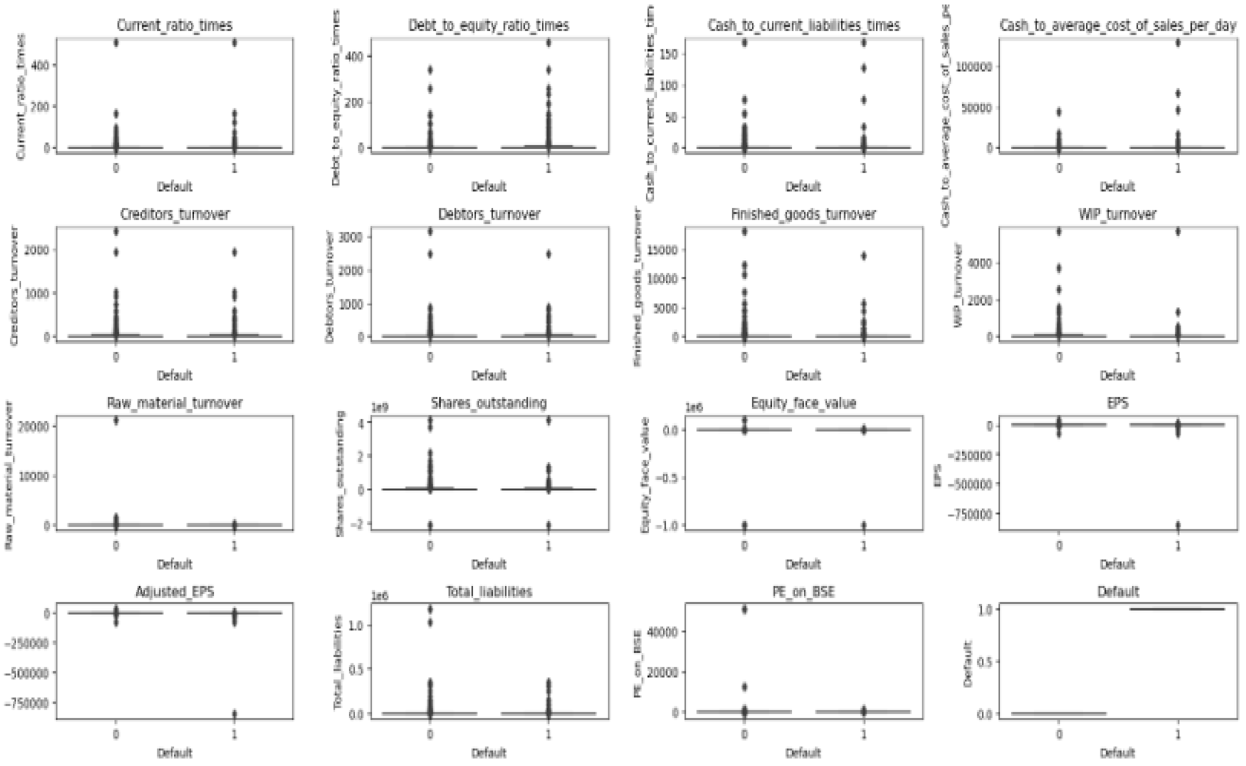


Figure A. 9 Boxplot for Financial Variables and Default Risk

Key Observations on Default Risk

1. Defaulters have lower financial health:
    - Lower Net Worth, Total Income, Profitability Ratios (PAT, PBDITA), and Cash Profits.
    - Higher Borrowings, Liabilities, and Debt-to-Equity Ratio indicate financial distress.
2. Profitability & Liquidity Issues:
    - Negative PBT, Cash Profits, and EPS suggest struggling businesses.
    - Low Current & Quick Ratios, Weak Net Working Capital indicate liquidity problems.
3. Market Indicators are not reliable predictors:
    - High variation in PE Ratio, Shares Outstanding, and EPS suggests weak correlation with default risk.

Takeaway:

High debt, low profitability, and poor liquidity strongly correlate with default risk.

4.2.2 Heatmap

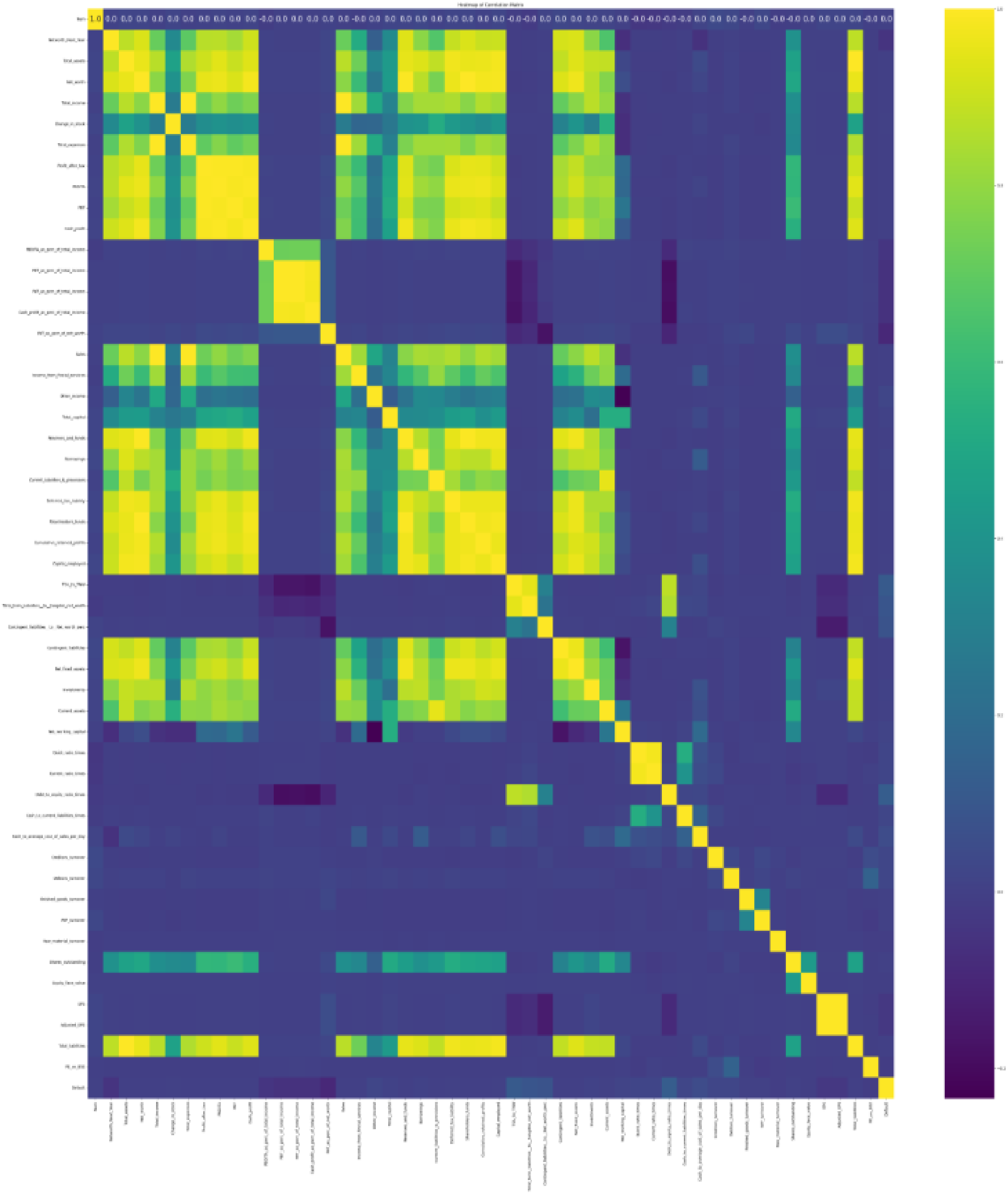


Figure A. 10 Heatmap for numeric columns

Multicollinearity Issues:

- Clusters of highly correlated variables suggest redundancy.
- Feature selection or dimensionality reduction (e.g., PCA or VIF analysis) may be needed.

4.2.3 Important Financial Indicators for Default Prediction

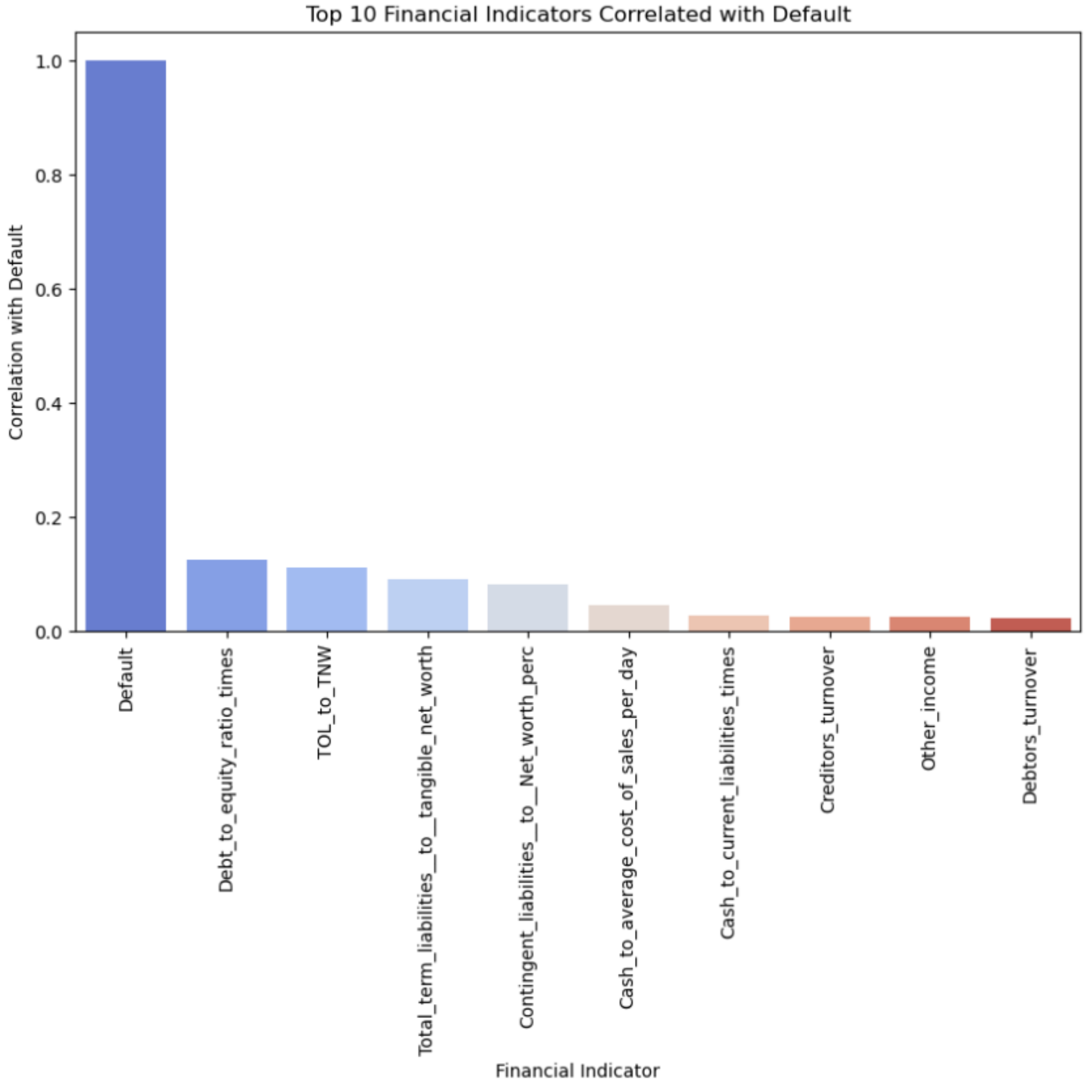


Figure A. 11 Important Financial Indicators for Default Prediction

Important Financial Indicators for Default Prediction – Observations

1. High Debt Levels Increase Default Risk:
   - Debt-to-Equity Ratio and Total Outside Liabilities (TOL) to Tangible Net Worth (TNW) are strong predictors.
   - Companies with excessive leverage are more likely to default.
2. Weakened Liquidity Raises Concerns:
   - Cash to Average Cost of Sales per Day and Cash to Current Liabilities indicate liquidity issues.
   - Insufficient cash reserves can lead to financial distress.
3. Liability Management is Crucial:

- Total Term Liabilities to Tangible Net Worth and Contingent Liabilities to Net Worth highlight the importance of managing long-term obligations.

4. Turnover Ratios Play a Role:
   - Debtors Turnover and Creditors Turnover indicate that inefficient working capital management can contribute to default.
5. Other Income Has Some Influence:
   - Companies relying more on Other Income rather than core business revenue might face sustainability issues.

Key Takeaway:

Businesses with high debt, poor liquidity, and weak liability management are at higher risk of default. Strong financial risk control is essential for sustainability.

4.2.4 Debt Management Practices of Defaulters vs. Non-Defaulters

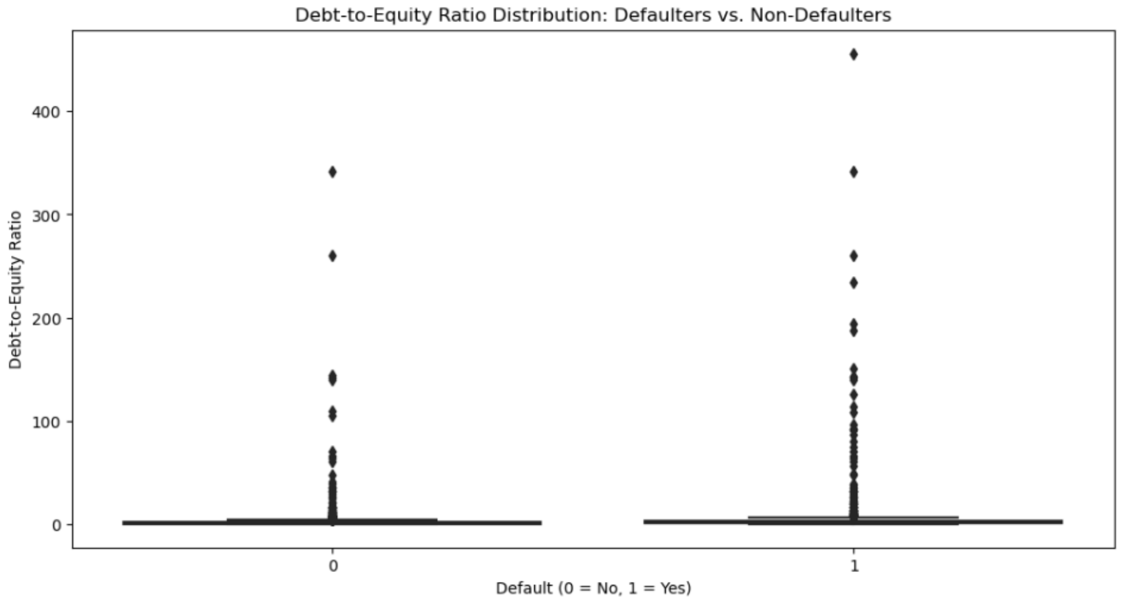


Figure A. 12 Boxplot of Debt Management Practices of Defaulters vs. Non-Defaulters

Observations on Debt Management Practices of Defaulters vs. Non-Defaulters

1. Extreme Debt-to-Equity Ratios in Both Groups:
   - Both defaulters and non-defaulters have outliers with very high Debt-to-Equity Ratios, indicating that some companies take on excessive leverage.
2. Higher Debt Exposure in Defaulters:
   - Although both groups show a wide range, defaulters tend to have more instances of extreme debt levels.
   - This suggests that excessive leverage is a key risk factor for default.

3. Majority of Companies Maintain Low Debt Levels:
    - Most companies in both categories have low Debt-to-Equity Ratios, clustered near zero.
    - This implies that while high debt increases risk, some firms with moderate leverage still manage to avoid default.

Key Takeaway:

- Excessive debt is a major risk factor for default. However, not all highly leveraged firms default, indicating that other financial health indicators also play a role.

4.2.5 Financial Ratios & Risk Assessment

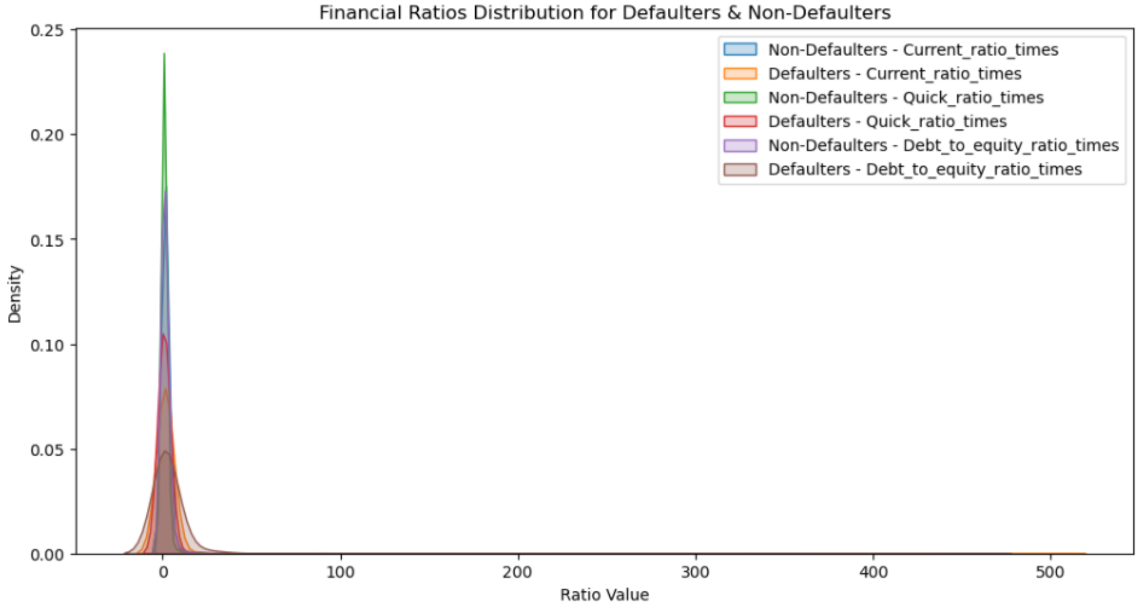


Figure A. 13 Financial Ratios Distribution for Defaulters & Non-Defaulters

Observations on Financial Ratios Distribution for Defaulters & Non-Defaulters

1. Highly Skewed Distribution:
    - All financial ratios (Current Ratio, Quick Ratio, Debt-to-Equity Ratio) exhibit extreme skewness with long right tails, indicating a few companies have very high values.
2. Similar Distribution for Defaulters & Non-Defaulters:
    - There is no sharp distinction between defaulters and non-defaulters in terms of ratio distribution, suggesting other factors influence default risk.
3. Debt-to-Equity Ratio is More Dispersed:
    - Compared to liquidity ratios, Debt-to-Equity Ratio shows greater variability, reinforcing its importance in assessing financial stability.

Key Takeaway:

- Liquidity and leverage ratios alone do not clearly separate defaulters from non-defaulters. Further analysis incorporating additional financial metrics is needed.

4.2.6 Patterns in Financial Data for Risk Assessment

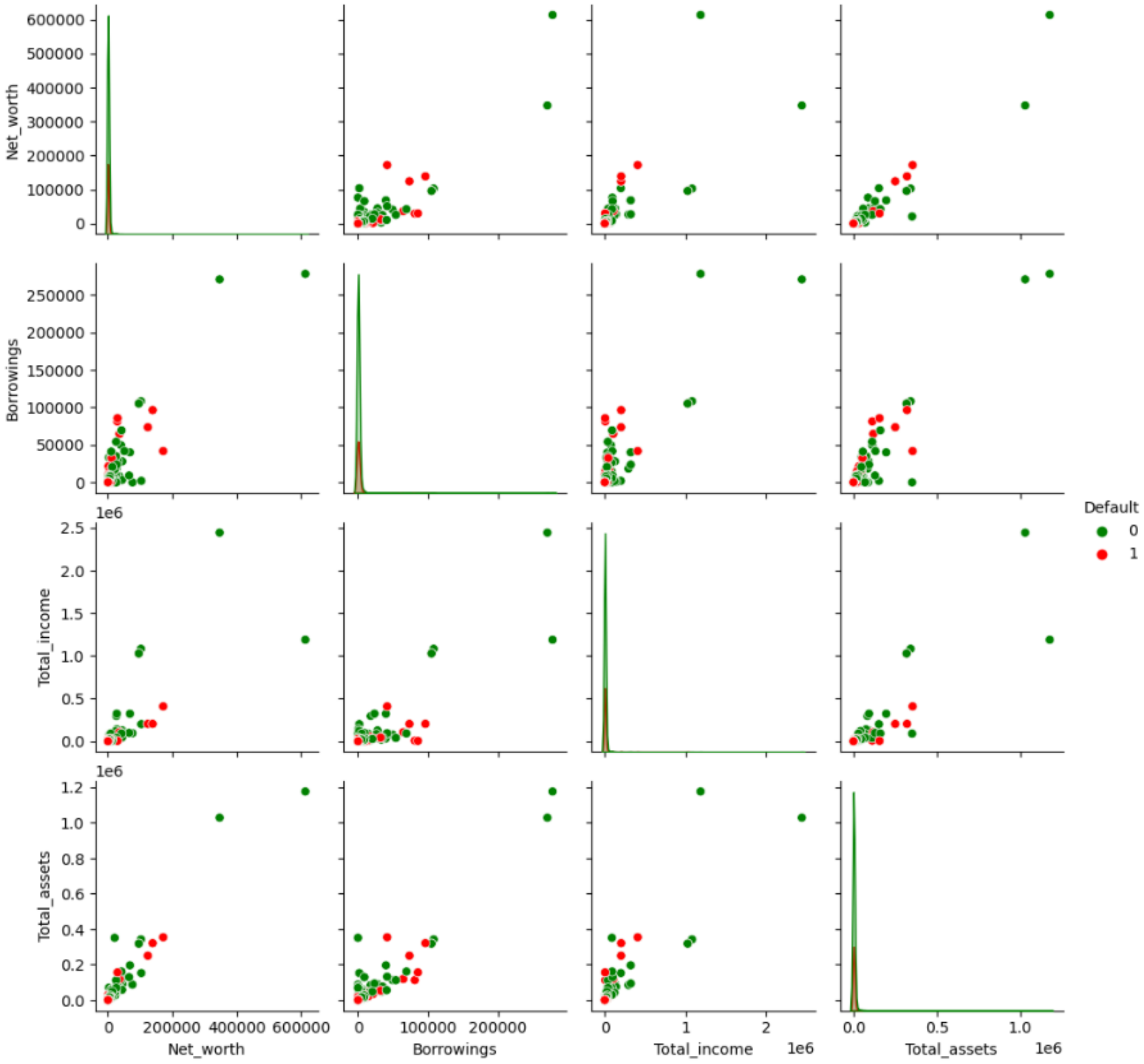


Figure A. 14 Pair Plot of Financial Data for Risk Assessment

Observations on Financial Metrics and Default Trends

1. Defaulters (Red) Tend to Have Lower Net Worth:
   - Many defaulters are concentrated in the lower range of net worth, indicating financial instability.
2. Borrowings are High Across Both Groups:

- Some defaulters have high borrowings, but so do non-defaulters, suggesting that borrowings alone do not directly signal default risk.

3. Positive Correlation Among Financial Metrics:
   - Net Worth, Total Income, and Total Assets show a general upward trend together, meaning financially stronger firms tend to have all three in higher ranges.
4. Defaulters Appear in Lower to Mid-Income & Asset Ranges:
   - Defaulting firms generally cluster in the lower-to-moderate income and asset range, reinforcing the need to assess financial health holistically.

Key Takeaway:

- Lower net worth and moderate borrowings could indicate a higher likelihood of default. However, more factors need to be considered for a complete risk assessment.

5.DATA PRE-PROCESSING

5.1 Feature Engineering

Data Preparation for Modelling:

Creating a binary target variable using 'Networth_Next_Year'

The new column Default categorizes the data into two groups:

0: Indicates that the Networth_Next_Year is positive (which may indicate financial stability or a non-default situation).

1: Indicates that the Networth_Next_Year is zero or negative (which might be a sign of potential financial distress or default risk).

The target variable, Default, derived from Networth_Next_Year', is separated from the rest of the dataset. The data is then split into training (75%) and testing (25%) sets to facilitate effective model training and evaluation for predicting defaulters.

5.1.1 Outlier Detection

Number of outliers in each column:

SI.NO	Column	No. of outliers
0	Num	0
1	Networth_Next_Year	624

SI.NO	Column	No. of outliers
2	Total_assets	585
3	Net_worth	595
4	Total_income	508
5	Change_in_stock	750
6	Total_expenses	518
7	Profit_after_tax	712
8	PBDITA	584
9	PBT	704
10	Cash_profit	627
11	PBDITA_as_perc_of_total_income	346
12	PBT_as_perc_of_total_income	546
13	PAT_as_perc_of_total_income	610
14	Cash_profit_as_perc_of_total_income	426
15	PAT_as_perc_of_net_worth	427
16	Sales	500
17	Income_from_fincial_services	517
18	Other_income	389
19	Total_capital	551

SI.NO	Column	No. of outliers
20	Reserves_and_funds	643
21	Borrowings	532
22	Current_liabilities_&_provisions	581
23	Deferred_tax_liability	406
24	Shareholders_funds	588
25	Cumulative_retained_profits	699
26	Capital_employed	572
27	TOL_to_TNW	414
28	Total_term_liabilities__to__tangible_net_worth	406
29	Contingent_liabilities__to__Net_worth_perc	478
30	Contingent_liabilities	393
31	Net_fixed_assets	569
32	Investments	451
33	Current_assets	532
34	Net_working_capital	806
35	Quick_ratio_times	371
36	Current_ratio_times	397
37	Debt_to_equity_ratio_times	381

SI.NO	Column	No. of outliers
38	Cash_to_current_liabilities_times	539
39	Cash_to_average_cost_of_sales_per_day	583
40	Creditors_turnover	442
41	Debtors_turnover	408
42	Finished_goods_turnover	399
43	WIP_turnover	378
44	Raw_material_turnover	296
45	Shares_outstanding	476
46	Equity_face_value	533
47	EPS	638
48	Adjusted_EPS	694
49	Total_liabilities	585
50	PE_on_BSE	237
51	Default	904

Figure A. 15 Number of Outliers in each column

Key Insights on Outliers

1. Financial Instability: High outliers in Net Worth (595), Total Assets (585), and Borrowings (532) indicate extreme variations in financial stability.
2. Profitability Fluctuations: Metrics like Profit After Tax (712) and PBT (704) show major deviations, suggesting inconsistent earnings across companies.
3. Liquidity & Debt Risks: Debt-to-Equity Ratio (381), Current Ratio (397), and Quick Ratio (371) highlight firms with high leverage or weak liquidity.

