Business Report - 11

PG Program in Data Science and Business Analytics

submitted by

Sangram Keshari Patro

BATCH:PGPDSBA.O.AUG24.B

Contents

List of Figures

List of Tables

Part I
PART-A

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.

2 Data Description

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

2.1 Data dictionary

- **Networth Next Year**: Net worth of the customer in the next year.
- **Total assets**: Total assets of the customer.
- **Net worth**: Net worth of the customer for the present year.
- **Total income**: Total income of the customer.
- **Change in stock**: Current stock value − Last trading day stock value.
- **Total expenses**: Total expenses incurred by the customer.
- **Profit after tax (PAT)**: 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**: $\frac{\text{PBDITA}}{\text{Total income}} \times 100$
- **PBT as % of total income**: $\frac{\text{PBT}}{\text{Total income}} \times 100$
- **PAT as % of total income**: $\frac{\text{PAT}}{\text{Total income}} \times 100$
- **Cash profit as % of total income**: $\frac{\text{Cash profit}}{\text{Total income}} \times 100$
- **PAT as % of net worth**: $\frac{\text{PAT}}{\text{Net worth}} \times 100$
- **Sales**: Sales made 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 and provisions**: Current liabilities of the customer.
- **Deferred tax liability**: Future income tax the customer will pay due to current transactions.
- **Shareholders funds**: Equity amount in a company belonging to shareholders.
- **Cumulative retained profits**: Total cumulative profit retained by the customer.
- **Capital employed**: Current assets − Current liabilities.
- **TOL/TNW**: $\frac{\text{Total liabilities}}{\text{Total net worth}}$
- **Total term liabilities / Tangible net worth**: $\frac{\text{Short-term liabilities+Long-term liabilities}}{\text{Tangible net worth}}$
- **Contingent liabilities / Net worth (%)**: $\frac{\text{Contingent liabilities}}{\text{Net worth}} \times 100$
- **Contingent liabilities**: Liabilities arising from uncertain events.
- **Net fixed assets**: Purchase price of all fixed assets.
- **Investments**: Total invested amount.
- **Current assets**: Assets expected to be converted to cash within a year.
- **Net working capital**: Current assets − Current liabilities.
- **Quick ratio (times)**: $\frac{\text{Total cash}}{\text{Current liabilities}}$
- **Current ratio (times)**: $\frac{\text{Current assets}}{\text{Current liabilities}}$
- **Debt to equity ratio (times)**: $\frac{\text{Total liabilities}}{\text{Shareholder equity}}$
- **Cash to current liabilities (times)**: $\frac{\text{Total liquid cash}}{\text{Current liabilities}}$
- **Cash to average cost of sales per day**: $\frac{\text{Total cash}}{\text{Average cost of sales}}$
- **Creditors turnover**: $\frac{\text{Net credit purchase}}{\text{Average trade creditors}}$
- **Debtors turnover**: $\frac{\text{Net credit sales}}{\text{Average accounts receivable}}$
- **Finished goods turnover**: $\frac{\text{Annual sales}}{\text{Average inventory}}$
- **WIP turnover**: $\frac{\text{Cost of goods sold for a period}}{\text{Average inventory for that period}}$
- **Raw material turnover**: $\frac{\text{Cost of goods sold}}{\text{Average inventory for the same period}}$
- **Shares outstanding**: Number of issued shares − Shares held in the company.
- **Equity face value**: Cost of equity at the time of issuance.
- **EPS**: $\frac{\text{Net income}}{\text{Total number of outstanding shares}}$
- **Adjusted EPS**: $\frac{\text{Adjusted net earnings}}{\text{Weighted average number of common shares outstanding}}$
- **Total liabilities**: Sum of all types of liabilities.
- **PE on BSE**: $\frac{\text{Current stock price}}{\text{Earnings per share}}$

3 Project Summary and Outline

1. Understanding the Dataset

Dataset Overview: We are working with a comprehensive dataset comprising various financial metrics from the balance sheets of different companies. Key variables include total assets, net worth, total income, expenses, and more.

2. Exploratory Data Analysis (EDA)

Initial Data Inspection:

- **Purpose**: To understand the structure and quality of the dataset.
- **Actions**: Load the dataset and inspect its structure, data types, and summary statistics.
- **Outcome**: Identified the need for handling missing values and outliers.

3. Binary Target Variable Creation

Default Variable: Create a binary target variable "Default" derived from "Networth Next Year":

- **Not Likely to Default**: If $NetworthNextYear > 0$, label as 0.
- **Likely to Default**: If $NetworthNextYear \leq 0$, label as 1.

4. Handling Missing Values

Checking for Missing Values:

- **Purpose**: To ensure data completeness and quality.
- **Actions**: Check for missing values across the dataset and visualize them using heatmaps.
- **Outcome**: Identified patterns and proportions of missing data.

5. Identifying and Handling Outliers

Identifying Outliers:

- **Purpose**: To detect anomalies that could skew the analysis.
- **Actions**: Calculate quartiles $Q1$, $Q3$, interquartile range (IQR), and determine the upper (UL) and lower limits (LL).
- **Outcome**: Identified outliers beyond UL and LL.

Handling Outliers:

- **Purpose**: To maintain data integrity without distorting true values.
- **Actions**: Convert outliers below LL and above UL to missing values instead of capping or removing them.
- **Outcome**: Ensured accurate imputation of missing values later.

6. Assessing Features with Missing Data

Feature Analysis:

- **Purpose**: To identify and handle features with significant missing data.
- **Actions**: Visualize the proportion of missing values for each feature.
- **Outcome**: Drop features with more than 30% missing values.

7. Evaluating Missing Data in Records

Row-Level Missing Data:

- **Purpose**: To ensure completeness of records.