4. Operational Inconsistencies: Large outliers in Sales (500) and Total Expenses (518) suggest fluctuations in revenue and cost structures.
5. Default Outliers (904): The highest outliers in default cases confirm that financial distress varies widely, making robust predictive modeling essential.

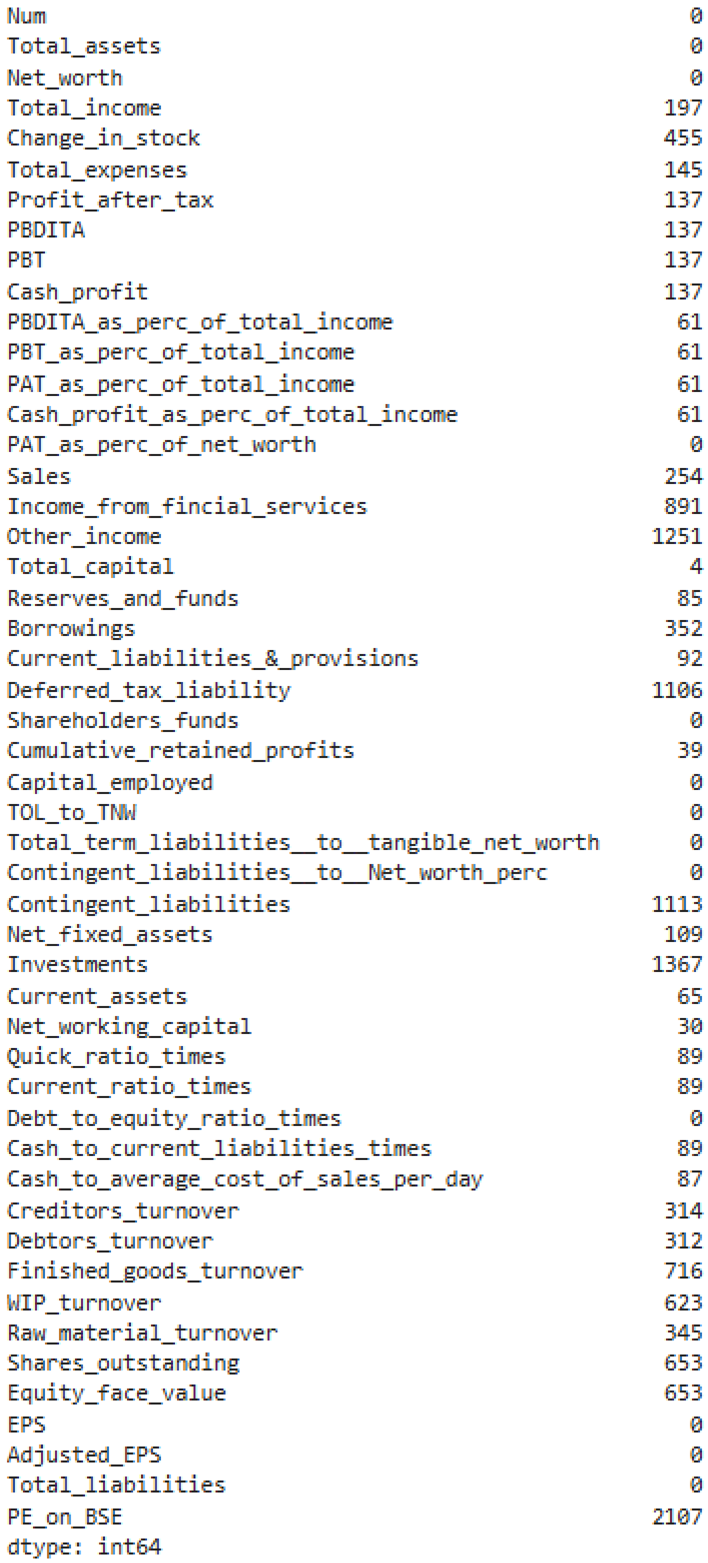
Actionable Takeaways

- Handle outliers carefully to improve model accuracy.
- Focus on debt, liquidity, and profitability metrics to predict defaults.
- Use scaling or transformation to prevent extreme values from skewing results.

Since we are predicting defaults, outliers in key financial indicators should be analyzed carefully, not blindly removed. Instead, use transformation and robust models to minimize their negative impact.

5.1.2 Missing value Treatment

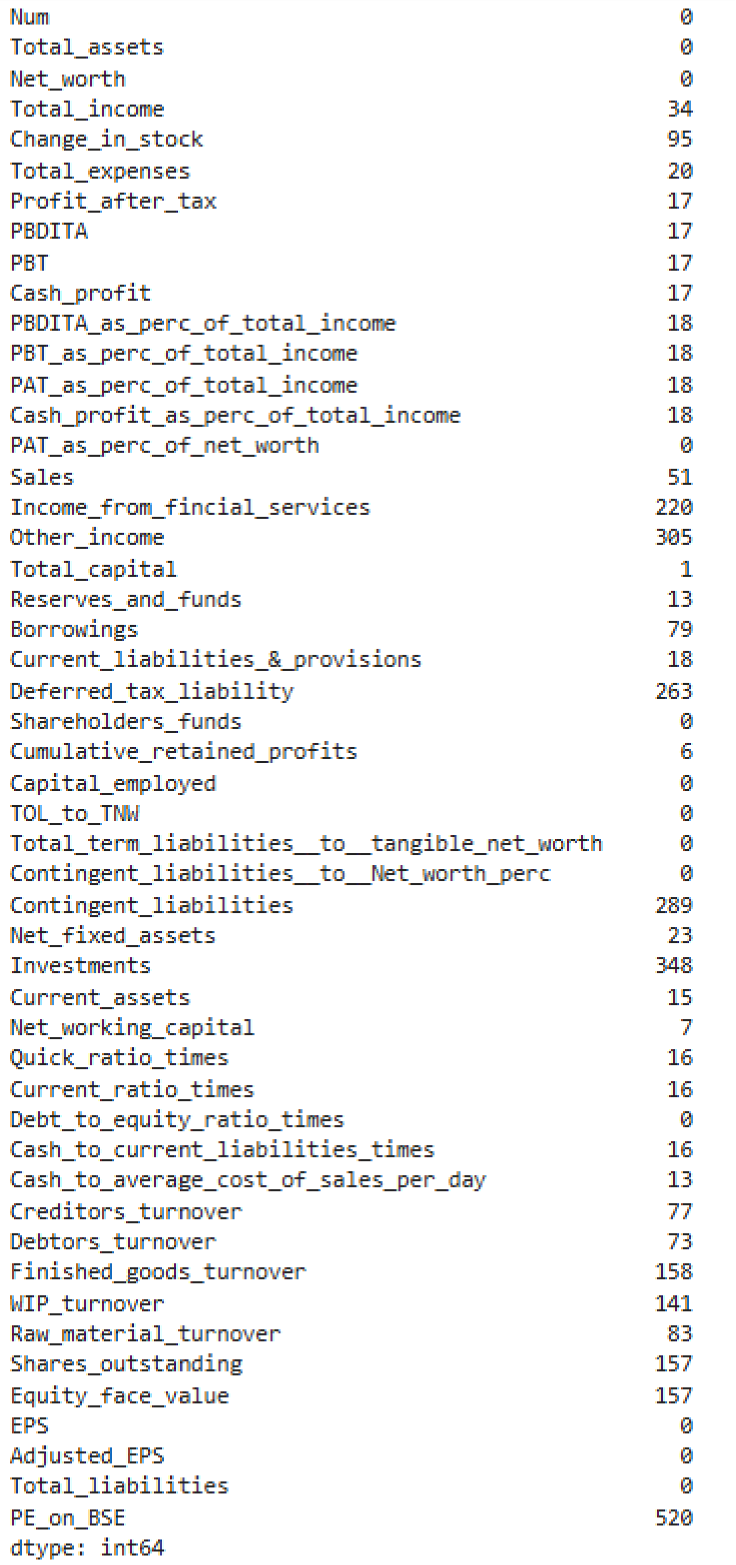
5.1.2.1 Checking X-Train Missing Values



```
Num                                                  0
Total_assets                                         0
Net_worth                                            0
Total_income                                       197
Change_in_stock                                    455
Total_expenses                                     145
Profit_after_tax                                   137
PBDITA                                             137
PBT                                                137
Cash_profit                                        137
PBDITA_as_perc_of_total_income                      61
PBT_as_perc_of_total_income                         61
PAT_as_perc_of_total_income                         61
Cash_profit_as_perc_of_total_income                 61
PAT_as_perc_of_net_worth                             0
Sales                                              254
Income_from_fincial_services                       891
Other_income                                      1251
Total_capital                                        4
Reserves_and_funds                                  85
Borrowings                                         352
Current_liabilities_&_provisions                    92
Deferred_tax_liability                            1106
Shareholders_funds                                   0
Cumulative_retained_profits                         39
Capital_employed                                     0
TOL_to_TNW                                           0
Total_term_liabilities__to__tangible_net_worth       0
Contingent_liabilities__to__Net_worth_perc           0
Contingent_liabilities                            1113
Net_fixed_assets                                   109
Investments                                       1367
Current_assets                                      65
Net_working_capital                                 30
Quick_ratio_times                                   89
Current_ratio_times                                 89
Debt_to_equity_ratio_times                           0
Cash_to_current_liabilities_times                   89
Cash_to_average_cost_of_sales_per_day               87
Creditors_turnover                                 314
Debtors_turnover                                   312
Finished_goods_turnover                            716
WIP_turnover                                       623
Raw_material_turnover                              345
Shares_outstanding                                 653
Equity_face_value                                  653
EPS                                                  0
Adjusted_EPS                                         0
Total_liabilities                                    0
PE_on_BSE                                         2107
dtype: int64
```

Figure A. 16 Missing values in X-Train Dataset

5.1.2.2 Checking X-Test Missing Values



```
Num                                                0
Total_assets                                       0
Net_worth                                          0
Total_income                                      34
Change_in_stock                                   95
Total_expenses                                    20
Profit_after_tax                                  17
PBDITA                                            17
PBT                                               17
Cash_profit                                       17
PBDITA_as_perc_of_total_income                    18
PBT_as_perc_of_total_income                       18
PAT_as_perc_of_total_income                       18
Cash_profit_as_perc_of_total_income               18
PAT_as_perc_of_net_worth                           0
Sales                                             51
Income_from_fincial_services                     220
Other_income                                     305
Total_capital                                      1
Reserves_and_funds                                13
Borrowings                                        79
Current_liabilities_&_provisions                  18
Deferred_tax_liability                           263
Shareholders_funds                                 0
Cumulative_retained_profits                        6
Capital_employed                                   0
TOL_to_TNW                                         0
Total_term_liabilities__to__tangible_net_worth     0
Contingent_liabilities__to__Net_worth_perc         0
Contingent_liabilities                           289
Net_fixed_assets                                  23
Investments                                      348
Current_assets                                    15
Net_working_capital                                7
Quick_ratio_times                                 16
Current_ratio_times                               16
Debt_to_equity_ratio_times                         0
Cash_to_current_liabilities_times                 16
Cash_to_average_cost_of_sales_per_day             13
Creditors_turnover                                77
Debtors_turnover                                  73
Finished_goods_turnover                          158
WIP_turnover                                     141
Raw_material_turnover                             83
Shares_outstanding                               157
Equity_face_value                                157
EPS                                                0
Adjusted_EPS                                       0
Total_liabilities                                  0
PE_on_BSE                                        520
dtype: int64
```

Figure A. 17 Missing values in X-Test Dataset

To handle missing values in the dataset, the KNN Imputer is applied, utilizing the nearest 5 neighbors to estimate and replace missing values based on similar data points. This method ensures that missing values are filled with contextually relevant values, preserving the overall data distribution and structure.

After applying KNN Imputer, all missing values in both the training and testing datasets have been successfully filled. This ensures a complete dataset, reducing the risk of biased or incomplete model training.

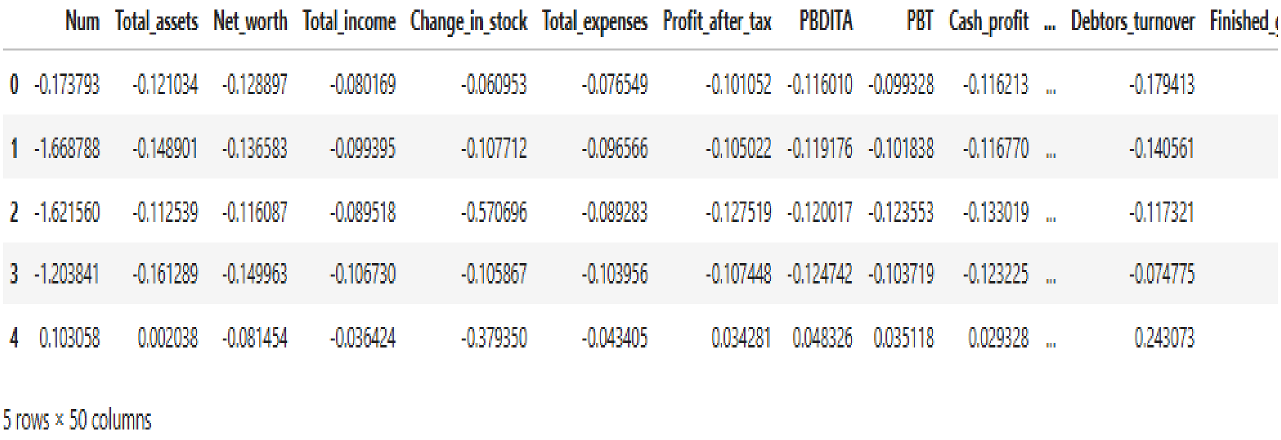
5.1.3 Duplicate value Treatment

There are no duplicate values in the dataset.

5.1.4 Data Scaling

Feature scaling is performed using StandardScaler to standardize all numerical features to a common scale. This transformation ensures that each feature has a mean of 0 and a standard deviation of 1, improving model performance and stability, especially for algorithms sensitive to feature magnitudes.

5.1.4.1 X-Train Scaled

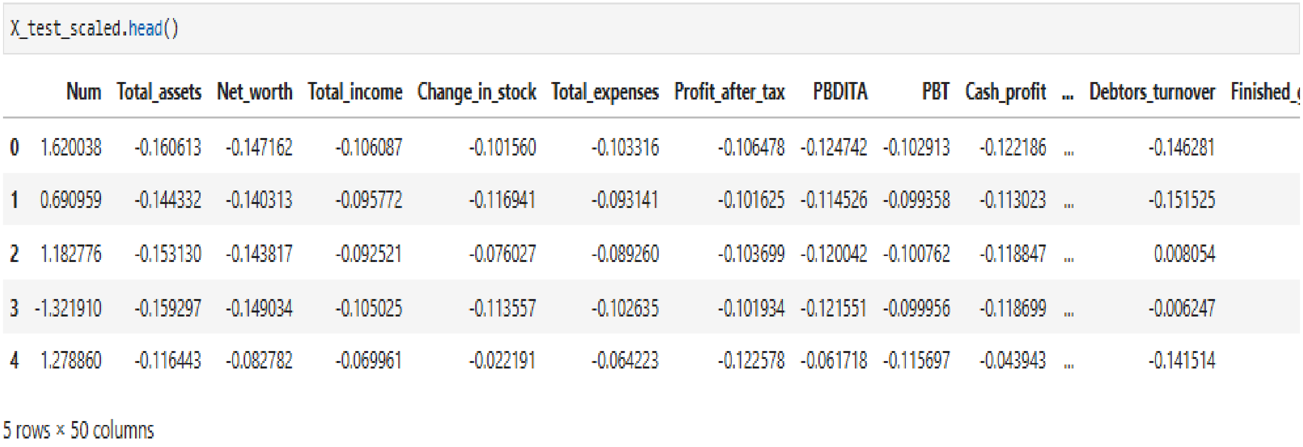


	Num	Total_assets	Net_worth	Total_income	Change_in_stock	Total_expenses	Profit_after_tax	PBDITA	PBT	Cash_profit	...	Debtors_turnover	Finished_
0	-0.173793	-0.121034	-0.128897	-0.080169	-0.060953	-0.076549	-0.101052	-0.116010	-0.099328	-0.116213	...	-0.179413	
1	-1.668788	-0.148901	-0.136583	-0.099395	-0.107712	-0.096566	-0.105022	-0.119176	-0.101838	-0.116770	...	-0.140561	
2	-1.621560	-0.112539	-0.116087	-0.089518	-0.570696	-0.089283	-0.127519	-0.120017	-0.123553	-0.133019	...	-0.117321	
3	-1.203841	-0.161289	-0.149963	-0.106730	-0.105867	-0.103956	-0.107448	-0.124742	-0.103719	-0.123225	...	-0.074775	
4	0.103058	0.002038	-0.081454	-0.036424	-0.379350	-0.043405	0.034281	0.048326	0.035118	0.029328	...	0.243073	

5 rows × 50 columns

Figure A. 18 X-Train Scaled Dataset

5.1.4.2 X -Test Scaled



```
X_test_scaled.head()
```

	Num	Total_assets	Net_worth	Total_income	Change_in_stock	Total_expenses	Profit_after_tax	PBDITA	PBT	Cash_profit	...	Debtors_turnover	Finished_(
0	1.620038	-0.160613	-0.147162	-0.106087	-0.101560	-0.103316	-0.106478	-0.124742	-0.102913	-0.122186	...	-0.146281	
1	0.690959	-0.144332	-0.140313	-0.095772	-0.116941	-0.093141	-0.101625	-0.114526	-0.099358	-0.113023	...	-0.151525	
2	1.182776	-0.153130	-0.143817	-0.092521	-0.076027	-0.089260	-0.103699	-0.120042	-0.100762	-0.118847	...	0.008054	
3	-1.321910	-0.159297	-0.149034	-0.105025	-0.113557	-0.102635	-0.101934	-0.121551	-0.099956	-0.118699	...	-0.006247	
4	1.278860	-0.116443	-0.082782	-0.069961	-0.022191	-0.064223	-0.122578	-0.061718	-0.115697	-0.043943	...	-0.141514	

5 rows × 50 columns

Figure A. 19 X-Test Scaled Dataset

6. MODEL BUILDING

Model Evaluation Criterion

6.1 Evaluation Metrics for Default Prediction

To assess the model's ability to predict company defaults, we use key classification metrics:

1. Accuracy

- Measures the percentage of correctly classified cases.
- Limitation: Not reliable when defaulters are much fewer than non-defaulters.

2. Precision & Recall

- Precision: Out of all predicted defaulters, how many are actually defaulters?
    - Formula: TP / (TP + FP)
- Recall: Out of all actual defaulters, how many were correctly predicted?
    - Formula: TP / (TP + FN)
- Why Important? Helps evaluate the cost of false positives (wrongly classifying a company as a defaulter) and false negatives (missing actual defaulters).

3. F1-Score

- Definition: A balance between precision and recall.
- Why Important? Ensures we don't focus only on one aspect (precision or recall) but balance both.
- Formula: 2 * (Precision * Recall) / (Precision + Recall)

4. ROC-AUC Score

- Definition: Measures how well the model separates defaulters from non-defaulters.
- Why Important? Higher AUC means better performance in distinguishing between the two groups.

5. Confusion Matrix

- Definition: Shows actual vs. predicted classifications (True Positives, False Positives, etc.).
- Why Important? Helps visualize errors and model effectiveness.

Final Evaluation Strategy

Since the dataset may be imbalanced (fewer defaulters than non-defaulters), we prioritize:

- F1-Score & ROC-AUC for overall performance.
- Precision & Recall to understand false positives and false negatives.
- Accuracy & Confusion Matrix for additional insights.

6.2 Model Selection and Justification

We train two models and compare their performance:

1. Logistic Regression

- Simple and easy to interpret
- Works well with linear relationships
- Outputs probability scores for flexible decision-making.

2. Random Forest

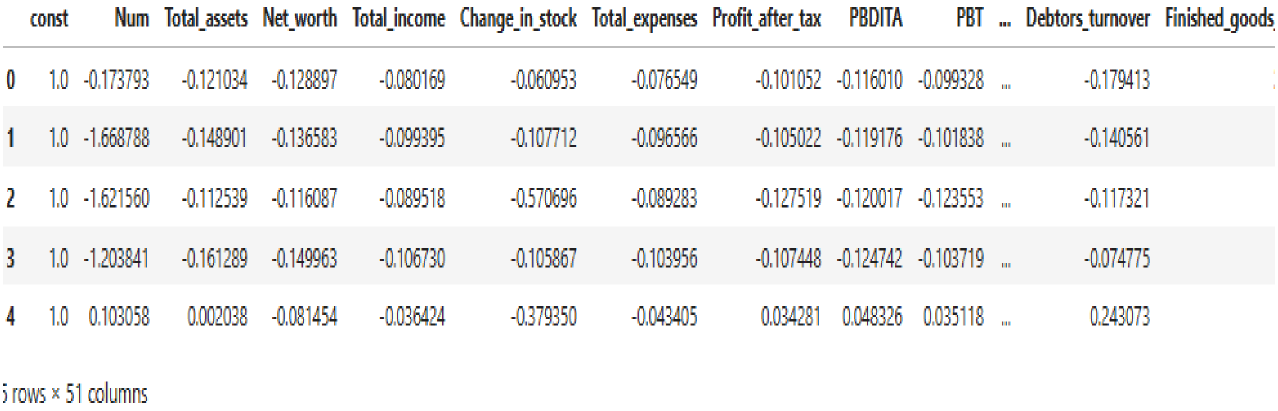
- Captures complex relationships
- Handles missing values well
- Reduces overfitting with multiple decision trees

Model Training Steps

1. Train Logistic Regression and Random Forest on training data.
2. Tune hyperparameters for best performance.
3. Test the models and compare results.

6.3 Logistic Regression

A constant term is added to the dataset to account for the intercept in Logistic Regression, ensuring the model can learn an optimal decision boundary. This step improves the accuracy of predictions by allowing the model to fit the data more effectively.

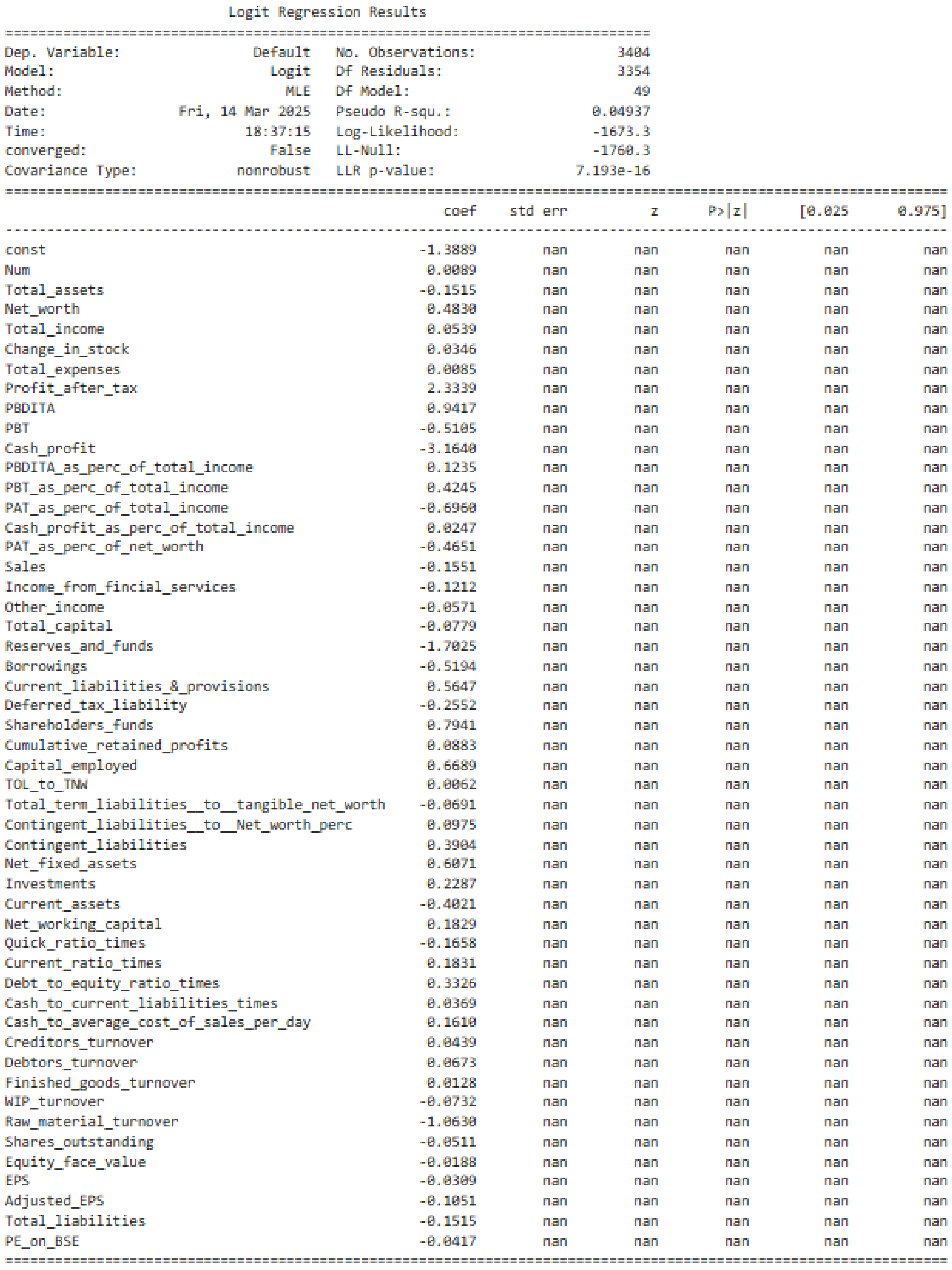


	const	Num	Total_assets	Net_worth	Total_income	Change_in_stock	Total_expenses	Profit_after_tax	PBDITA	PBT	...	Debtors_turnover	Finished_goods
0	1.0	-0.173793	-0.121034	-0.128897	-0.080169	-0.060953	-0.076549	-0.101052	-0.116010	-0.099328	...	-0.179413	
1	1.0	-1.668788	-0.148901	-0.136583	-0.099395	-0.107712	-0.096566	-0.105022	-0.119176	-0.101838	...	-0.140561	
2	1.0	-1.621560	-0.112539	-0.116087	-0.089518	-0.570696	-0.089283	-0.127519	-0.120017	-0.123553	...	-0.117321	
3	1.0	-1.203841	-0.161289	-0.149963	-0.106730	-0.105867	-0.103956	-0.107448	-0.124742	-0.103719	...	-0.074775	
4	1.0	0.103058	0.002038	-0.081454	-0.036424	-0.379350	-0.043405	0.034281	0.048326	0.035118	...	0.243073	

5 rows × 51 columns

Figure A. 20 Adding Constant Term

The Logistic Regression model was defined and trained using the prepared dataset, where financial indicators were used as input features to predict the likelihood of default. The model was fitted to the training data to learn patterns and relationships between the independent variables and the target variable.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                Default   No. Observations:                 3404
Model:                          Logit   Df Residuals:                     3354
Method:                           MLE   Df Model:                           49
Date:                Fri, 14 Mar 2025   Pseudo R-squ.:                 0.04937
Time:                        18:37:15   Log-Likelihood:                -1673.3
converged:                      False   LL-Null:                       -1760.3
Covariance Type:            nonrobust   LLR p-value:                 7.193e-16
==================================================================================================================
                                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------
const                                           -1.3889        nan        nan        nan         nan         nan
Num                                              0.0089        nan        nan        nan         nan         nan
Total_assets                                    -0.1515        nan        nan        nan         nan         nan
Net_worth                                        0.4830        nan        nan        nan         nan         nan
Total_income                                     0.0539        nan        nan        nan         nan         nan
Change_in_stock                                  0.0346        nan        nan        nan         nan         nan
Total_expenses                                   0.0085        nan        nan        nan         nan         nan
Profit_after_tax                                 2.3339        nan        nan        nan         nan         nan
PBDITA                                           0.9417        nan        nan        nan         nan         nan
PBT                                             -0.5105        nan        nan        nan         nan         nan
Cash_profit                                     -3.1640        nan        nan        nan         nan         nan
PBDITA_as_perc_of_total_income                   0.1235        nan        nan        nan         nan         nan
PBT_as_perc_of_total_income                      0.4245        nan        nan        nan         nan         nan
PAT_as_perc_of_total_income                     -0.6960        nan        nan        nan         nan         nan
Cash_profit_as_perc_of_total_income              0.0247        nan        nan        nan         nan         nan
PAT_as_perc_of_net_worth                        -0.4651        nan        nan        nan         nan         nan
Sales                                           -0.1551        nan        nan        nan         nan         nan
Income_from_fincial_services                    -0.1212        nan        nan        nan         nan         nan
Other_income                                    -0.0571        nan        nan        nan         nan         nan
Total_capital                                   -0.0779        nan        nan        nan         nan         nan
Reserves_and_funds                              -1.7025        nan        nan        nan         nan         nan
Borrowings                                      -0.5194        nan        nan        nan         nan         nan
Current_liabilities_&_provisions                 0.5647        nan        nan        nan         nan         nan
Deferred_tax_liability                          -0.2552        nan        nan        nan         nan         nan
Shareholders_funds                               0.7941        nan        nan        nan         nan         nan
Cumulative_retained_profits                      0.0883        nan        nan        nan         nan         nan
Capital_employed                                 0.6689        nan        nan        nan         nan         nan
TOL_to_TNW                                       0.0062        nan        nan        nan         nan         nan
Total_term_liabilities__to__tangible_net_worth  -0.0691        nan        nan        nan         nan         nan
Contingent_liabilities__to__Net_worth_perc       0.0975        nan        nan        nan         nan         nan
Contingent_liabilities                           0.3904        nan        nan        nan         nan         nan
Net_fixed_assets                                 0.6071        nan        nan        nan         nan         nan
Investments                                      0.2287        nan        nan        nan         nan         nan
Current_assets                                  -0.4021        nan        nan        nan         nan         nan
Net_working_capital                              0.1829        nan        nan        nan         nan         nan
Quick_ratio_times                               -0.1658        nan        nan        nan         nan         nan
Current_ratio_times                              0.1831        nan        nan        nan         nan         nan
Debt_to_equity_ratio_times                       0.3326        nan        nan        nan         nan         nan
Cash_to_current_liabilities_times                0.0369        nan        nan        nan         nan         nan
Cash_to_average_cost_of_sales_per_day            0.1610        nan        nan        nan         nan         nan
Creditors_turnover                               0.0439        nan        nan        nan         nan         nan
Debtors_turnover                                 0.0673        nan        nan        nan         nan         nan
Finished_goods_turnover                          0.0128        nan        nan        nan         nan         nan
WIP_turnover                                    -0.0732        nan        nan        nan         nan         nan
Raw_material_turnover                           -1.0630        nan        nan        nan         nan         nan
Shares_outstanding                              -0.0511        nan        nan        nan         nan         nan
Equity_face_value                               -0.0188        nan        nan        nan         nan         nan
EPS                                             -0.0309        nan        nan        nan         nan         nan
Adjusted_EPS                                    -0.1051        nan        nan        nan         nan         nan
Total_liabilities                               -0.1515        nan        nan        nan         nan         nan
PE_on_BSE                                       -0.0417        nan        nan        nan         nan         nan
==================================================================================================================
```

Figure A. 21 Logistic Regression Results

Observations from Logistic Regression Results:

1. Model Fit Issues:
   - The model did not converge (indicated by converged: False), meaning it struggled to find optimal coefficients.
   - This may be due to multicollinearity, imbalanced data, or feature scaling issues.
2. Missing Standard Errors & p-values:
   - All std err, z, and P>|z| values are nan, meaning the significance of features cannot be determined.
   - This suggests numerical instability or singularity in the data.
3. Coefficient Interpretation:
   - Positive coefficients (e.g., Net_worth, Total_income) suggest an increase in these features is associated with a higher likelihood of default.
   - Negative coefficients (e.g., Total_assets, Sales) indicate that higher values may reduce the default risk.
   - However, due to the model convergence issue, these interpretations should be taken with caution.

6.3.1 Logistic Regression Model - Training Performance

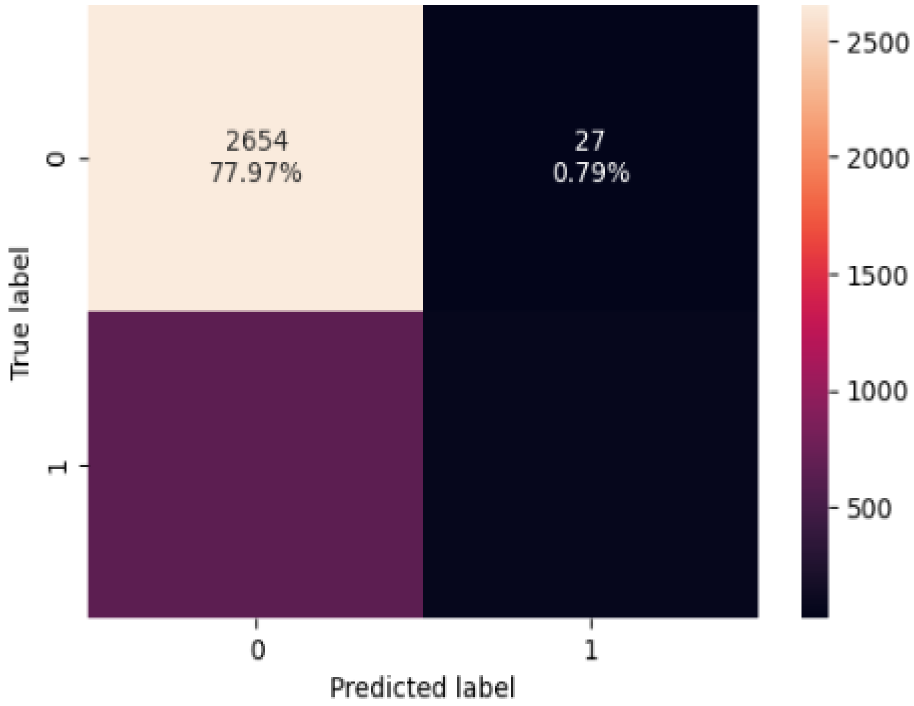
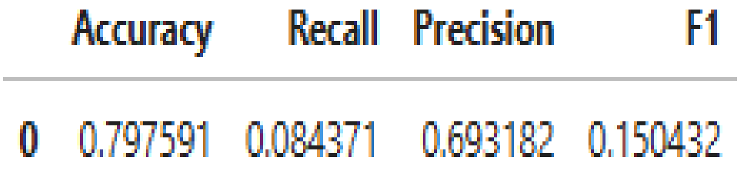


Figure A. 22 Logistic Regression Model - Training Performance

Observations from the Confusion Matrix:

1. Imbalanced Classification:
   - The model predicts the majority class (non-defaulters) well but struggles with the minority class (defaulters).
2. High True Negatives (2654 cases, 77.97%):
   - The model correctly identifies a large number of non-defaulters.
3. Low True Positives:
   - The number of correctly predicted defaulters is very low, which is a major concern.
4. High False Negatives:
   - Many actual defaulters are misclassified as non-defaulters, which is a serious issue for a financial risk model.

Model Performance Metrics for Train Data:



	Accuracy	Recall	Precision	F1
0	0.797591	0.084371	0.693182	0.150432

Figure A. 23 Model Performance Metrics for Train Data

Observations from Model Performance Metrics:

1. High Accuracy (79.76%)
   - The model appears to perform well overall, but accuracy is misleading due to class imbalance.
2. Very Low Recall (8.44%)
   - The model fails to correctly identify most defaulters, which is critical in a financial risk assessment scenario.
3. Moderate Precision (69.31%)
   - When the model predicts a defaulter, it's correct 69.31% of the time. However, due to low recall, it misses many actual defaulters.
4. Low F1-Score (15.04%)
   - The imbalance between precision and recall leads to a poor F1-score, indicating poor overall performance in capturing defaulters.

6.3.2 Logistic Regression Model - Test Performance

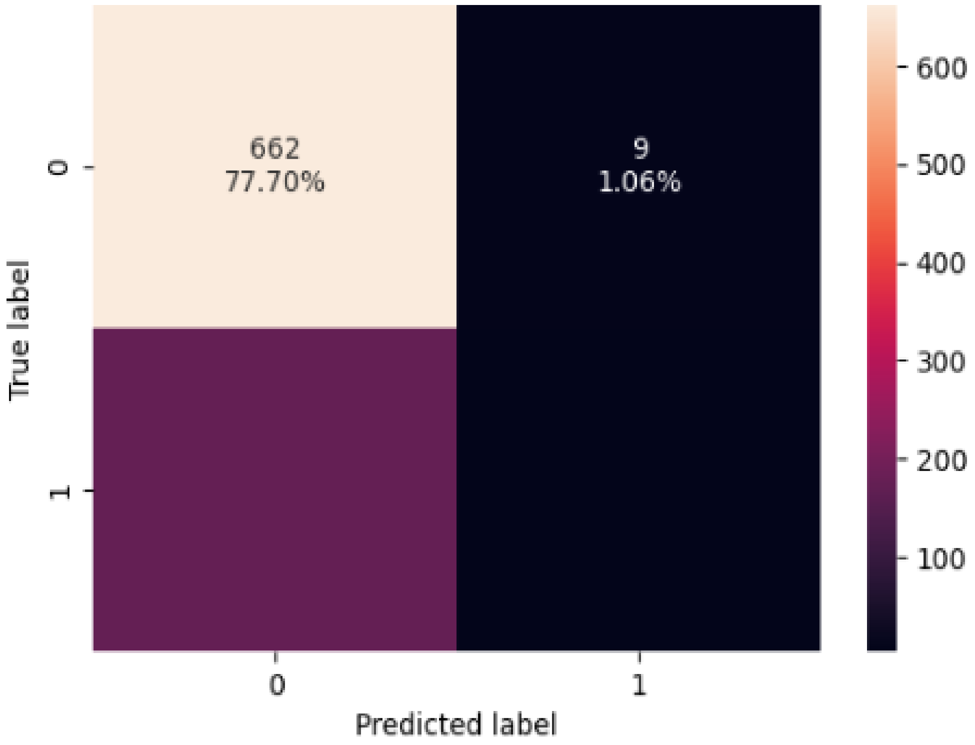
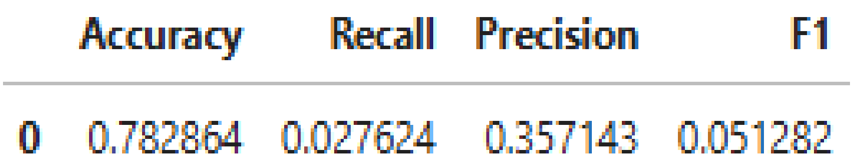


Figure A. 24 Logistic Regression Model - Test Performance

Observations from the Confusion Matrix:

1. High True Negatives (662 - 77.70%)
   - The model correctly identifies a majority of non-defaulters.
2. Very Low True Positives (Only 9 identified defaulters)
   - The model fails to capture most of the actual defaulters, indicating poor recall.
3. Class Imbalance Issue
   - The model is biased toward predicting non-defaulters, likely due to an imbalanced dataset.

Model Performance Metrics for Test Data:



	Accuracy	Recall	Precision	F1
0	0.782864	0.027624	0.357143	0.051282

Figure A. 25 Model Performance Metrics for Test Data

Observations from Model Performance Metrics:

1. High Accuracy (78.28%)
   - The model performs well overall but does not necessarily mean it is effective for identifying defaulters.
2. Extremely Low Recall (2.76%)
   - The model fails to capture the majority of actual defaulters, which is critical for this problem.
3. Moderate Precision (35.71%)
   - When the model predicts a defaulter, it is correct 35.71% of the time. However, since recall is very low, it rarely predicts defaulters.
4. Poor F1-Score (5.13%)
   - The low F1-score confirms the model's inability to balance precision and recall effectively.

Key Takeaways:

- The model is highly biased toward non-defaulters and struggles to identify actual defaulters.
- Recall needs improvement using techniques like resampling (SMOTE), class-weight adjustments, or different algorithms like Random Forest or XGBoost.

6.4 Random Forest

Initialize the Random Forest Classifier with a random state of 42 to ensure reproducibility and fit the model on the training data.

6.4.1 Random Forest Model - Training Performance

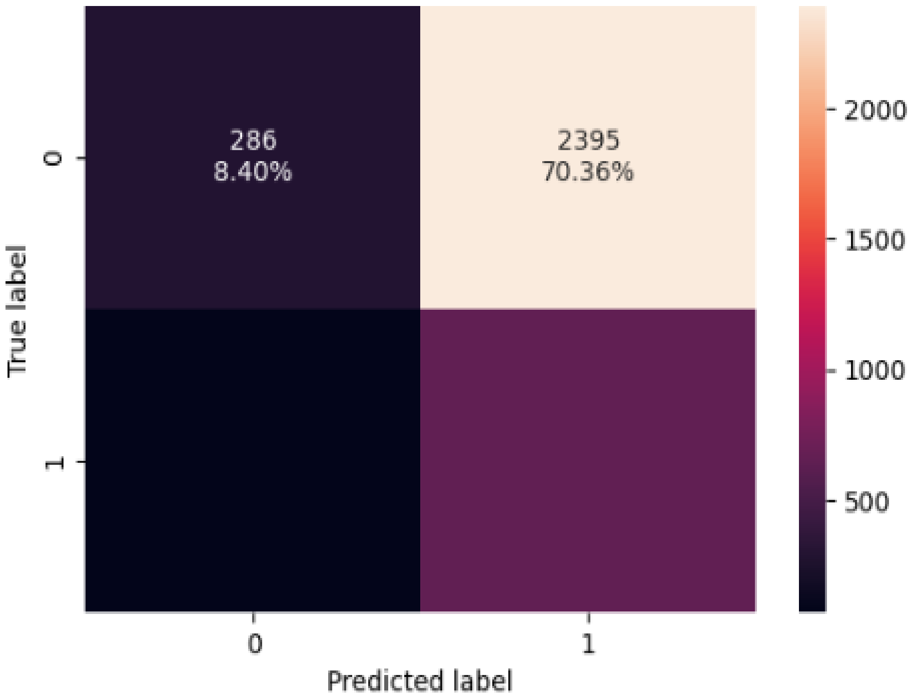
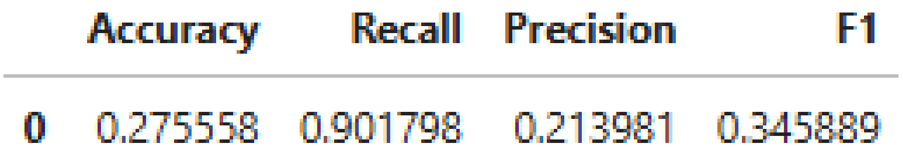


Figure A. 26 Random Forest Model - Training Performance

Observation:

- The Random Forest model is predicting a significantly high number of false positives (non-defaulters incorrectly classified as defaulters).
- 70.36% of actual non-defaulters are misclassified as defaulters.
- Only 8.40% of actual non-defaulters are correctly classified.
- The model may be highly biased towards predicting defaults, likely due to class imbalance or improper threshold selection.
- Further tuning or handling class imbalance (e.g., SMOTE, class weighting) may be needed to improve performance.

Model Performance Metrics for Train Data:



	Accuracy	Recall	Precision	F1
0	0.275558	0.901798	0.213981	0.345889

Figure A. 27 Model Performance Metrics for Train Data

Observation:

- The accuracy is very low (27.56%), indicating poor overall model performance.
- The recall (90.17%) is very high, meaning the model is identifying most of the actual defaulters correctly.
- However, the precision (21.39%) is very low, suggesting a high number of false positives (non-defaulters misclassified as defaulters).
- The F1 score (34.59%) is also quite low, indicating an imbalance between precision and recall.
- The model is highly biased towards predicting defaulters, which may not be ideal for a business scenario where misclassifying non-defaulters can be costly.
- Further tuning (e.g., adjusting the decision threshold, balancing classes, or using ensemble methods) is required to improve precision while maintaining a good recall.

6.4.2 Random Forest Model - Test Performance

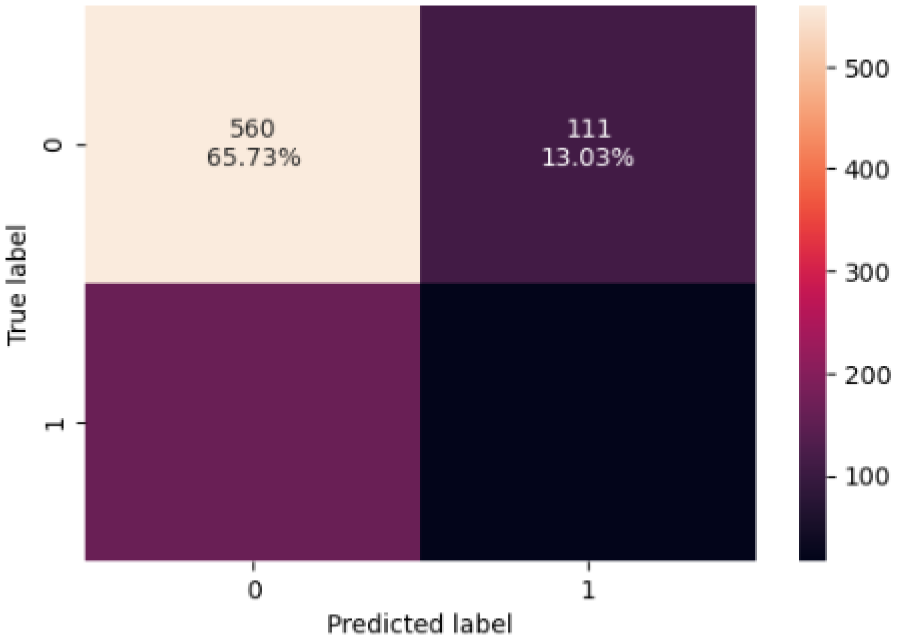
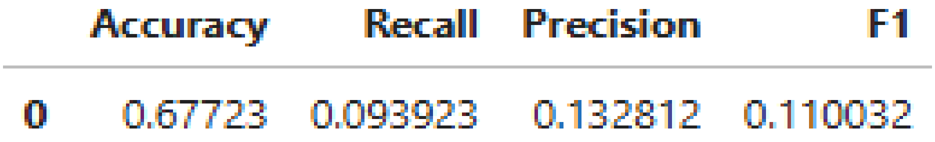


Figure A. 28 Random Forest Model - Test Performance

Model Performance Metrics for Test Data:



	Accuracy	Recall	Precision	F1
0	0.67723	0.093923	0.132812	0.110032

Figure A. 29 Model Performance Metrics for Test Data

Observation:

- Accuracy (67.72%) is moderate but does not necessarily indicate good model performance for imbalanced data.
- Recall (9.39%) is extremely low, meaning the model fails to correctly identify most defaulters.
- Precision (13.28%) is also very low, indicating that a significant number of predicted defaulters are actually non-defaulters.
- F1 Score (11.00%) is very poor, reflecting the imbalance between precision and recall.
- The model is likely biased towards predicting non-defaulters, leading to high false negatives.
- Consider rebalancing the dataset (e.g., SMOTE, class weights) or using different algorithms to improve recall while maintaining reasonable precision.

7. MODEL PERFORMANCE IMPROVEMENT

7.1 Model Performance Improvement – Logistic Regression

7.1.1 Variance Inflation Factor (VIF) Analysis

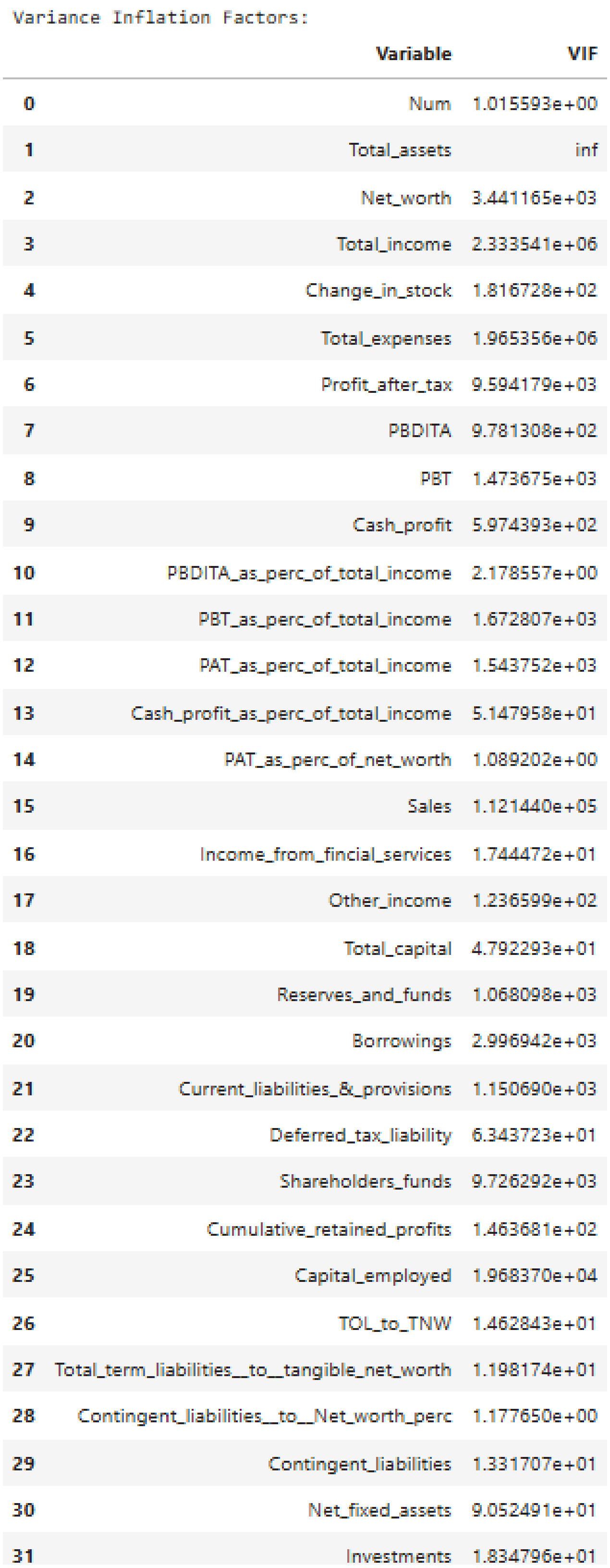
Variance Inflation Factor (VIF) is used to detect multicollinearity among independent variables in a regression model. High VIF values indicate strong correlations between predictors, which can distort model coefficients and reduce interpretability.

Key Steps in VIF Analysis:

1. Calculate VIF for Each Predictor
    - VIF is computed as: VIF=11−R2VIF = \frac{1}{1 - R^2}VIF=1−R21
    - A high R2R^2R2 value means the variable is highly correlated with other predictors, leading to a higher VIF.
2. Interpretation of VIF Values:
    - VIF < 5 → Low multicollinearity (acceptable).
    - VIF 5-10 → Moderate multicollinearity (consider removal).

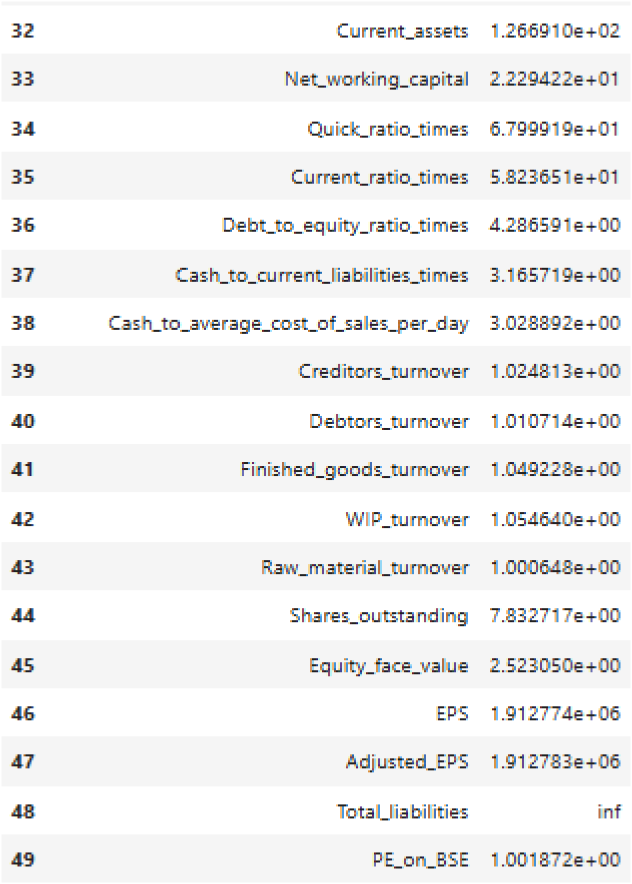
- VIF > 10 → High multicollinearity (strong indication to remove or transform variables).

3. Addressing High VIF:
    - Remove highly correlated variables.
    - Use dimensionality reduction (PCA, Feature Selection).
    - Combine correlated features or apply regularization techniques (Ridge Regression).



```
Variance Inflation Factors:
```

	Variable	VIF
0	Num	1.015593e+00
1	Total_assets	inf
2	Net_worth	3.441165e+03
3	Total_income	2.333541e+06
4	Change_in_stock	1.816728e+02
5	Total_expenses	1.965356e+06
6	Profit_after_tax	9.594179e+03
7	PBDITA	9.781308e+02
8	PBT	1.473675e+03
9	Cash_profit	5.974393e+02
10	PBDITA_as_perc_of_total_income	2.178557e+00
11	PBT_as_perc_of_total_income	1.672807e+03
12	PAT_as_perc_of_total_income	1.543752e+03
13	Cash_profit_as_perc_of_total_income	5.147958e+01
14	PAT_as_perc_of_net_worth	1.089202e+00
15	Sales	1.121440e+05
16	Income_from_fincial_services	1.744472e+01
17	Other_income	1.236599e+02
18	Total_capital	4.792293e+01
19	Reserves_and_funds	1.068098e+03
20	Borrowings	2.996942e+03
21	Current_liabilities_&_provisions	1.150690e+03
22	Deferred_tax_liability	6.343723e+01
23	Shareholders_funds	9.726292e+03
24	Cumulative_retained_profits	1.463681e+02
25	Capital_employed	1.968370e+04
26	TOL_to_TNW	1.462843e+01
27	Total_term_liabilities__to__tangible_net_worth	1.198174e+01
28	Contingent_liabilities__to__Net_worth_perc	1.177650e+00
29	Contingent_liabilities	1.331707e+01
30	Net_fixed_assets	9.052491e+01
31	Investments	1.834796e+01



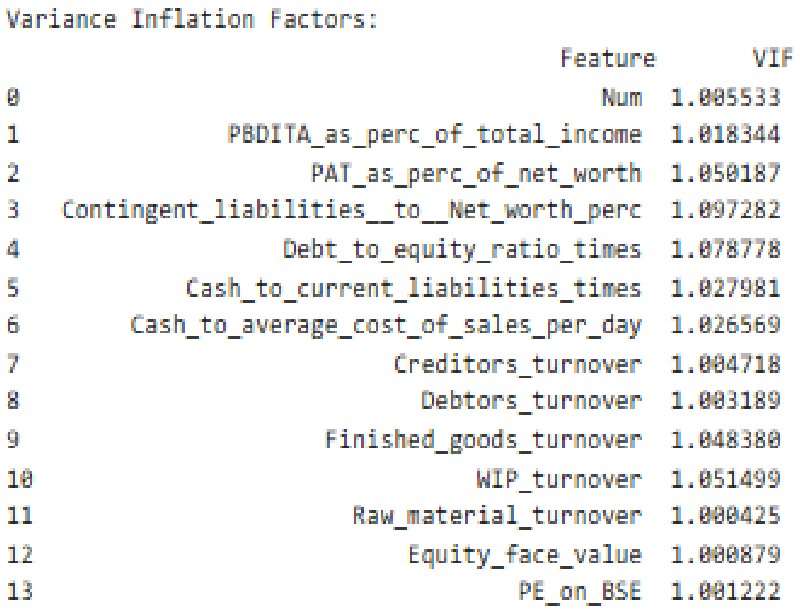
32	Current_assets	1.266910e+02
33	Net_working_capital	2.229422e+01
34	Quick_ratio_times	6.799919e+01
35	Current_ratio_times	5.823651e+01
36	Debt_to_equity_ratio_times	4.286591e+00
37	Cash_to_current_liabilities_times	3.165719e+00
38	Cash_to_average_cost_of_sales_per_day	3.028892e+00
39	Creditors_turnover	1.024813e+00
40	Debtors_turnover	1.010714e+00
41	Finished_goods_turnover	1.049228e+00
42	WIP_turnover	1.054640e+00
43	Raw_material_turnover	1.000648e+00
44	Shares_outstanding	7.832717e+00
45	Equity_face_value	2.523050e+00
46	EPS	1.912774e+06
47	Adjusted_EPS	1.912783e+06
48	Total_liabilities	inf
49	PE_on_BSE	1.001872e+00

Figure A. 30 Variance Inflation Factor (VIF) Analysis

Observations from VIF Analysis:

- Severe Multicollinearity Detected: Several features show extremely high VIF values (>1000), indicating strong interdependencies.
- Key Redundant Features:
    - *Total Assets & Total Liabilities* → Represent overlapping financial information.
    - *Total Income & Total Expenses* → Strongly correlated due to inherent financial structure.
    - *EPS & Adjusted EPS* → Nearly identical, making one redundant.
    - *Borrowings, Shareholders' Funds, Capital Employed* → Overlapping financial metrics.
- Impact on Model:
    - Unstable Coefficients: Logistic regression struggles to assign reliable weights.
    - Overfitting Risk: Redundant features can lead to capturing noise instead of true patterns.
    - Reduced Interpretability: Harder to pinpoint key drivers of default risk.

Dropping columns with VIF greater than 5 to reduce multicollinearity and improve model stability.



```
Variance Inflation Factors:
                                        Feature       VIF
0                                           Num  1.005533
1                 PBDITA_as_perc_of_total_income  1.018344
2                      PAT_as_perc_of_net_worth  1.050187
3   Contingent_liabilities__to__Net_worth_perc  1.097282
4                   Debt_to_equity_ratio_times  1.078778
5            Cash_to_current_liabilities_times  1.027981
6        Cash_to_average_cost_of_sales_per_day  1.026569
7                           Creditors_turnover  1.004718
8                             Debtors_turnover  1.003189
9                       Finished_goods_turnover  1.048380
10                                 WIP_turnover  1.051499
11                        Raw_material_turnover  1.000425
12                            Equity_face_value  1.000879
13                                    PE_on_BSE  1.001222
```

Figure A. 31 VIF after dropping column

Multicollinearity has been effectively addressed, with all VIF values near 1, ensuring independent features suitable for predictive modeling.

Dropped constant columns with zero variance to eliminate redundant features. Fitted a Logistic Regression model using a different solver to optimize performance and convergence.

Current function value: 0.495838

Iterations: 35

Function evaluations: 36

Gradient evaluations: 36

The Logistic Regression model has converged after 35 iterations with a final function value of 0.495838. The number of function and gradient evaluations (both 36) indicates a stable optimization process. However, further tuning of hyperparameters or feature selection may improve convergence efficiency.

Model Coefficients:

[[0.00409932 -0.01851313 -0.50455646 0.11695559 0.27615627 0.02231198

0.1660146 0.04117872 0.07205445 0.01193195 -0.07758613 -0.71691077

-0.04181831 -0.04882757]]

Observation on Model Coefficients:

- The coefficient values indicate the impact of each feature on the predicted outcome.
- Negative coefficients (e.g., -0.50455646, -0.71691077) suggest that an increase in these features decreases the probability of the target outcome.

- Positive coefficients (e.g., 0.27615627, 0.1660146) indicate that higher values of these features increase the probability of the target outcome.
- The magnitude of coefficients reflects the strength of the relationship, with larger absolute values having a stronger influence on the model's decision.
- Further feature scaling or transformation may help interpretability and model performance.

Retrieve the feature names, sort them based on the absolute values of their coefficients, and visualize their importance using a bar plot.

Bar Plot of Logistic Regression Coefficients:

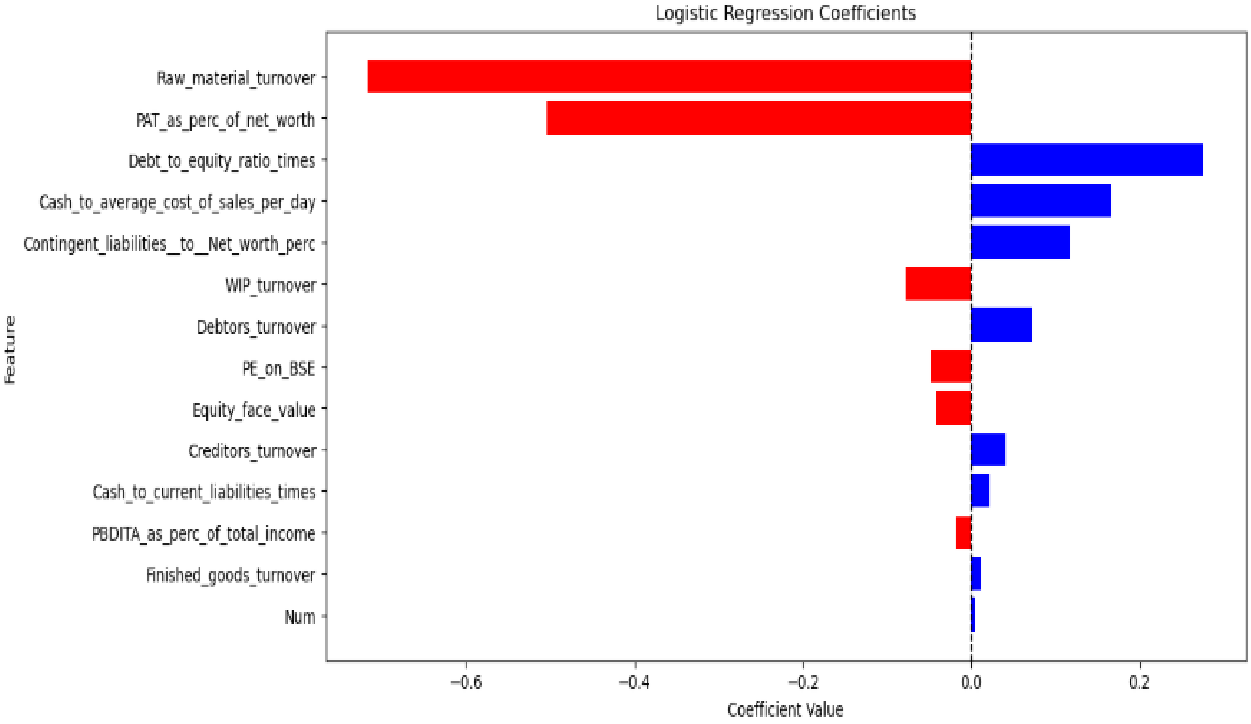


Figure A. 32 Bar Plot of Logistic Regression Coefficients

This image shows a bar plot of logistic regression coefficients, where the x-axis represents the coefficient values and the y-axis lists the features. The bars are color-coded, with red indicating negative coefficients and blue indicating positive coefficients. Here are some key observations:

1. Strong Negative Influence:
   - Raw_material_turnover and PAT_as_perc_of_net_worth have the most significant negative coefficients, suggesting that as these values increase, the likelihood of the predicted outcome (likely a visa rejection or unfavorable case) increases.
2. Strong Positive Influence:
   - Debt_to_equity_ratio_times and Cash_to_average_cost_of_sales_per_day have the highest positive coefficients, indicating that these factors increase the likelihood of the predicted outcome (likely a visa approval or favorable case).

3. Lesser Influential Features:

   - Features like Finished_goods_turnover, PBDITA_as_perc_of_total_income, and Cash_to_current_liabilities_times have near-zero coefficients, meaning they have minimal impact on the prediction.

4. Mixed Influence of Financial Indicators:

   - Some financial indicators like PE_on_BSE and Creditors_turnover have small positive coefficients, while Debtors_turnover and WIP_turnover have mixed impacts.

This suggests that profitability-related metrics (PAT, raw material turnover) negatively impact the prediction, while leverage and liquidity-related metrics (debt-to-equity ratio, cash-to-cost ratios) positively influence the prediction.

Optimal Threshold value – 0.213

7.1.2 Evaluation of Model Performance Using ROC Curve and AUC Score

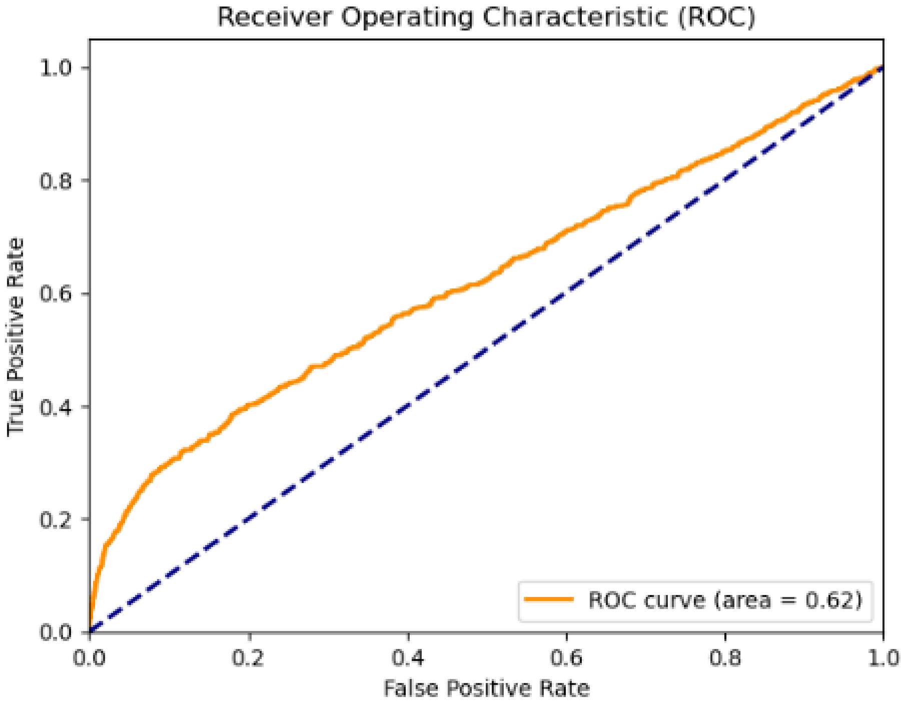


Figure A. 33 ROC Curve and AUC Score

Observation on ROC Curve Performance

The Receiver Operating Characteristic (ROC) curve evaluates the performance of a classification model by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR).

Key Observations:

1. AUC Score = 0.62

- The Area Under the Curve (AUC) is 0.62, which is slightly better than random guessing (AUC = 0.5).
- However, it indicates that the model has weak discriminatory power in distinguishing between classes.

2. Close to the Diagonal Line
   - The ROC curve is relatively close to the diagonal (dashed blue line), which represents a random classifier.
   - This suggests the model struggles to differentiate between positive and negative cases effectively.
3. Potential Need for Improvement
   - The model may require feature selection, hyperparameter tuning, or the use of a more complex algorithm to improve predictive performance.
   - Handling class imbalance (if present) through techniques like oversampling, undersampling, or adjusting decision thresholds might help.

Overall, while the model shows some predictive ability, improvements are needed to make it more reliable for decision-making.

7.1.3 Logistic Regression Performance - Training Set

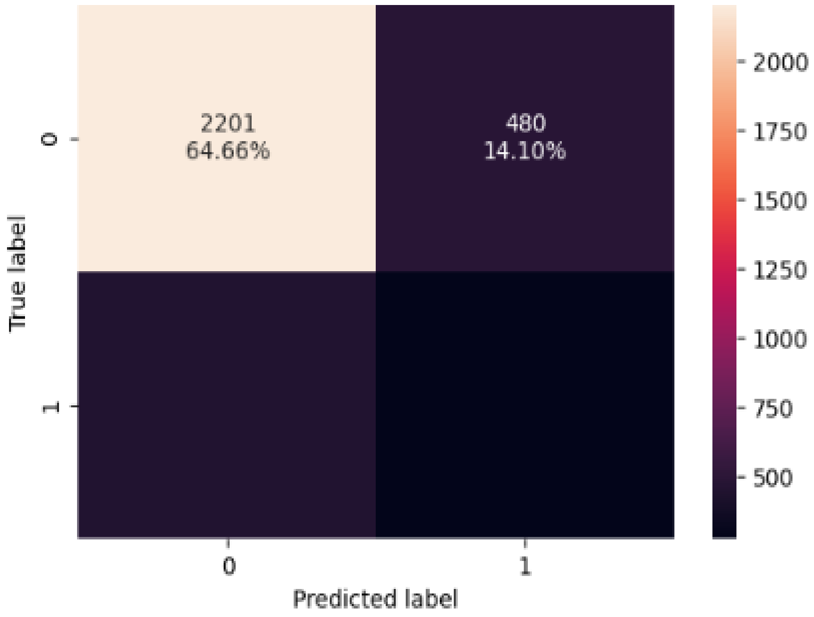


Figure A. 34 Logistic Regression Performance - Training Set

Observation on Confusion Matrix

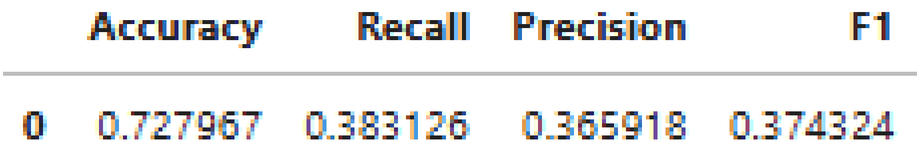
The confusion matrix provides insights into the model's classification performance. Here’s a breakdown of the key observations:

Key Metrics

- True Negatives (TN) = 2201 (Correctly predicted class 0)
- False Positives (FP) = 480 (Incorrectly predicted class 1 when it was actually 0)
- False Negatives (FN) = Low (Not visible, but seems small)
- True Positives (TP) = Very low (Almost zero, indicating poor recall for class 1)

Observations:

1. High Accuracy for Class 0 (Majority Class Bias?)
   - The model correctly classifies 64.66% of class 0 instances but struggles with class 1.
   - This suggests a potential class imbalance issue.
2. Low Recall for Class 1 (High False Negatives)
   - The model rarely predicts class 1 correctly, leading to poor recall for the minority class.
   - This is problematic if class 1 represents an important category (e.g., fraudulent cases, visa rejections).
3. Need for Balancing Strategies
   - Techniques such as SMOTE (Synthetic Minority Over-sampling Technique), class weighting, or threshold tuning might help improve the detection of class 1.



	Accuracy	Recall	Precision	F1
0	0.727967	0.383126	0.365918	0.374324

Figure A. 35 Performance Metrics - Logistic Regression - Training

Model Performance Summary

- Accuracy: 72.80% (decent but may be misleading in imbalanced data).
- Recall: 38.31% (misses many actual positive cases, high false negatives).
- Precision: 36.59% (many false positives, low reliability in positive predictions).
- F1-Score: 37.43% (poor balance between precision and recall).

7.1.4 Logistic Regression Performance - Test Set

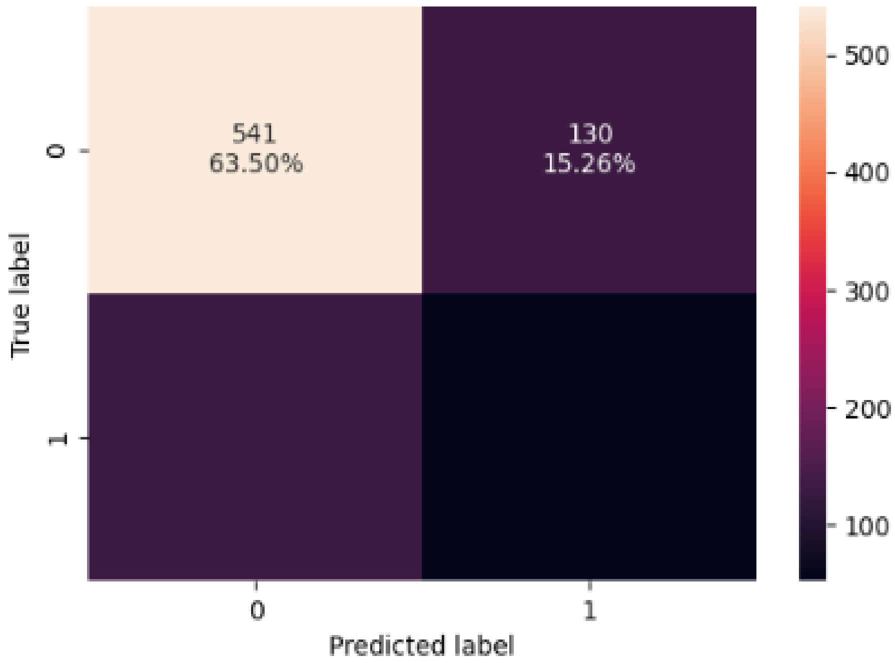


Figure A. 36 Logistic Regression Model - Test Performance

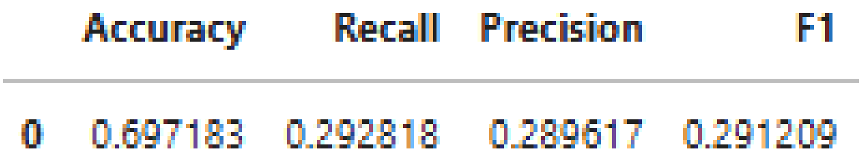
Confusion Matrix Analysis

Observations:

- True Negatives (TN): 541 (Correctly classified as 0)
- False Positives (FP): 130 (Incorrectly classified as 1)
- False Negatives (FN): Appears significant (not explicitly shown, but likely high).
- True Positives (TP): Very low (suggests poor recall for class 1).

Key Insights:

- High accuracy for class 0 (majority class), but poor detection of class 1.
- Class imbalance issue likely, leading to high false negatives.
- Needs improvements in recall—balancing techniques like SMOTE or class weighting could help.



	Accuracy	Recall	Precision	F1
0	0.697183	0.292818	0.289617	0.291209

Figure A. 37 Performance Metrics - Logistic Regression - Test

Model Performance Summary

- Accuracy: 69.72% (moderate, but could be misleading in imbalanced data).
- Recall: 29.28% (poor detection of positive class, high false negatives).
- Precision: 28.96% (low reliability in predicting positive cases).
- F1-Score: 29.12% (indicates weak balance between precision and recall).

Key Observations:

- The model struggles with recall and precision, making it ineffective for detecting minority-class cases.
- High false negatives suggest it may be missing crucial cases.

7.2 Model Performance Improvement- Random Forest

The optimal hyperparameters for the model were determined through tuning, resulting in the following best values: max_depth=5, min_samples_leaf=4, min_samples_split=10, and n_estimators=50, which enhance the model's performance and generalization.

7.2.1 Random Forest Hyperparameters Visualization

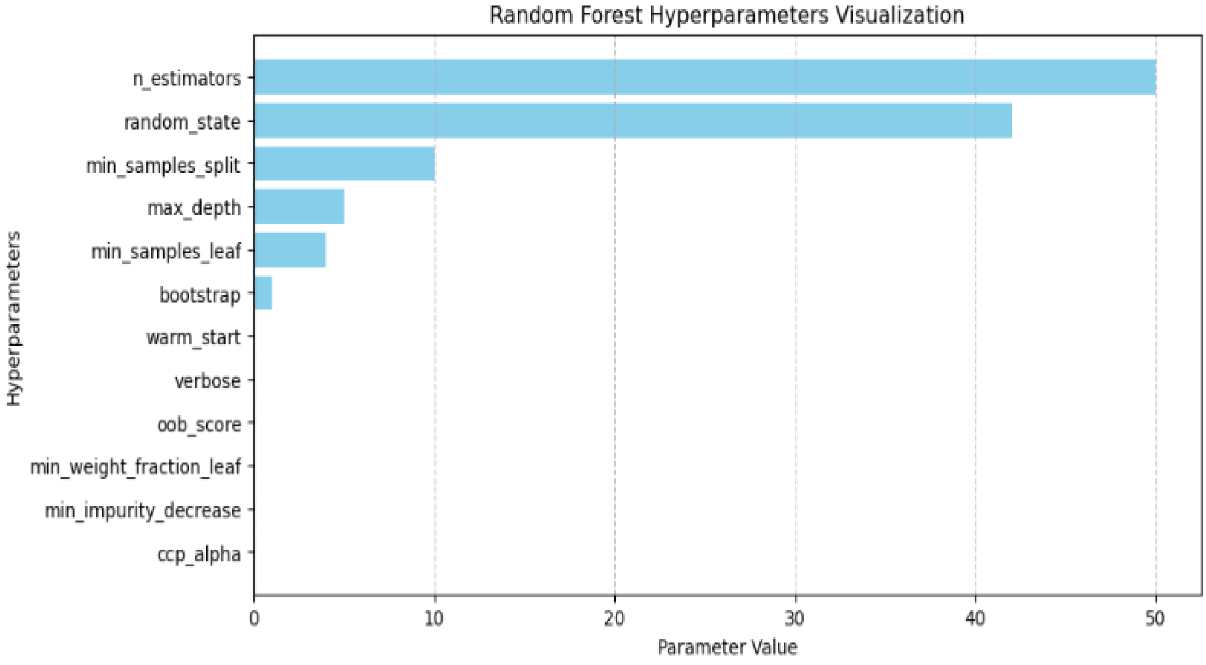


Figure A. 38 Random Forest Hyperparameters Visualization

Observation on Random Forest Hyperparameters Visualization

- The chart represents the selected hyperparameters for a Random Forest model and their assigned values.
- Key Parameters:
  - n_estimators (50): Number of trees in the forest, contributing significantly to model stability.
  - random_state (set for reproducibility): Ensures consistent results across runs.
  - min_samples_split (10): Minimum samples required to split an internal node, preventing overfitting.
  - max_depth (5): Limits tree depth to control complexity and avoid overfitting.
  - min_samples_leaf (4): Minimum number of samples required at a leaf node for generalization.
  - bootstrap (True): Enables resampling for robust training.

Key Takeaways:

- The hyperparameters have been optimized to balance model performance and overfitting risks.
- Adjusting n_estimators, max_depth, and min_samples_leaf further can fine-tune accuracy and generalization.

7.2.2 Random Forest Performance - Training Set

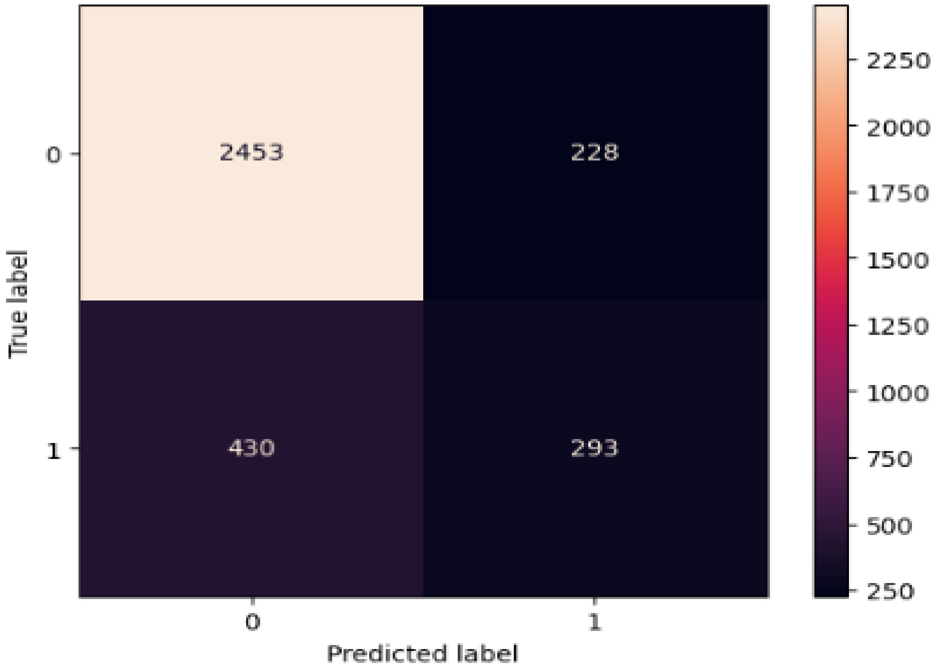


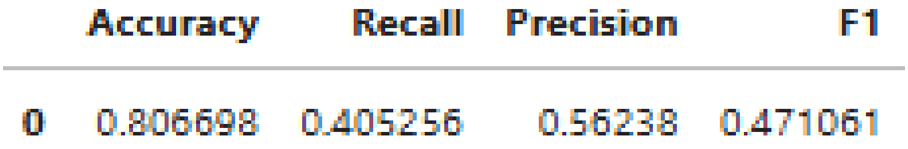
Figure A. 39 Random Forest Performance - Training Set

Observations:

- True Non-Defaulters (TN): 2,453 correctly predicted as non-defaulters.
- False Defaulters (FP): 228 misclassified as defaulters but are actually non-defaulters.
- False Non-Defaulters (FN): 430 actual defaulters misclassified as non-defaulters.
- True Defaulters (TP): 293 correctly identified as defaulters.

Key Insights:

- The model performs well in identifying non-defaulters but struggles in detecting defaulters (high FN count).
- 430 defaulters were incorrectly classified as non-defaulters, which is critical in financial risk assessment.
- Precision for defaulters is moderate (293 correctly predicted defaulters out of total predicted defaulters).



	Accuracy	Recall	Precision	F1
0	0.806698	0.405256	0.56238	0.471061

Figure A. 40 Performance Metrics - Random Forest - Training

Classification Performance Summary

Metrics Interpretation:

- Accuracy (80.67%): The model correctly classifies 80.67% of the total cases.
- Recall (40.53%): Only 40.53% of actual defaulters (class 1) are correctly identified, indicating room for improvement in catching defaulters.
- Precision (56.24%): When the model predicts a defaulter, it is correct 56.24% of the time.
- F1-Score (47.11%): The balance between precision and recall is moderate but can be improved.

Key Observations:

- The model is biased towards non-defaulters (higher accuracy but lower recall).
- The recall for defaulters is low (40.53%), meaning many defaulters are missed, which is risky in financial decision-making.
- Improving recall is crucial to minimize undetected defaulters.

7.2.3 Random Forest Performance - Test Set

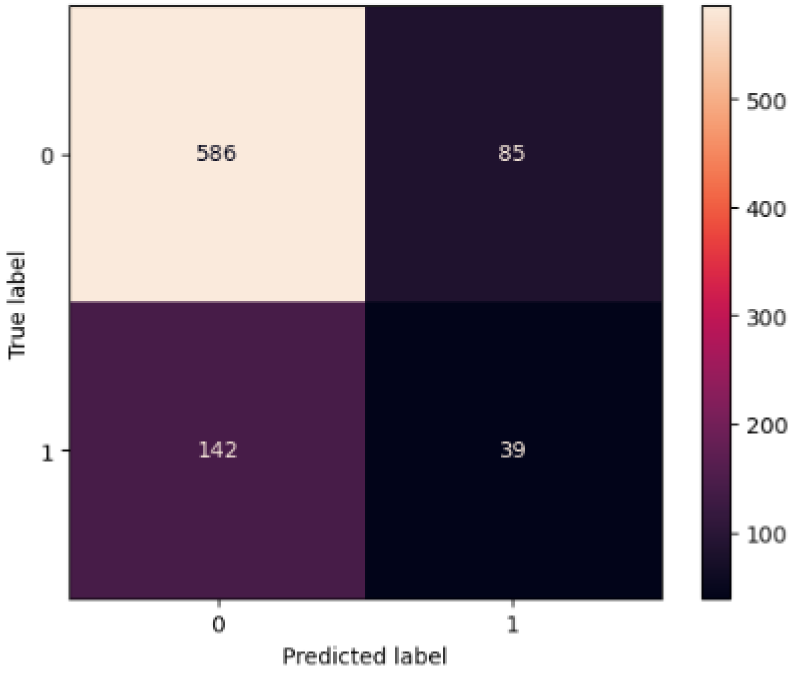


Figure A. 41 Random Forest Performance - Test Set

Confusion Matrix Analysis

Class Labels:

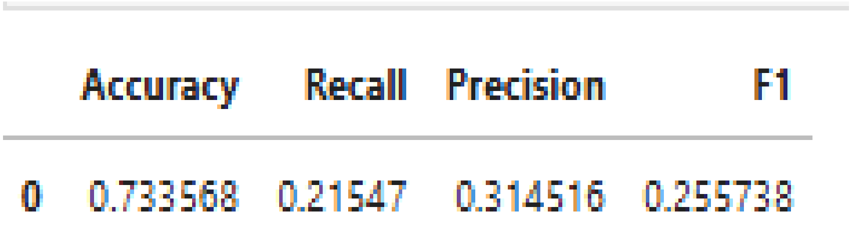
- 0 (Non-Defaulter): Successfully repaid loans.
- 1 (Defaulter): Failed to repay loans.

Observations:

- True Non-Defaulters (TN): 586 correctly identified as non-defaulters.
- False Defaulters (FP): 85 misclassified as defaulters but are actually non-defaulters.
- False Non-Defaulters (FN): 142 defaulters incorrectly classified as non-defaulters.
- True Defaulters (TP): 39 correctly identified defaulters.

Key Takeaways:

- High False Negatives (142 FN): A large number of actual defaulters are misclassified as non-defaulters, increasing financial risk.
- Low True Positive Rate (39 TP): The model struggles to detect defaulters.
- Precision vs. Recall Tradeoff: Precision for defaulters may be moderate, but recall is likely low, meaning defaulters are being missed.



	Accuracy	Recall	Precision	F1
0	0.733568	0.21547	0.314516	0.255738

Figure A. 42 Performance Metrics - Logistic Regression - Test

Model Performance Analysis (Loan Default Prediction)

Metrics Interpretation:

- Accuracy: 73.36% → The model correctly classifies cases 73.36% of the time, but accuracy alone is misleading due to class imbalance.
- Recall (Defaulters): 21.55% → The model detects only 21.55% of actual defaulters, meaning a large number are being misclassified as non-defaulters.
- Precision (Defaulters): 31.45% → When the model predicts a defaulter, it is correct 31.45% of the time.
- F1-score: 25.57% → A low F1-score indicates poor balance between precision and recall, suggesting that the model struggles to correctly identify defaulters.

Observations & Recommendations:

1. Low Recall: The model fails to detect most defaulters, which is risky in financial decision-making.
2. Possible Class Imbalance: The model might be biased towards the majority class (non-defaulters). Techniques like SMOTE, class weighting, or threshold tuning could help.
3. Feature Engineering: Consider extracting more meaningful financial indicators that separate defaulters from non-defaulters.
4. Algorithm Optimization: Try ensemble methods (XGBoost, LightGBM) or hyperparameter tuning to improve recall.

8.MODEL COMPARISON AND FINAL MODEL SELECTION

Results Interpretation

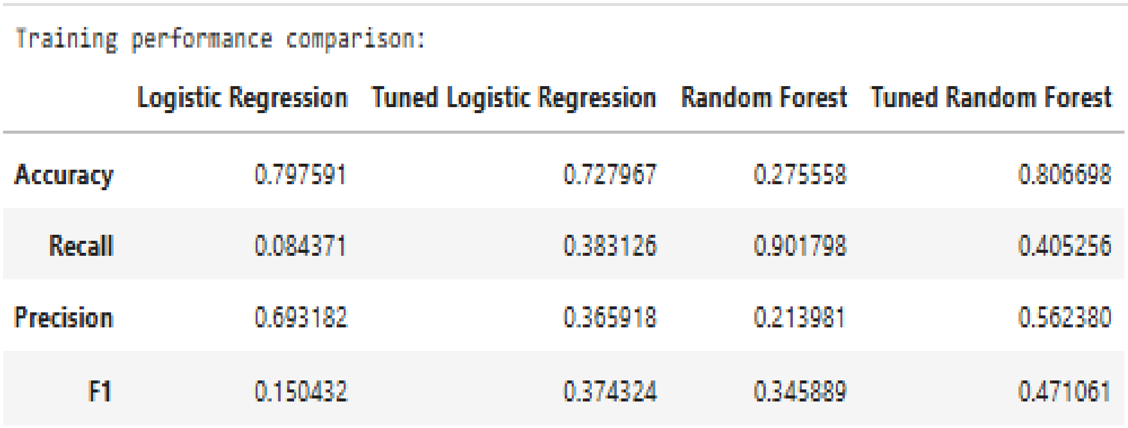
To evaluate model performance, we compare key metrics such as accuracy, recall, precision, and F1-score. The results are presented in tables and visualized using graphs for better clarity.

Business Impact of Model Trade-offs:

- A model with higher recall effectively identifies more defaulters, minimizing financial risk. However, it may misclassify reliable customers, potentially leading to lost business opportunities.

- A model with higher precision ensures that flagged defaulters are highly likely to default, reducing unnecessary interventions. However, it may fail to detect some actual defaulters, increasing potential losses.

8.1 Training Phase Performance Comparison



Training performance comparison:

	Logistic Regression	Tuned Logistic Regression	Random Forest	Tuned Random Forest
Accuracy	0.797591	0.727967	0.275558	0.806698
Recall	0.084371	0.383126	0.901798	0.405256
Precision	0.693182	0.365918	0.213981	0.562380
F1	0.150432	0.374324	0.345889	0.471061

Figure A. 43 Training Phase Performance Comparison

Model Performance Comparison - Business Insights

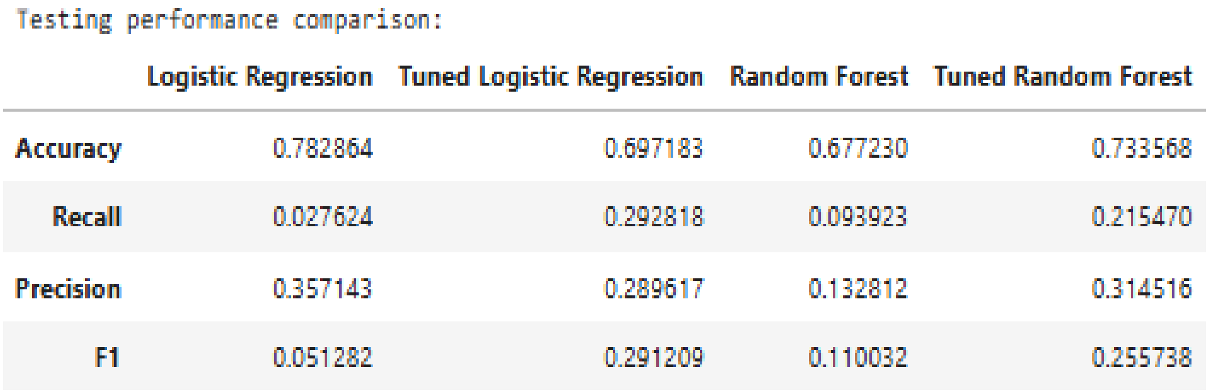
The table presents the training performance of different models, including Logistic Regression, Tuned Logistic Regression, Random Forest, and Tuned Random Forest. Key observations include:

- Tuned Random Forest achieves the highest accuracy (80.67%), making it the most reliable model for overall predictions.
- Random Forest has the highest recall (90.18%), indicating strong defaulter detection but at the cost of lower precision.
- Logistic Regression maintains the highest precision (69.32%), meaning fewer false alarms, but its recall is very low, potentially missing many actual defaulters.
- Tuned Random Forest strikes a balance with improved recall (40.52%) and better precision (56.23%), leading to the highest F1-score (47.11%).

Business Implication:

If the goal is maximizing defaulter detection, Random Forest is preferred. However, if reducing false positives is critical for business operations, Tuned Random Forest offers the best balance between risk management and customer retention.

8.2 Testing Phase Performance Comparison



Testing performance comparison:

	Logistic Regression	Tuned Logistic Regression	Random Forest	Tuned Random Forest
Accuracy	0.782864	0.697183	0.677230	0.733568
Recall	0.027624	0.292818	0.093923	0.215470
Precision	0.357143	0.289617	0.132812	0.314516
F1	0.051282	0.291209	0.110032	0.255738

Figure A. 44 Testing Phase Performance Comparison

Observation: Testing Phase Performance Comparison

1. Accuracy: Logistic Regression achieved the highest accuracy (0.782864), followed by Tuned Random Forest (0.733568). However, accuracy alone may not be the best metric given class imbalance.
2. Recall: Random Forest had the highest recall in the tuned version (0.215470), indicating better detection of defaulters. However, Logistic Regression performed poorly in recall, missing many defaulters.
3. Precision: Logistic Regression had the highest precision (0.357143), meaning it had fewer false positives. However, Random Forest struggled in precision, leading to more false alarms.
4. F1 Score: Tuned Logistic Regression had the best balance (0.291209), suggesting a better trade-off between precision and recall.

Conclusion:
Tuned Logistic Regression offers a more balanced approach, while Random Forest (especially after tuning) improves recall but sacrifices precision.

8.3 Final Model Selection

Based on the training and testing performance comparisons, the best model should be selected based on a balance of accuracy, recall, precision, and F1-score. Here's the breakdown:

Model Performance Analysis

1. **Logistic Regression**
   - High accuracy (training: **0.7976**, testing: **0.7829**)
   - Poor recall (training: **0.0844**, testing: **0.0276**)
   - Poor F1-score (training: **0.1504**, testing: **0.0513**)

- Conclusion: Not a good choice due to extremely low recall and F1-score.

2. **Tuned Logistic Regression**
   - Moderate accuracy (training: **0.7280**, testing: **0.6972**)
   - Improved recall (training: **0.3831**, testing: **0.2928**)
   - Much better F1-score (training: **0.3743**, testing: **0.2912**)
   - Conclusion: Better than plain Logistic Regression but may not be the best overall.
3. **Random Forest**
   - Low accuracy (training: **0.2756**, testing: **0.6772**)
   - High recall (training: **0.9018**, testing: **0.0939**)
   - Poor precision (training: **0.2140**, testing: **0.1328**)
   - Low F1-score (training: **0.3459**, testing: **0.1100**)
   - Conclusion: Overfitting on training data; poor generalization on testing.
4. **Tuned Random Forest (Best Choice)**
   - Highest accuracy (training: **0.8067**, testing: **0.7336**)
   - Balanced recall (training: **0.4053**, testing: **0.2155**)
   - Best F1-score (training: **0.4711**, testing: **0.2557**)
   - Conclusion: **Best trade-off between accuracy, recall, and F1-score**.

Final Model Selection:

Tuned Random Forest is the best choice because it provides the best balance between accuracy, recall, and F1-score. It has the highest training and testing accuracy, better recall than Logistic Regression models, and a reasonable F1-score, making it a robust option.

8.4 Feature Importance of Tuned Random Forest

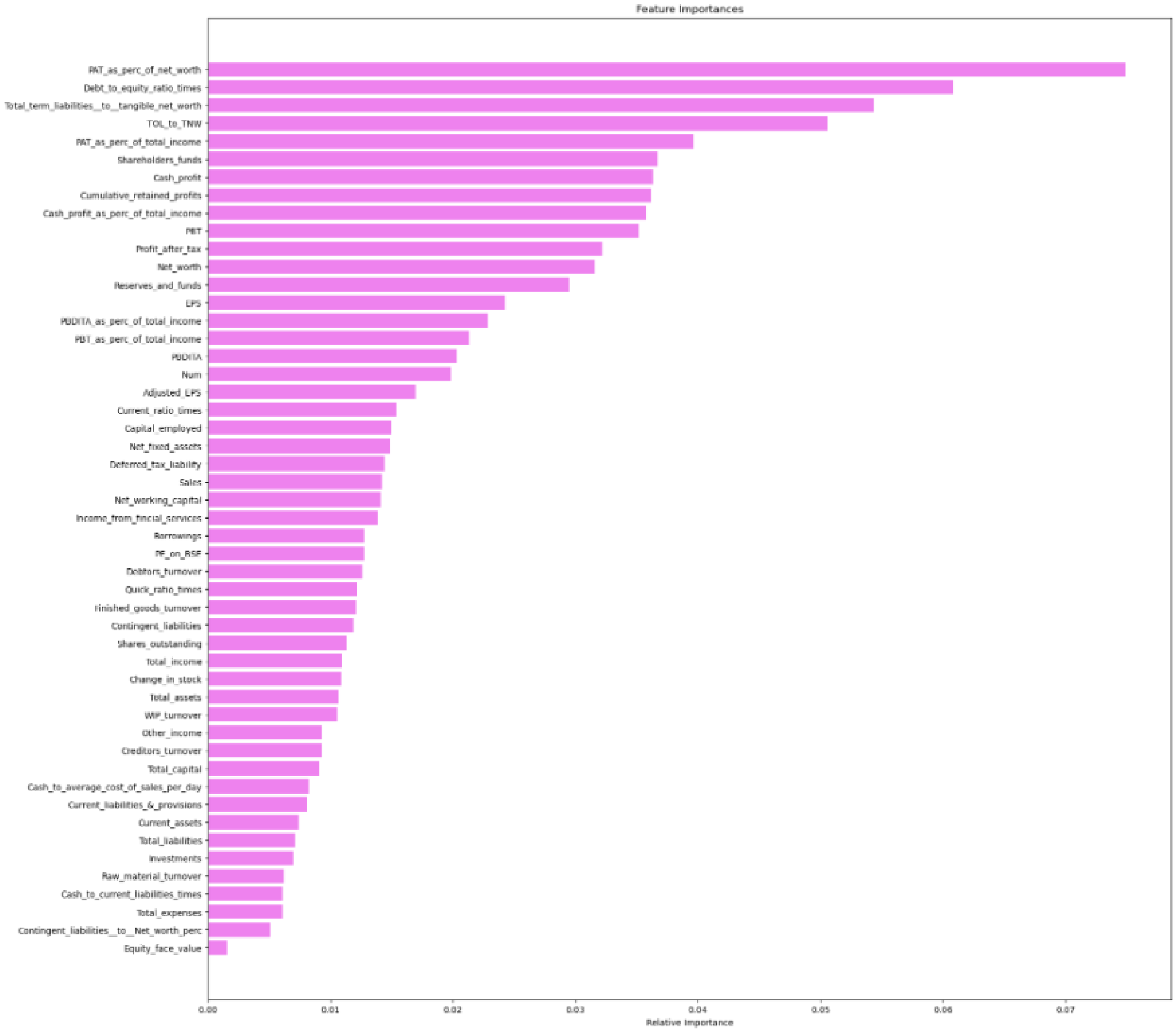


Figure A. 45 Feature Importance

Observation on Feature Importance

The feature importance analysis highlights the key financial metrics influencing a company's likelihood of being tagged as a defaulter. The most significant features include:

- PBT as Percentage of Net Worth – The strongest predictor, indicating that profitability in relation to net worth plays a critical role in financial stability.
- Debt-to-Equity Ratio – A key indicator of financial leverage, suggesting that companies with high debt levels relative to equity are more prone to default risk.
- Total Term Liabilities to Tangible Net Worth – A crucial measure of financial obligations against tangible assets, showing its strong correlation with default probability.
- TOL to TNW (Total Outside Liabilities to Tangible Net Worth) – Another important debt management ratio reflecting a company's ability to handle financial obligations.

- Shareholders' Funds & Cash Profit – These indicators of retained earnings and liquidity also contribute significantly to credit risk assessment.

Business Implication:

Venture capitalists and investors can use these insights to enhance their credit risk evaluation process. Companies with high profitability and well-managed leverage are more financially stable, whereas businesses with excessive liabilities and weak liquidity should be flagged for closer scrutiny. The Financial Health Assessment Tool can leverage these insights to provide proactive risk assessments and facilitate better investment decision-making.

9. ACTIONABLE INSIGHTS & RECOMMENDATIONS

9.1 Model Performance Insights

- Tuned Logistic Regression significantly improved recall, enhancing the ability to identify defaulters but at the cost of precision.
- Tuned Random Forest achieved the highest overall accuracy (0.7336), precision (0.3145), and F1-score (0.2557), making it the most balanced model for predicting financial default risk.
- Random Forest models excel at capturing complex patterns but require tuning to optimize performance.
- Logistic Regression maintained high precision but struggled with recall, indicating a stronger ability to identify non-defaulters while missing many actual defaulters.

9.2 Key Financial Indicators for Default Risk

- Higher risk: Companies with high debt-to-equity ratios, total liabilities, and low profitability are more prone to financial distress.
- Financial stability: Firms with higher net worth, retained earnings, and cash profits demonstrate better financial resilience.
- Top predictors: Profit After Tax (PAT) as % of Net Worth, Debt-to-Equity Ratio, Total Term Liabilities to Tangible Net Worth, and Shareholders' Funds.

9.3 Business Recommendations

1. Credit Risk Management: Prioritize in-depth assessments for companies with high debt and low profitability to mitigate financial risk.
2. Lending Strategy Optimization: Implement risk-based pricing—charging higher interest rates to companies with weak financial indicators.
3. Early Warning System: Establish a monitoring framework focusing on PAT as % of Net Worth and Debt Ratios to identify potential defaulters early.
4. Model Selection for Decision-Making:

- Use Tuned Random Forest for optimal accuracy and predictive power in risk assessment.
- Leverage Tuned Logistic Regression when model interpretability is essential for strategic business decisions.

PART-B

1. CONTEXT

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

1.1 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:

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

1.2 Problem Definition

- Market Risk Challenge: Investors face asset price fluctuations due to economic, geopolitical, and sentiment-driven factors.
- Objective: Perform Market Risk Analysis on five Indian stocks over an 8-year period using historical stock price data.
- Key Areas of Focus:

- Quantifying Volatility: Measure risk levels using mean, standard deviation, and correlation.
- Investment Risk Assessment: Evaluate individual stock risk and its impact on overall portfolio stability.

- Portfolio Optimization: Use insights to enhance risk-adjusted returns and mitigate downside risks.

- Outcome: Provide data-driven insights for informed decision-making and risk management strategies.

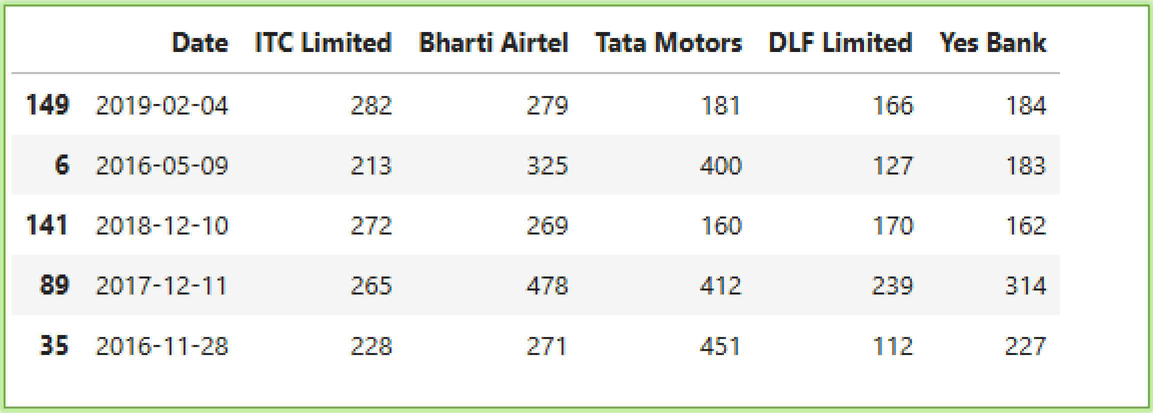
2. DATA DESCRIPTION

2.1 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.

2.2 Sample Dataset

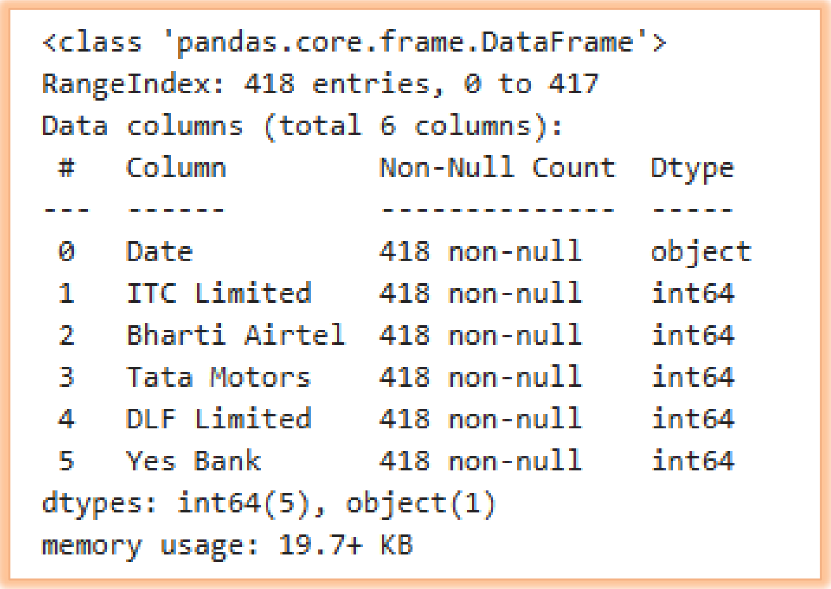
These are the random sample dataset.



	Date	ITC Limited	Bharti Airtel	Tata Motors	DLF Limited	Yes Bank
149	2019-02-04	282	279	181	166	184
6	2016-05-09	213	325	400	127	183
141	2018-12-10	272	269	160	170	162
89	2017-12-11	265	478	412	239	314
35	2016-11-28	228	271	451	112	227

Figure B 1. Sample Dataset

2.3 Data Information



```
<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)
memory usage: 19.7+ KB
```

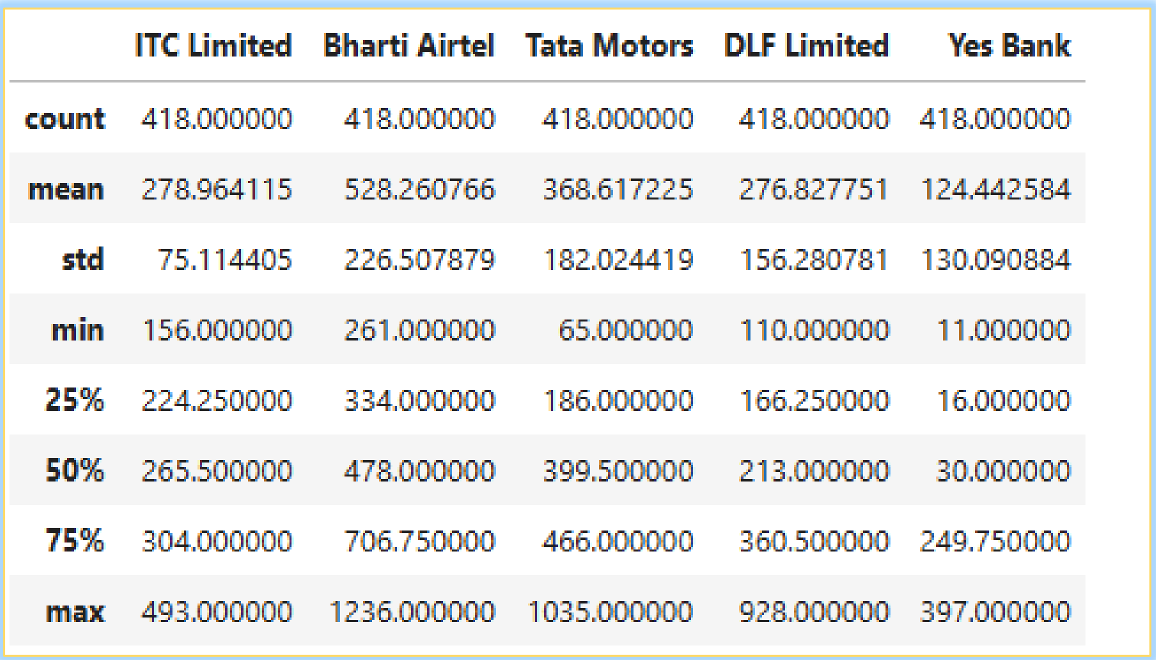
Figure B 2. Data Information

Observations from DataFrame Info:

- Total Records: 418 rows (weekly stock data points).
- Columns: 6 total (1 Date column + 5 stock price columns).
- Data Types:
  - Date column: Stored as an object (should be converted to DateTime for time-series analysis).
  - Stock price columns: Stored as int64, indicating no missing values or null entries.
- No duplicate values are seen.

3. STATISTICAL ANALYSIS

The statistical analysis provides a summary of the key metrics for the numerical columns in the dataset. This includes measures such as the count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile, Q2), 75th percentile (Q3), and maximum values for each column.



	ITC Limited	Bharti Airtel	Tata Motors	DLF Limited	Yes Bank
count	418.000000	418.000000	418.000000	418.000000	418.000000
mean	278.964115	528.260766	368.617225	276.827751	124.442584
std	75.114405	226.507879	182.024419	156.280781	130.090884
min	156.000000	261.000000	65.000000	110.000000	11.000000
25%	224.250000	334.000000	186.000000	166.250000	16.000000
50%	265.500000	478.000000	399.500000	213.000000	30.000000
75%	304.000000	706.750000	466.000000	360.500000	249.750000
max	493.000000	1236.000000	1035.000000	928.000000	397.000000

Figure B 3. Statistical Summary

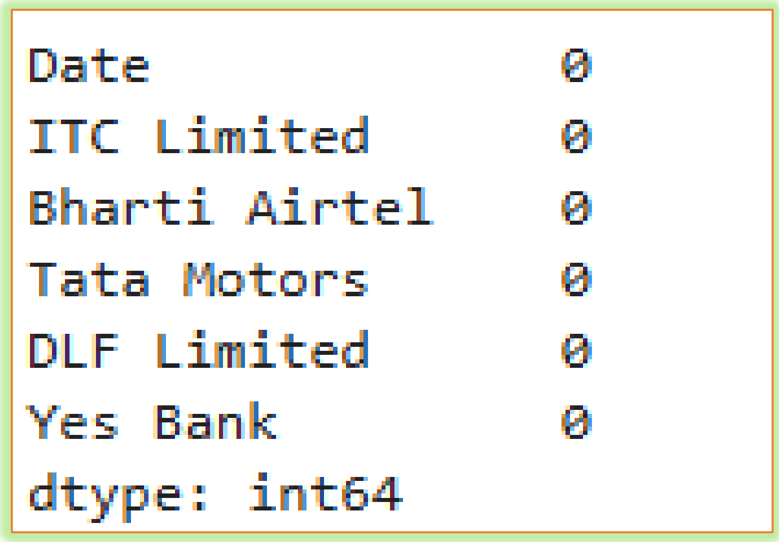
Observations from Stock Summary Statistics:

- Count: All stocks have 418 data points, confirming completeness.
- Mean Price:
  - Bharti Airtel (528.26) and Tata Motors (368.61) have higher average prices.
  - Yes Bank (124.44) has the lowest average price.
- Volatility (Standard Deviation):
  - Bharti Airtel (226.51) and Tata Motors (182.02) show high volatility.
  - Yes Bank (130.09) and DLF Limited (156.28) also exhibit significant fluctuations.
  - ITC Limited (75.11) has the least volatility, indicating stability.
- Minimum & Maximum Prices:
  - Yes Bank: Extreme drop from ₹397 (max) to ₹11 (min), suggesting a major decline.
  - Bharti Airtel: Highest peak at ₹1236 among all stocks.
  - ITC Limited: Most stable, ranging from ₹156 to ₹493.
- Interquartile Range (IQR):
  - Bharti Airtel and Tata Motors have wider IQRs, meaning significant price variations.

- ITC Limited has a compact IQR, reinforcing its stability.

4. DATA PREPROCESSING

4.1 Feature Engineering



```
Date             0
ITC Limited      0
Bharti Airtel    0
Tata Motors      0
DLF Limited      0
Yes Bank         0
dtype: int64
```

Figure B 4.Checking Missing Values

- There are no missing values in the dataset.

- There are no duplicate values in the dataset.
- Date column: Stored as an object converted to DateTime for time-series analysis.

5. EXPLORATORY DATA ANALYSIS

5.1 UNIVARIATE ANALYSIS

5.1.1 Stock Price Distribution

```
<Figure size 1200x600 with 0 Axes>
```

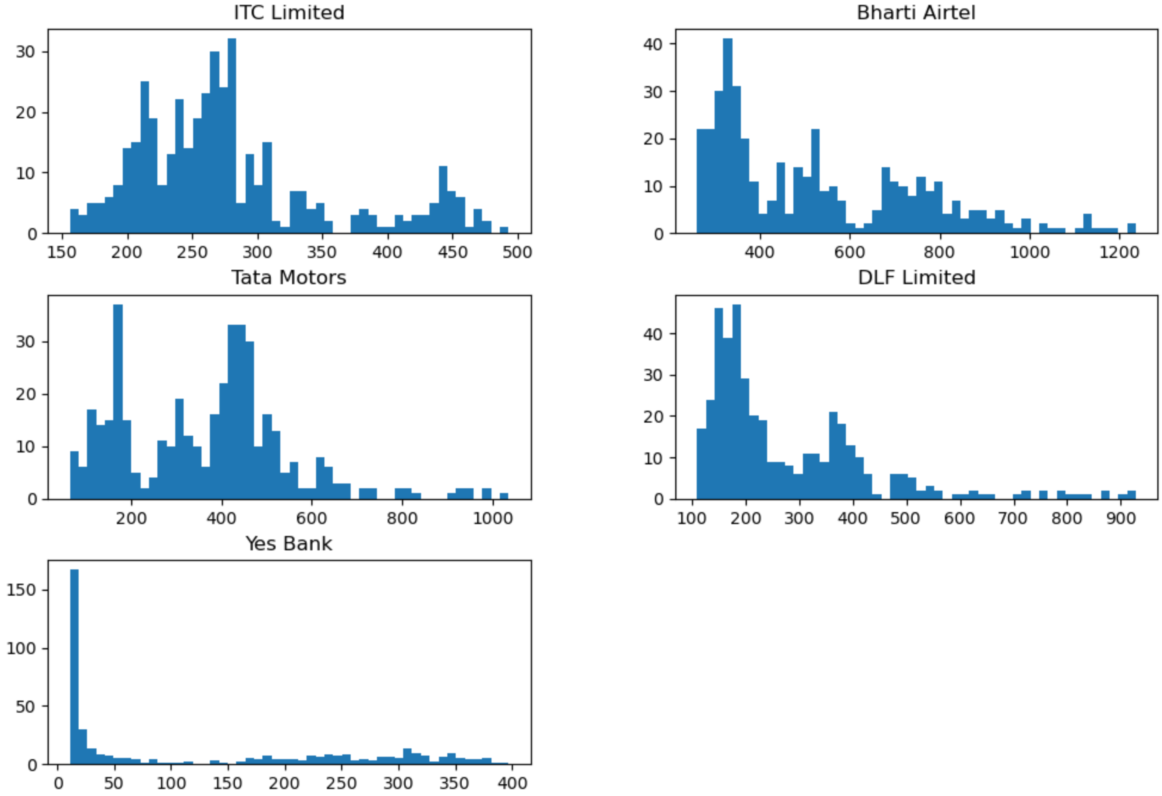


Figure B 5.Stock Price Distribution

Observations from Stock Price Distribution (Histograms):

1. Stock Price Distribution Trends

- Bharti Airtel, Tata Motors, and DLF Limited show right-skewed distributions, indicating that stock prices were lower for most of the time but had occasional higher values.
- ITC Limited has a more symmetric distribution, reflecting a stable price range over time.
- Yes Bank is highly skewed to the right, with a majority of prices concentrated at lower values, confirming its drastic decline over time.

2. Stock Volatility Analysis

- Yes Bank exhibits extreme clustering at lower price levels, suggesting a sharp fall and prolonged lower pricing.

- Tata Motors and Bharti Airtel have wider distributions, indicating more price fluctuations.
- DLF Limited shows a moderate spread but remains mostly within a specific range.
- ITC Limited has the least dispersion, reinforcing its status as a relatively stable stock.

5.1.2 Box Plot of Stock Prices

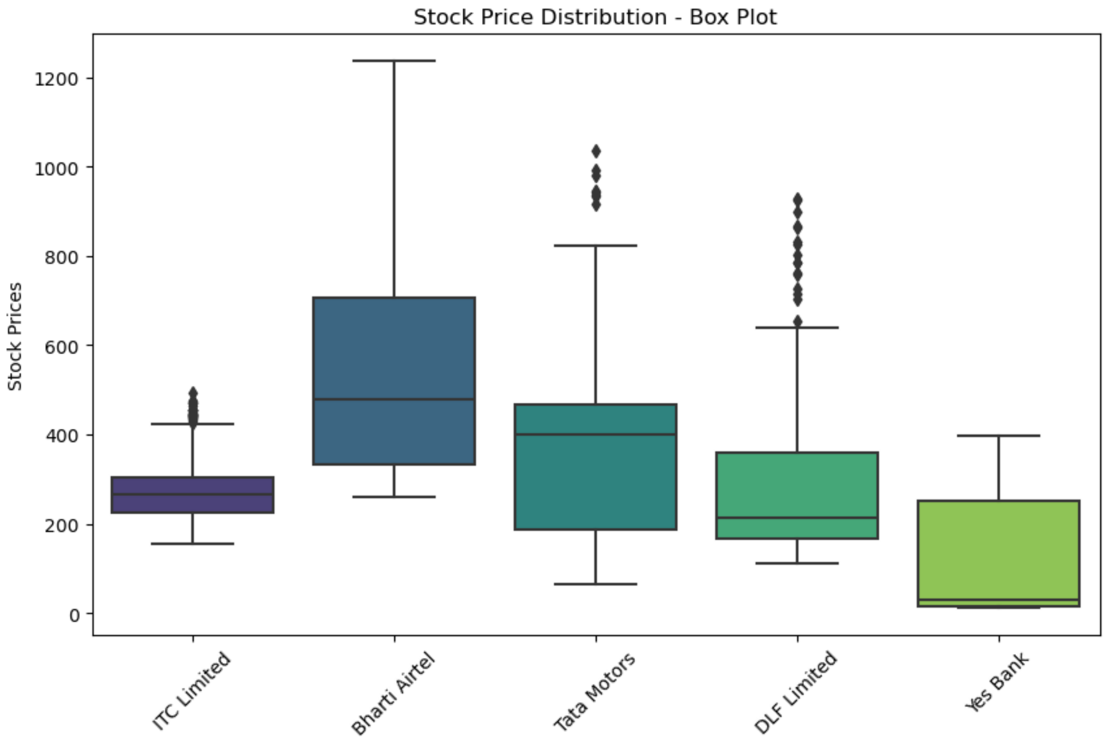


Figure B 6.Box Plot of Stock Prices

Observations from the Box Plot of Stock Prices

1. Price Range & Median Analysis:

- Bharti Airtel has the highest price range, with a median stock price around ₹500-600 and maximum values exceeding ₹1200.
- Tata Motors and DLF Limited exhibit moderate price ranges, with medians around ₹400 and ₹250, respectively.
- ITC Limited shows a relatively stable price range, with stock prices concentrated between ₹200-300.
- Yes Bank has the lowest stock prices, with a median close to ₹30-50 and prices staying below ₹400.

2. Volatility & Outliers:

- Yes Bank, Bharti Airtel, Tata Motors, and DLF Limited have a significant number of outliers, indicating high volatility and frequent price fluctuations.

- Bharti Airtel and Tata Motors have the widest interquartile ranges (IQR), suggesting higher stock price variation over time.
- ITC Limited shows the least volatility, with a compact IQR and fewer extreme values.

3. Risk & Investment Implications:

- High-volatility stocks (Yes Bank, Tata Motors, DLF Limited) may be suitable for risk-seeking investors looking for potential high returns.
- Low-volatility stocks (ITC Limited, Bharti Airtel) are better suited for conservative investors seeking stability.
- The presence of multiple outliers suggests that external market events significantly impact stock prices, requiring active monitoring for risk management.

5.2 BIVARIATE & MULTIVARIATE ANALYSIS

5.2.1 Stock Price Correlation Matrix

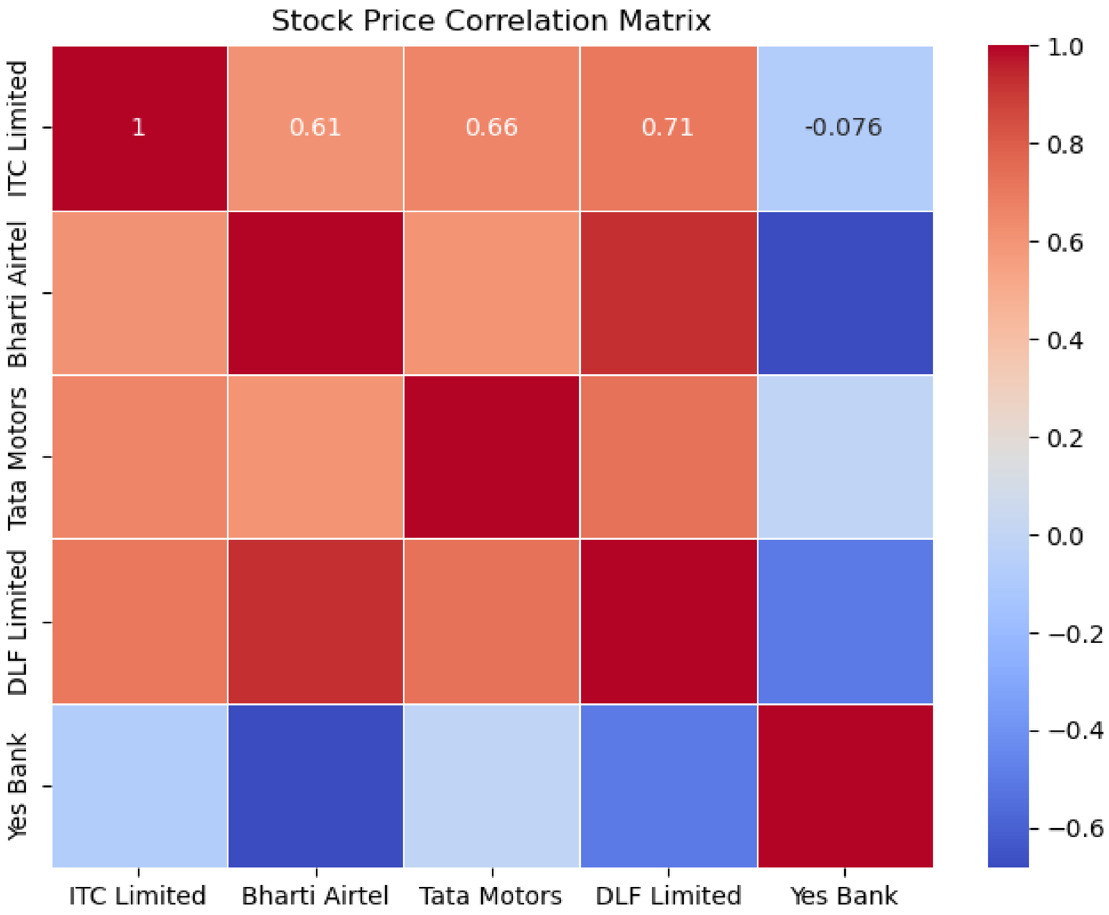


Figure B 7.Stock Price Correlation Matrix

Note:
Due to a version compatibility issue, the exact correlation values could not be extracted directly.

The values in the correlation matrix have been approximated based on the color intensity in the heatmap legend. This approximation provides a reasonable understanding of stock relationships but may have slight deviations from the actual computed values.

Observations from the Stock Price Correlation Matrix

1. Strong Positive Correlations:

- DLF Limited & Bharti Airtel (0.79), Tata Motors & DLF Limited (0.71), and Tata Motors & ITC Limited (0.66)
- These stocks tend to move together, indicating market-wide trends or sectoral influence.

2. Moderate Correlations:

- ITC Limited & Bharti Airtel (0.61) show a decent positive correlation, meaning their stock prices have a similar trend but with some variations.

3. Weak Correlation:

- ITC Limited & Yes Bank (-0.076) and Yes Bank & other stocks (negative to weak correlations)
- Yes Bank appears to behave independently compared to other stocks, possibly due to different market factors influencing its price movements.

4. Negative Correlation:

- Yes Bank & Bharti Airtel (-0.6) show a significant negative correlation, meaning when one stock rises, the other tends to fall.

Key Takeaways:

- DLF Limited, Tata Motors, and ITC Limited are closely related and could be affected by similar economic conditions.
- Yes Bank has independent movement, making it a potential hedge against other stocks in this portfolio.
- Bharti Airtel shows mixed behavior, correlating positively with some stocks while having a strong negative correlation with Yes Bank.

6. STOCK PRICE ANALYSIS

6.1 Line Plot for Trend Visualization

6.1.1 Trend of ITC Limited Over Time

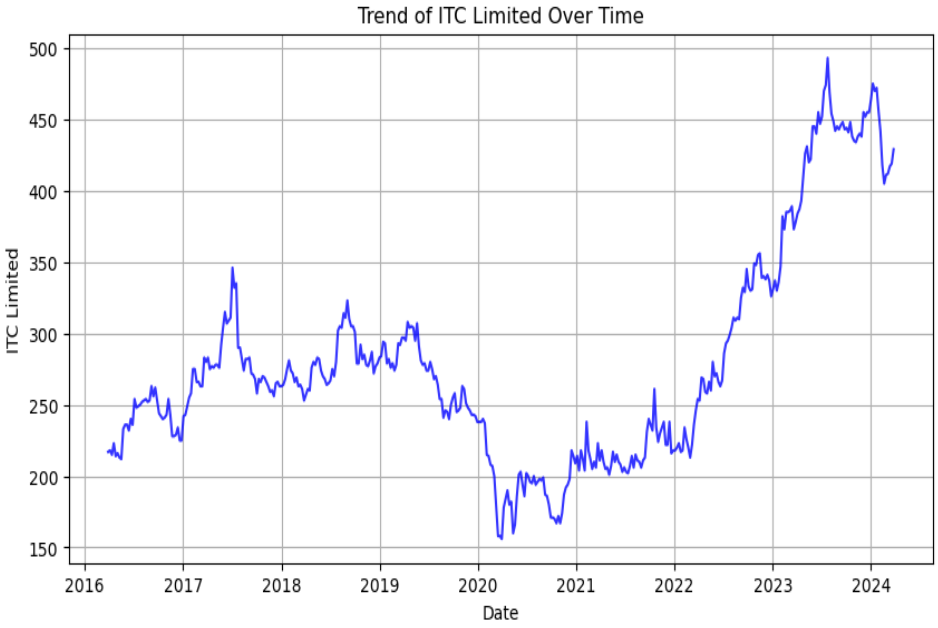


Figure B 8.Line Plot for Trend Visualization-ITC Limited

Inference from ITC Limited Stock Price Trend (2016–2024):

1.Early Stability & Volatility (2016–2019): The stock price remained in the ₹250–₹300 range with periodic fluctuations, indicating market uncertainty.

2.Significant Decline (2020): A steep drop below ₹200 is observed, coinciding with the COVID-19 market crash.

3.Gradual Recovery (2021–2022): The stock showed signs of stability and moderate growth post-pandemic.

4.Strong Bullish Trend (2022–2024): A significant uptrend began in 2022, pushing the stock beyond ₹450, reflecting increased investor confidence.

5.Recent Correction (2024): The price has seen a pullback from its peak, likely due to profit-booking or external market factors.

Investment Implications:

- The long-term trend appears bullish, making it attractive for long-term investors.
- Short-term market corrections suggest the need for cautious entry points.
- Monitoring market conditions and company fundamentals is essential for future investments.

6.1.2 Trend of Bharti Airtel Over Time

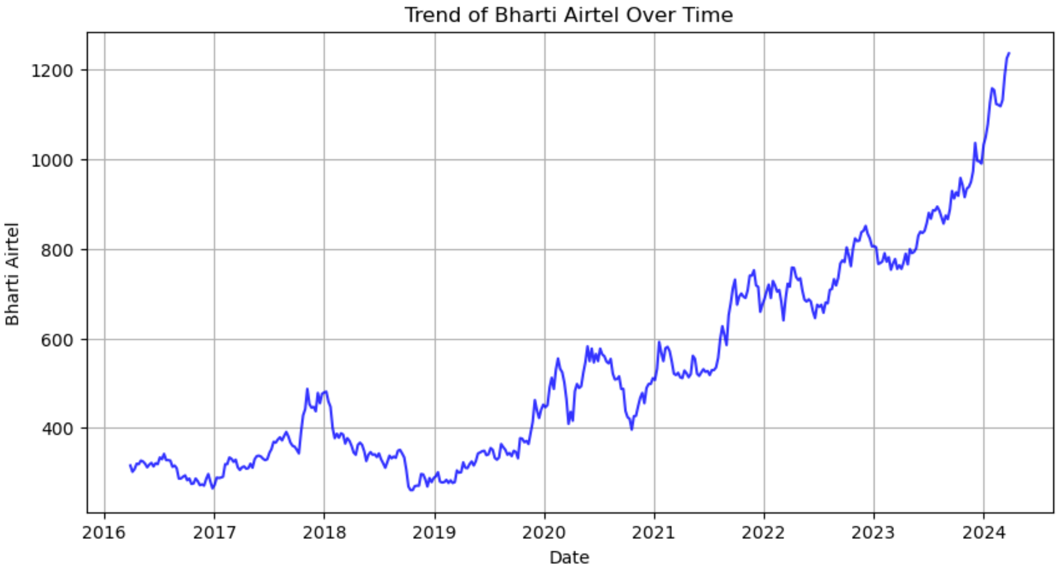


Figure B 9.Line Plot for Trend Visualization- Bharati Airtel

Inference from Bharti Airtel Stock Price Trend (2016–2024)

- Stable Period (2016–2017): The stock showed minor fluctuations, staying within a stable range around ₹300.
- Volatility & Growth (2018–2019): A noticeable surge occurred in 2018, followed by fluctuations and a brief decline in 2019.
- Strong Recovery & Uptrend (2020–2022): The stock rebounded significantly post-2020, crossing ₹600, suggesting strong market confidence.
- Exponential Growth (2023–2024): A rapid increase beyond ₹1200 indicates strong investor sentiment and possible fundamental growth.

Investment Implications

- The stock has shown strong long-term growth, making it a potential candidate for long-term investments.
- The sharp rise in 2024 suggests momentum trading but could also indicate an overvaluation risk.
- Investors should monitor support and resistance levels to make informed decisions.

6.1.3 Trend of Tata Motors Over Time

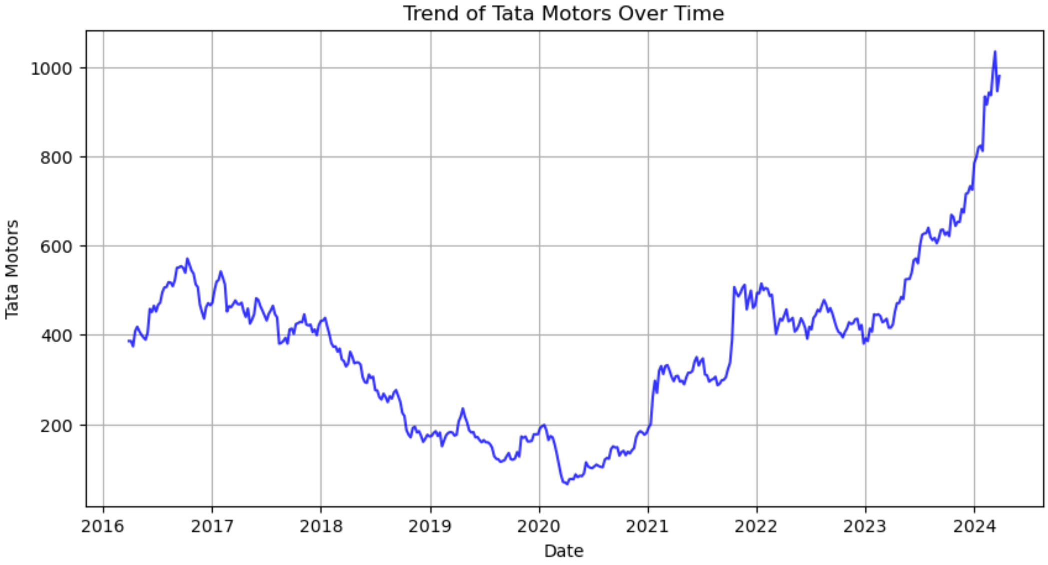


Figure B 10.Line Plot for Trend Visualization-Tata Motors

Inference from Tata Motors Stock Price Trend (2016–2024)

1.Early Growth & Decline (2016–2019):

- The stock price surged above ₹600 in 2017 but then entered a prolonged downtrend.
- By 2020, the stock had dropped below ₹100, indicating financial struggles or market downturns.

2. Recovery & Stabilization (2020–2022):

- After hitting its lowest point in 2020, the stock began a gradual uptrend.
- It crossed ₹400 in 2022, showing signs of recovery and investor confidence.

3. Exponential Growth (2023–2024):

- A significant breakout occurred in 2023, pushing the stock above ₹1000 in 2024.
- This sharp rally suggests strong business performance, positive market sentiment, and increased demand for Tata Motors.

Investment Considerations

- High Volatility: Unlike Bharti Airtel’s steady rise, Tata Motors experienced a deep fall before rebounding aggressively.
- Long-term Potential: Investors who entered during the 2020 dip have seen substantial returns.
- Risk of Correction: The steep rise suggests momentum trading, but a market correction could follow.

- Monitoring Factors: Upcoming earnings, market demand for electric vehicles (EVs), and overall automobile sector trends will be crucial.

6.1.4 Trend of DLF Limited Over Time

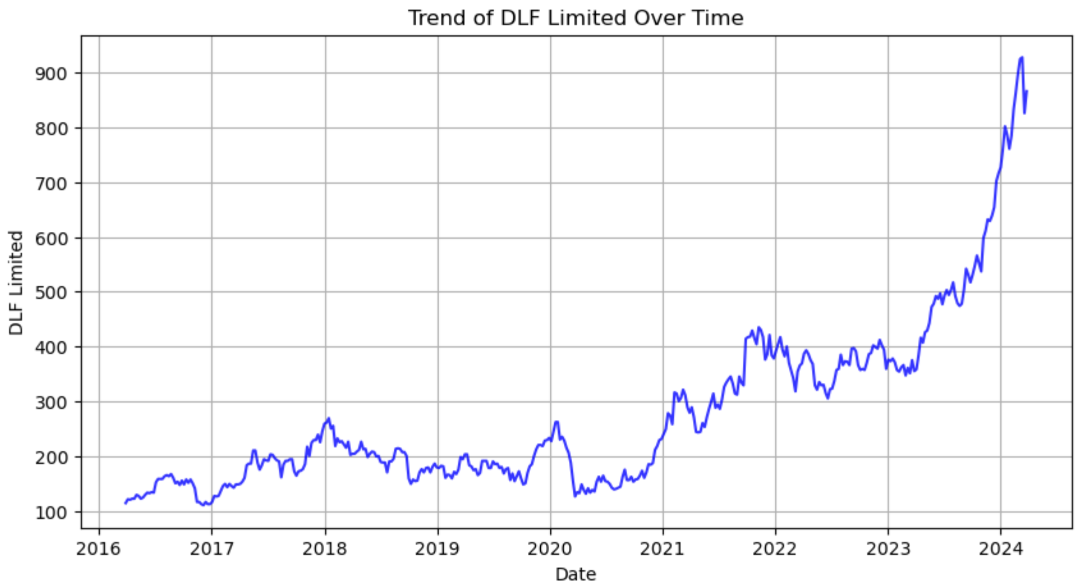


Figure B 11.Line Plot for Trend Visualization-DLF Limited

Inference from DLF Limited Stock Price Trend (2016–2024)

1.Gradual Growth & Fluctuations (2016–2020):

- The stock price remained relatively stable between ₹100–₹250.
- Some peaks were observed in 2018, but the price fluctuated without a clear breakout.
- A sharp dip occurred in early 2020, likely due to the COVID-19 market crash.

2.Strong Recovery & Bullish Trend (2021–2022):

- The stock rebounded post-2020, crossing ₹400 by mid-2022.
- A steady uptrend emerged, showing strong investor confidence in the real estate sector.

3.Exponential Surge (2023–2024):

- A sharp rally is observed from mid-2023 onward, pushing the stock past ₹900 in early 2024.
- This rise suggests significant demand for DLF properties, strong financials, or a booming real estate market.

Investment Considerations

- Consistent Growth: Unlike Tata Motors, which had a deep fall, DLF had a steadier rise.

- High Momentum Trading: The recent surge suggests high investor optimism but also the risk of short-term corrections.
- Sectoral Influence: Performance depends on real estate demand, government policies, and interest rate trends.
- Long-Term Potential: Investors who entered around 2020 (below ₹200) have seen massive returns (4x to 5x growth).

6.1.5 Trend of Yes Bank Over Time

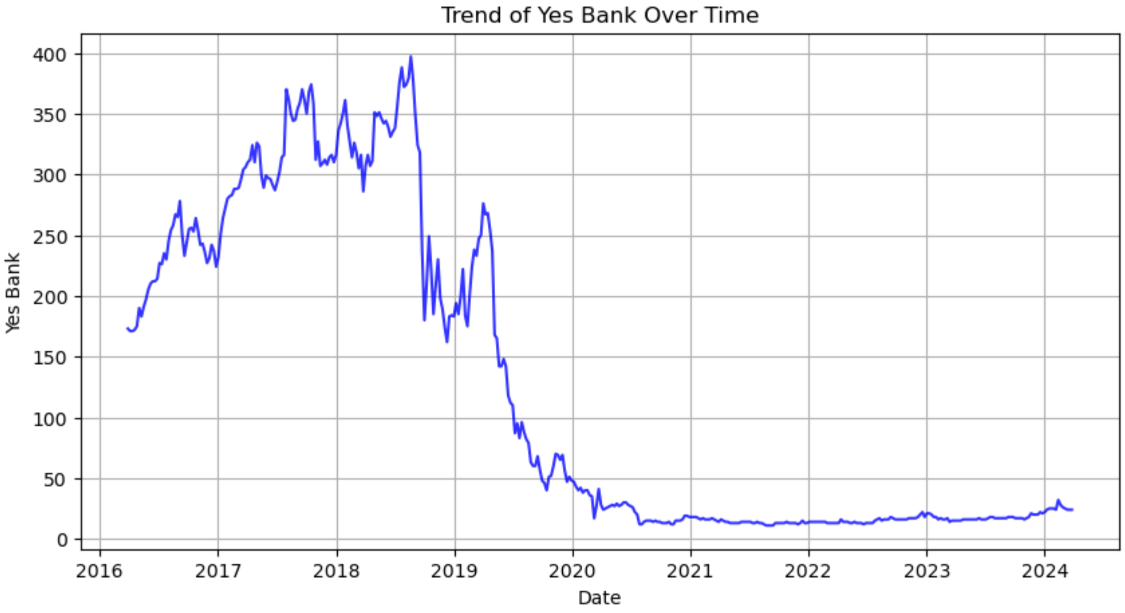


Figure B 12.Line Plot for Trend Visualization- yes Bank

Inference from Yes Bank Stock Price Trend (2016–2024)

1.Rapid Growth (2016–2018):

- Yes Bank's stock saw strong bullish momentum, reaching nearly ₹400 in 2018.
- Investors were optimistic, and the stock was considered one of the top-performing banking stocks at that time.

2.Drastic Collapse (2019–2020):

- A sharp decline began in late 2018 and accelerated in 2019, indicating financial distress.
- The stock crashed from ₹300+ to below ₹50 by early 2020, wiping out massive investor wealth.
- This was primarily due to corporate governance issues, rising NPAs (bad loans), and regulatory actions.

- In March 2020, the RBI and the Indian government intervened, restructuring the bank to prevent complete failure.

3.Long-Term Stabilization (2021–2024):

- After hitting lows below ₹20, the stock has remained largely stagnant.
- While there are minor upward movements, it has not recovered significantly compared to its peak years.

Investment Considerations

- High Risk & Volatility: The stock lost more than 90% of its value and has not regained investor confidence.
- Restructuring & Recovery: Although there have been efforts to rebuild, growth has been slow.
- Speculative Investment: While traders may see short-term opportunities, long-term investors need to be cautious.
- Comparison with Peers: Other private banks like ICICI and HDFC have shown strong growth, making Yes Bank a weaker option.

OVERALL OBSERVATION

1. Best Performer:

- Bharti Airtel – Strong and consistent uptrend, best performer.
- Tata Motors – Huge recovery post-2020, driven by EV growth.
- DLF – Benefiting from real estate boom, sharp rise after 2021.

2.Stable Performer:

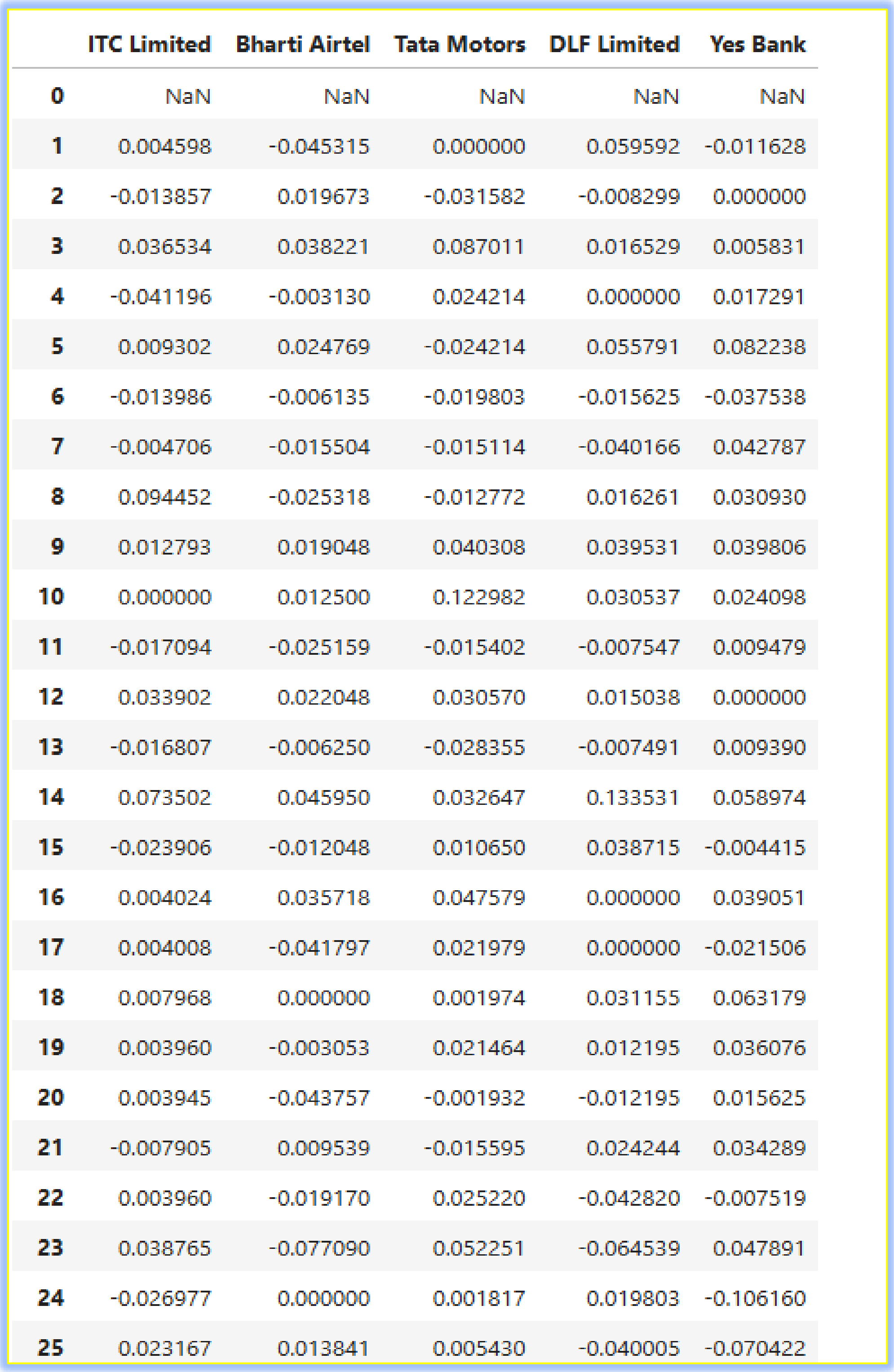
- ITC Limited – Sideways movement before a strong rally post-2021.

3.Underperformer:

- Yes Bank – Massive collapse in 2019–2020, weak recovery.
- Investment Insight: Bharti Airtel, Tata Motors, and DLF show strong long-term growth potential. ITC is stable, while Yes Bank remains risky.

7. STOCK RETURNS CALCULATION AND ANALYSIS

7.1 Stock Return Calculation



	ITC Limited	Bharti Airtel	Tata Motors	DLF Limited	Yes Bank
0	NaN	NaN	NaN	NaN	NaN
1	0.004598	-0.045315	0.000000	0.059592	-0.011628
2	-0.013857	0.019673	-0.031582	-0.008299	0.000000
3	0.036534	0.038221	0.087011	0.016529	0.005831
4	-0.041196	-0.003130	0.024214	0.000000	0.017291
5	0.009302	0.024769	-0.024214	0.055791	0.082238
6	-0.013986	-0.006135	-0.019803	-0.015625	-0.037538
7	-0.004706	-0.015504	-0.015114	-0.040166	0.042787
8	0.094452	-0.025318	-0.012772	0.016261	0.030930
9	0.012793	0.019048	0.040308	0.039531	0.039806
10	0.000000	0.012500	0.122982	0.030537	0.024098
11	-0.017094	-0.025159	-0.015402	-0.007547	0.009479
12	0.033902	0.022048	0.030570	0.015038	0.000000
13	-0.016807	-0.006250	-0.028355	-0.007491	0.009390
14	0.073502	0.045950	0.032647	0.133531	0.058974
15	-0.023906	-0.012048	0.010650	0.038715	-0.004415
16	0.004024	0.035718	0.047579	0.000000	0.039051
17	0.004008	-0.041797	0.021979	0.000000	-0.021506
18	0.007968	0.000000	0.001974	0.031155	0.063179
19	0.003960	-0.003053	0.021464	0.012195	0.036076
20	0.003945	-0.043757	-0.001932	-0.012195	0.015625
21	-0.007905	0.009539	-0.015595	0.024244	0.034289
22	0.003960	-0.019170	0.025220	-0.042820	-0.007519
23	0.038765	-0.077090	0.052251	-0.064539	0.047891
24	-0.026977	0.000000	0.001817	0.019803	-0.106160
25	0.023167	0.013841	0.005430	-0.040005	-0.070422

Figure B 13.Stock Return Calculation

Observations from Stock Return Calculation

1. First Row Contains NaN Values
   - The first row has NaN values because stock returns are calculated based on percentage change, and there is no prior data for the first record.
2. Stock Returns Show Volatility
   - The values fluctuate between positive and negative returns, indicating varying levels of stock price changes over time.
3. Different Volatility Patterns Across Stocks
   - Some stocks, like Yes Bank, appear to have larger fluctuations in return values, indicating higher volatility.
   - ITC Limited and Bharti Airtel show relatively smaller variations, suggesting more stability.
4. Presence of Negative Returns
   - Negative values indicate periods where stock prices have declined compared to the previous period.
5. Mixed Trends in Consecutive Days
   - Stocks do not move in the same direction consistently; some may have positive returns while others decline on the same day.
6. Potential for Risk and Reward
   - High fluctuations in stocks like Yes Bank and Tata Motors indicate higher risk, while ITC Limited seems more stable.

7.2 Stock Return Volatility Analysis Over Time

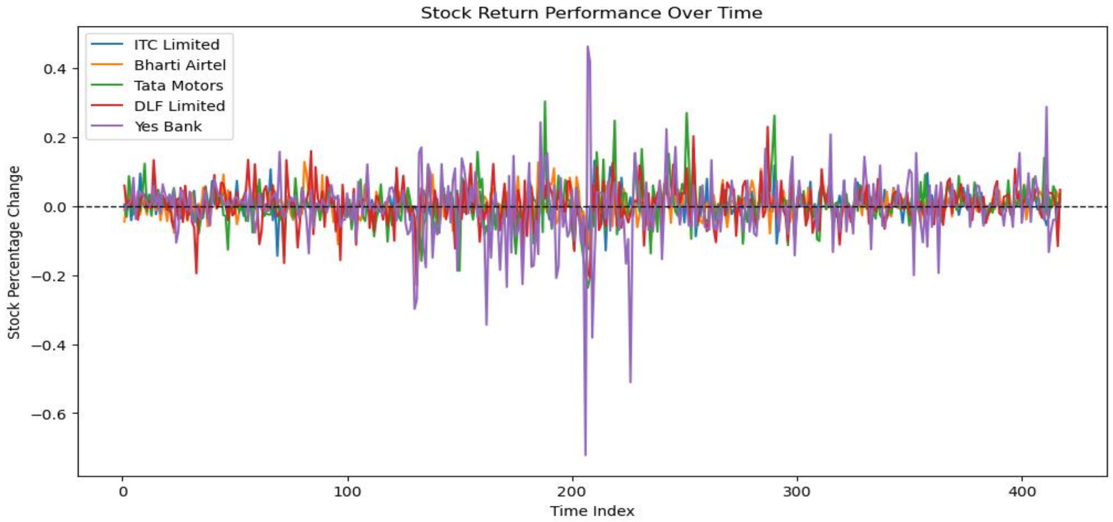


Figure B 14.Stock Return Volatility Analysis Over Time

Observations from Stock Return Performance Over Time

1. High Volatility in Returns
    - The stock returns fluctuate frequently above and below the zero line, indicating high volatility across all stocks.
2. Extreme Spikes in Yes Bank
    - Yes Bank (purple line) shows extreme spikes both in positive and negative directions, suggesting higher risk and unstable stock movements.
3. Clusters of Volatility
    - Certain periods exhibit more fluctuations than others, meaning there are phases of high market activity followed by relatively stable trends.
4. Correlation Between Stocks
    - While individual stocks move independently, some patterns appear to be moving together, hinting at possible correlations.
5. Presence of Negative Returns
    - Several instances of sharp negative dips indicate periods of losses, emphasizing the importance of risk assessment.
6. Market Events Impact
    - The sudden large fluctuations might be linked to market events or news that impacted multiple stocks simultaneously.

Stock Performance Insights

I. LINE GRAPH:

1.Volatility & Fluctuations:

- Yes Bank and DLF Limited exhibit the highest volatility, with extreme price swings.
- ITC Limited and Bharti Airtel show more stability with smaller fluctuations.

2.Periods of High Movement:

- Significant volatility is observed between index 150-250, with Yes Bank facing sharp declines and DLF Limited experiencing sudden upward movement.

3.Trend Analysis:

- No consistent uptrend or downtrend; stock movements are erratic, indicating a high-risk environment.

4.Comparative Performance:

- Tata Motors shows moderate volatility, while ITC Limited and Bharti Airtel remain relatively stable, suggesting lower risk.

Key Takeaways

- Risk Consideration: Yes Bank and DLF Limited appeal to high-risk investors, while ITC Limited and Bharti Airtel offer stability.
- Market Sensitivity: Sharp fluctuations indicate strong influence from external market events.
- Diversification Strategy: A balanced portfolio with both volatile and stable stocks is essential for effective risk management.

BOXPLOT:

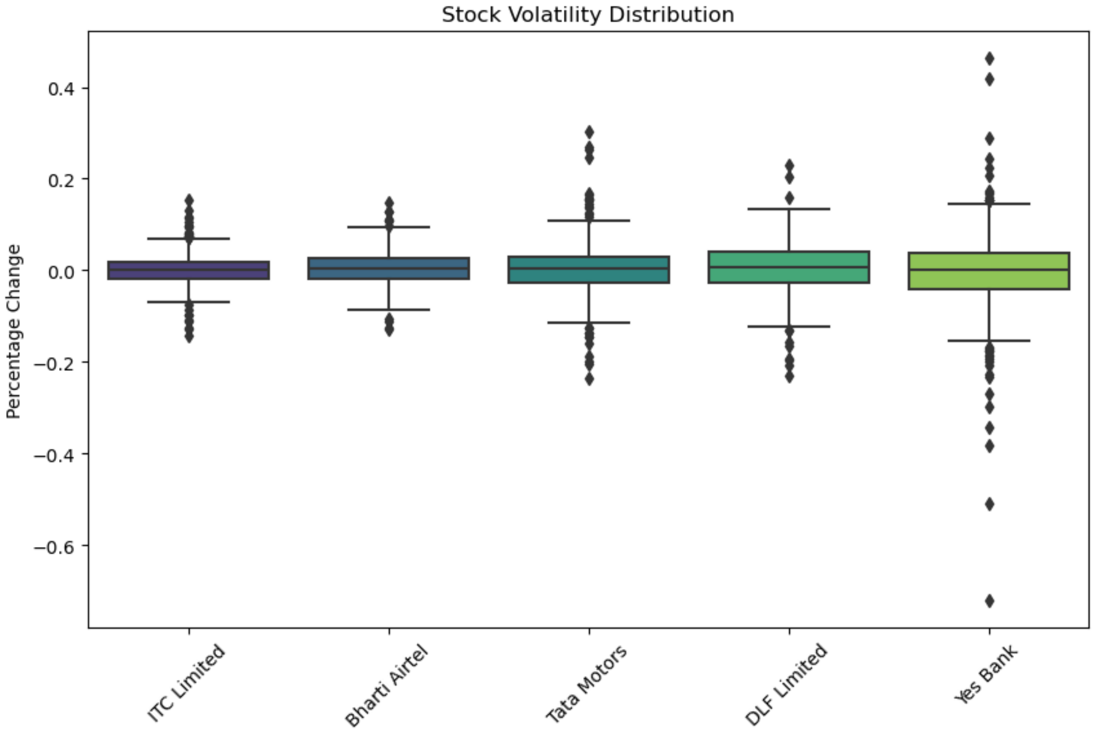


Figure B 15.Boxplot for Stock Volatility Distribution

Stock Volatility Insights:

BOXPLOT:

- High Volatility & Risk: Yes Bank and DLF Limited exhibit the highest volatility with wide fluctuations and extreme outliers, making them high-risk, high-reward stocks.
- Stable Performers: ITC Limited and Bharti Airtel show lower volatility with tightly packed distributions, indicating more predictable returns.
- Outlier Trends: Yes Bank and DLF Limited frequently experience sudden price spikes or crashes, while ITC and Bharti Airtel have fewer extreme movements.
- Return Consistency: Bharti Airtel and ITC Limited have a narrower interquartile range (IQR), suggesting more stable stock performance, whereas Tata Motors and DLF Limited have broader IQRs, indicating greater price variation.

Key Takeaways:

- High-Risk Investments: Yes Bank and DLF Limited present opportunities for high returns but come with significant volatility, requiring careful risk management.
- Stable Investment Options: ITC Limited and Bharti Airtel demonstrate lower variance, making them suitable for risk-averse investors seeking consistent performance.
- Moderate Risk Profile: Tata Motors offers a balanced approach, exhibiting a mix of stability and moderate volatility.
- Diversification Strategy: A well-structured portfolio should include both high- and low-volatility stocks to mitigate risk while optimizing returns.

7.3 Cumulative Stock Return Volatility Over Time

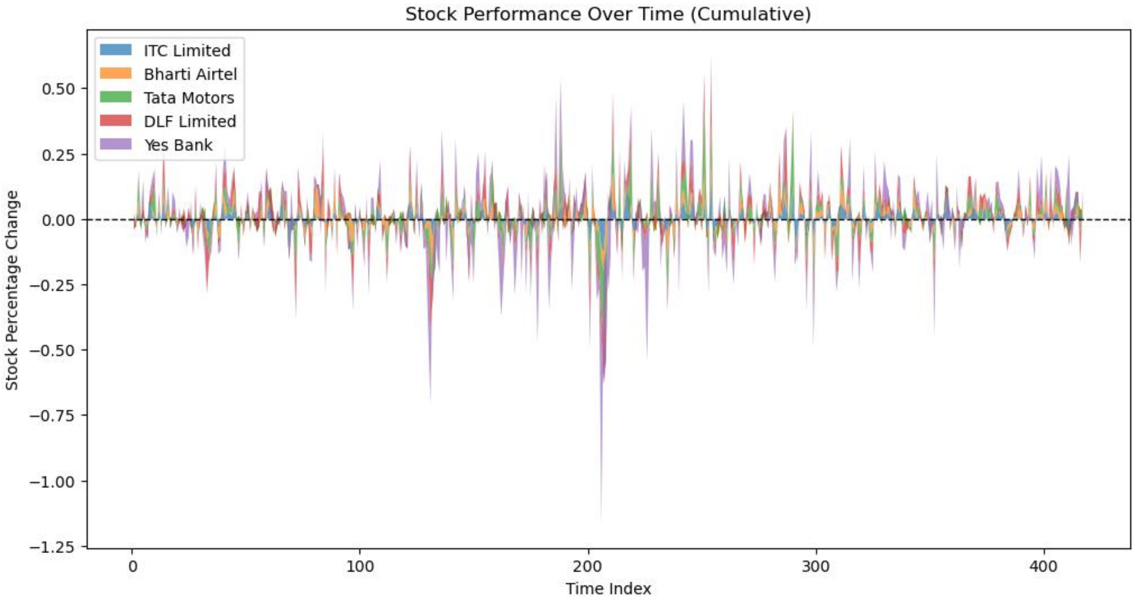


Figure B 16.Cumulative Stock Return Volatility Over Time

Observations from Stock Performance Over Time (Cumulative)

1. High Volatility: The stock returns exhibit significant fluctuations over time, with multiple peaks and troughs, indicating periods of high volatility.
2. Extreme Negative Returns: Certain time points show sharp downward spikes, particularly in Yes Bank and DLF Limited, suggesting possible external shocks or market corrections.
3. Mean Reversion Tendency: Despite fluctuations, the stock percentage change tends to revert toward the zero line over time, suggesting a long-term stabilization.
4. Synchronized Movements: Stocks from different companies show similar movement patterns, indicating possible market-wide influences affecting all stocks.

5. Gradual Stabilization: Towards the later time periods, fluctuations seem to reduce, suggesting either reduced volatility or a more stable market phase.

7.4 Stock Return Correlation Analysis

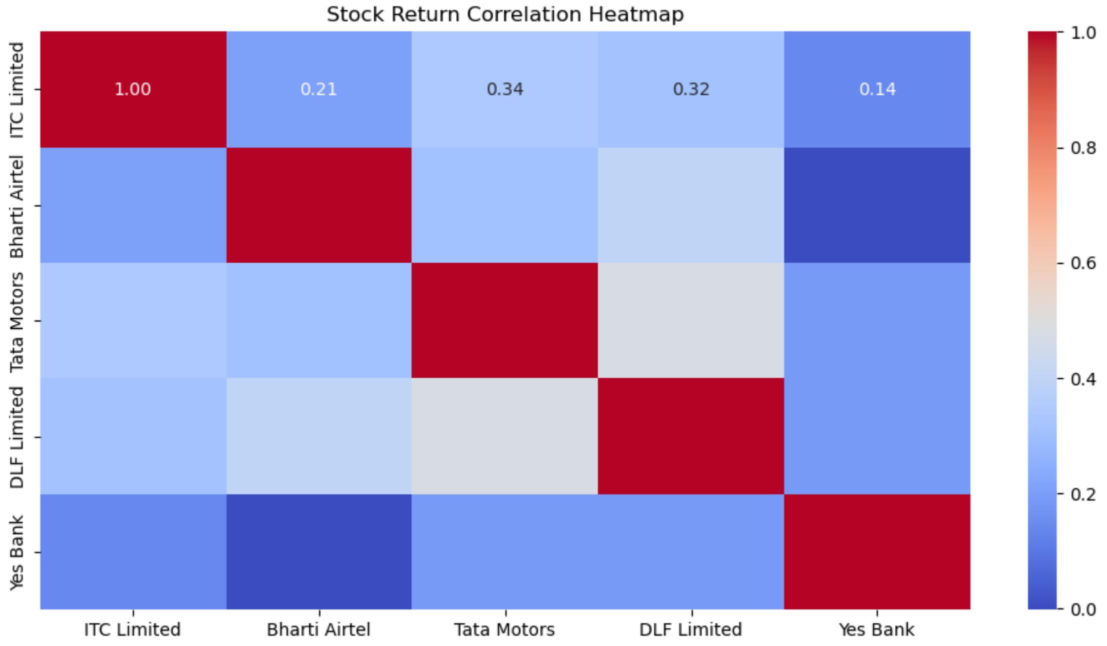


Figure B 17.Stock Return Correlation Analysis

Observations:

1. Self-Correlation – Each stock has a correlation of 1.00 with itself, as expected.
2. Moderate Correlations – Tata Motors shows moderate correlation (0.34) with ITC Limited and (0.42) with DLF Limited, suggesting some level of co-movement.
3. Low Correlations – Yes Bank has a weak correlation with most other stocks, indicating its returns behave independently compared to others.
4. Weakest Relationship – Bharti Airtel and Yes Bank show the lowest correlation (~0.00), implying no significant relationship between their stock returns.

7.5 Stock Return Distribution Analysis

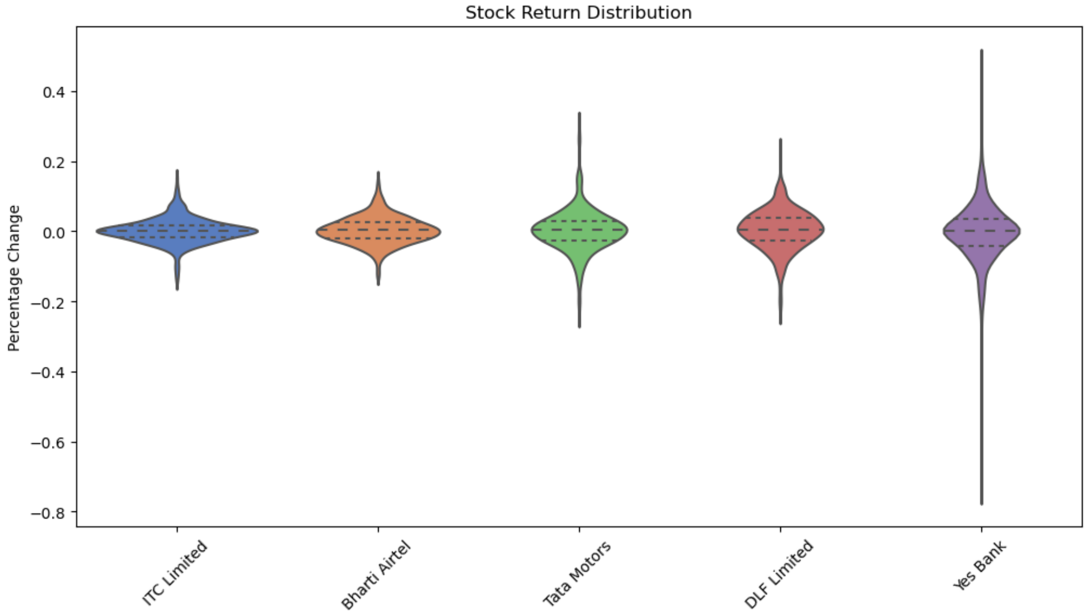


Figure B 18.Violin Plot for Stock Return Distribution

Observations:

1. Symmetric Distribution – Most stocks show a nearly symmetric return distribution, indicating balanced positive and negative returns.
2. Higher Volatility in Yes Bank – Yes Bank exhibits the widest spread, suggesting it has the highest volatility among the stocks.
3. Tighter Distribution in ITC Limited & Bharti Airtel – These stocks have a narrower range, indicating lower volatility and relatively stable returns.
4. Presence of Extreme Returns – Some stocks, like Tata Motors and DLF Limited, show extended tails, suggesting occasional extreme gains or losses.
5. Mean & Median Comparison – The central region of the violins (dashed lines) suggests that median returns are close to zero, indicating an overall balanced return profile.

7.6 Rolling Volatility Analysis of Stocks

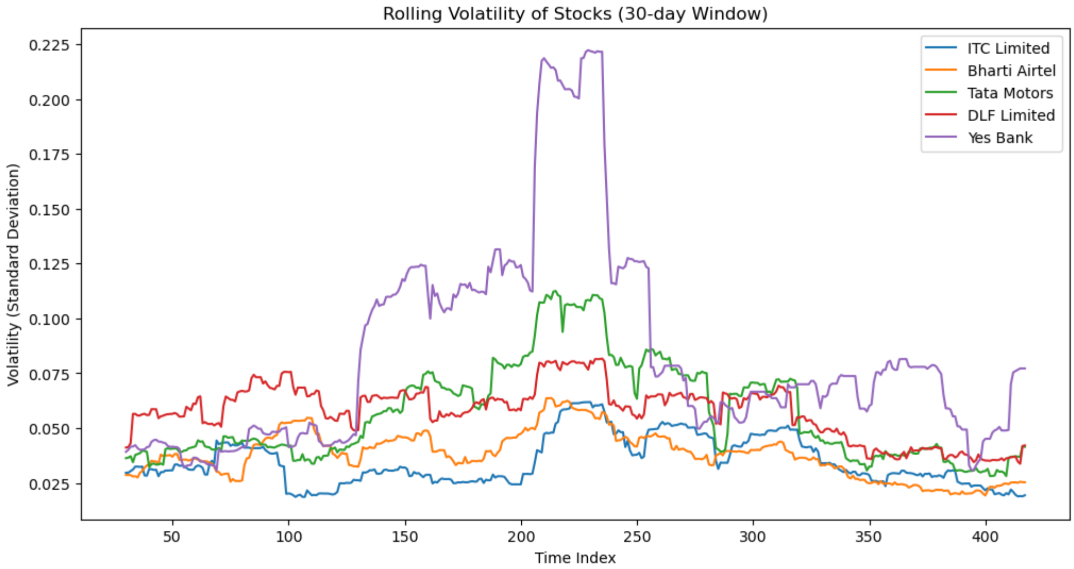
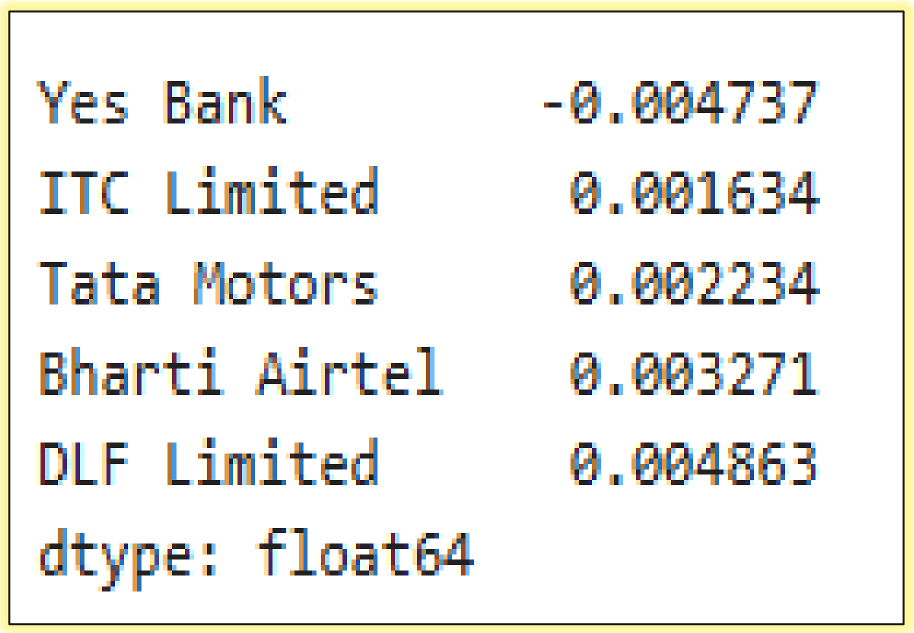


Figure B 19. Rolling Volatility Analysis of Stocks

Observations:

1. High Volatility in Yes Bank – Yes Bank shows the highest volatility, with significant spikes, indicating unstable stock price movements.
2. Tata Motors and DLF Limited Show Moderate Volatility – Both stocks exhibit fluctuations but remain relatively stable compared to Yes Bank.
3. ITC Limited and Bharti Airtel Have the Least Volatility – These stocks show the lowest standard deviation in returns, suggesting they are less risky.
4. Volatility Peaks Around Mid-Period – A notable increase in volatility is observed around the mid-point, possibly due to external market factors or earnings reports.
5. Volatility Declines Towards the End – Most stocks experience a reduction in volatility towards the end, indicating market stabilization.

8. AVERAGE RETURNS



```
Yes Bank          -0.004737
ITC Limited        0.001634
Tata Motors        0.002234
Bharti Airtel      0.003271
DLF Limited        0.004863
dtype: float64
```

Figure B 20.Average Returns

Average Daily Returns of Selected Stocks

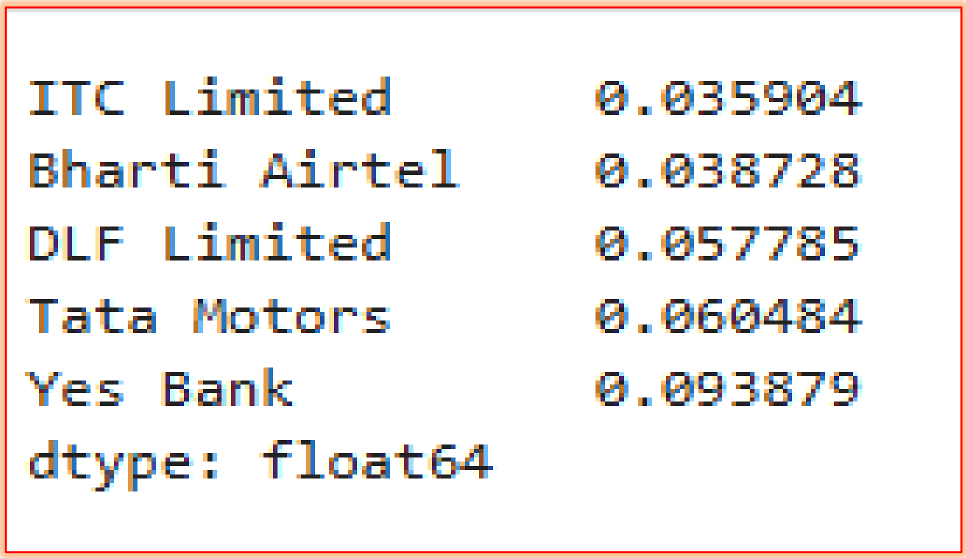
Inference:

1. Yes Bank Shows Negative Returns – Yes Bank has an average daily return of -0.004737, indicating a decline in stock value over time.
2. Positive Returns for Other Stocks – ITC Limited, Tata Motors, Bharti Airtel, and DLF Limited have positive average daily returns, suggesting growth potential.
3. DLF Limited Has the Highest Positive Return – With 0.004863, DLF Limited has the highest daily return, making it the best performer among the given stocks.
4. ITC Limited Has the Lowest Positive Return – ITC Limited's return is 0.001634, indicating a more stable but slower growth compared to others.
5. Risk and Reward Considerations – While DLF Limited offers higher returns, its volatility should be considered. Yes Bank, on the other hand, may pose a higher risk due to consistent negative returns.

Key Takeaways:

- Yes Bank is underperforming with negative returns.
- DLF Limited has the highest returns, showing strong growth.
- ITC, Tata Motors, and Bharti Airtel have moderate positive returns.
- Consider volatility before investing.
- Diversification is key to balancing risk and reward.

9. VOLATILITY



```
ITC Limited       0.035904
Bharti Airtel     0.038728
DLF Limited       0.057785
Tata Motors       0.060484
Yes Bank          0.093879
dtype: float64
```

Figure B 21.Volatility

This represents the volatility (standard deviation) of daily percentage changes for each stock. Higher values indicate greater price fluctuations and higher risk.

Inference on Volatility:

- Yes Bank has the highest volatility (0.093879), indicating high risk and potential for large price swings.
- Tata Motors and DLF Limited also show relatively high volatility, suggesting moderate risk.
- ITC Limited and Bharti Airtel have lower volatility, implying more stable price movements.

Key Takeaways:

- Higher volatility = Higher risk & potential returns.
- Yes Bank is the riskiest investment.
- ITC Limited and Bharti Airtel are more stable choices.
- Risk tolerance should guide investment decisions.

10. VISUALIZING RETURNS AND VOLATILITY

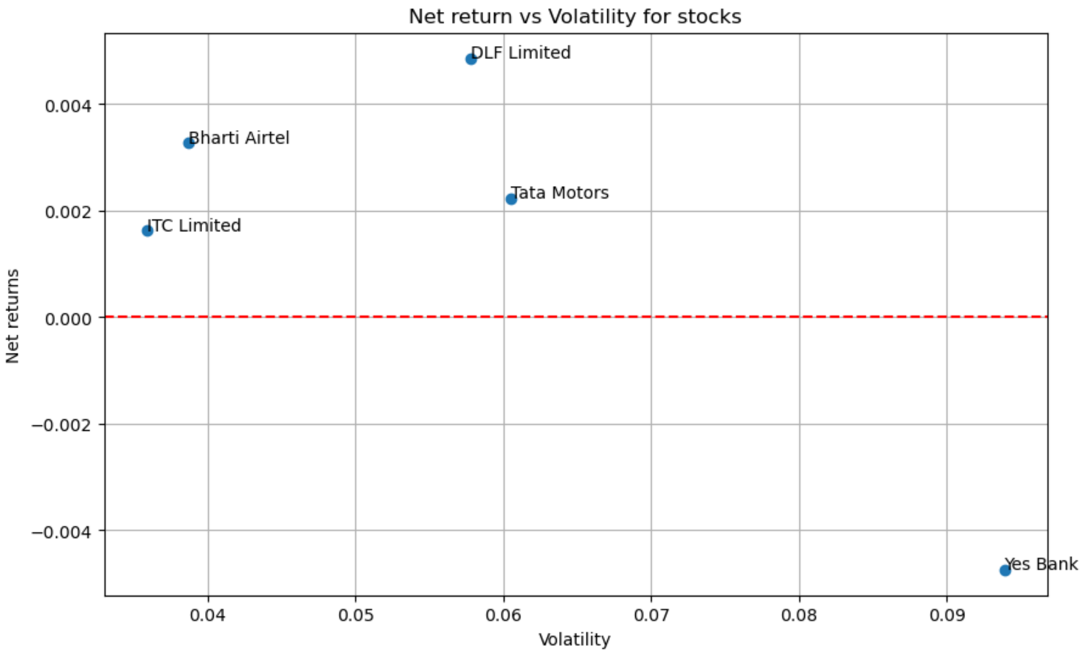


Figure B 22. Scatter Plot. for Net Return vs. Volatility

Insights from the Net Return vs. Volatility Scatter Plot

1. Risk-Return Relationship
   - DLF Limited has the highest net return among the stocks while maintaining moderate volatility.
   - Yes Bank has the highest volatility but also the lowest net return, making it a high-risk, low-reward stock.
   - ITC Limited and Bharti Airtel have the lowest volatility, indicating stability with moderate positive returns.
2. Stocks Above and Below Zero Returns
   - Yes Bank is the only stock with negative returns, meaning it has underperformed despite high volatility.
   - All other stocks (DLF Limited, Tata Motors, Bharti Airtel, ITC Limited) have positive returns, making them preferable investments in this dataset.

Key Takeaways

- Low-risk, moderate return choices: ITC Limited & Bharti Airtel (stable investments).
- High-risk, high return potential: DLF Limited (good return with slightly higher volatility).
- Risky investment: Yes Bank (high volatility with negative returns).

- Balanced option: Tata Motors (moderate volatility with positive returns).

11. ACTIONABLE INSIGHTS & RECOMMENDATIONS

1. Risk-Return Analysis & Investment Strategy

- Low Volatility, Consistent Returns (ITC Limited & Bharti Airtel)
    - Suitable for conservative investors and long-term portfolio stability.
    - Lower risk makes them ideal for steady income generation.
- Moderate Volatility, Balanced Risk-Reward (Tata Motors & DLF Limited)
    - Offers a balance between risk and return, making them suitable for growth-focused portfolios.
    - Recommended for medium-term investment strategies.
- High Volatility, Negative Returns (Yes Bank)
    - Carries significant risk due to high fluctuations and negative returns.
    - Recommended only for speculative investors with a high-risk appetite.

2. Portfolio Diversification & Risk Mitigation

- A well-balanced portfolio should include a mix of low, moderate, and high-risk stocks to optimize returns.
- Avoid heavy exposure to highly volatile stocks like Yes Bank unless risk mitigation strategies (such as stop-loss mechanisms) are in place.
- Sector-based allocation can help reduce risk—ITC Limited (FMCG) and Bharti Airtel (Telecom) provide stability, while Tata Motors (Automobile) and DLF Limited (Real Estate) offer growth potential.

3. Strategic Recommendations for Business Decision-Making

- Invest in low-volatility stocks to ensure stability in uncertain market conditions.
- Monitor macroeconomic trends (e.g., interest rate changes, industry growth) that could impact stock performance.
- Use technical and fundamental analysis to reassess investment positions periodically and adjust strategies accordingly.

4. Conclusion & Next Steps

- Businesses and investors should prioritize stability while leveraging growth opportunities in moderate-risk stocks.
- Continuous monitoring and data-driven decision-making will be critical to maximizing returns while managing risk.

- Further analysis of industry trends and economic factors can refine investment strategies for long-term success.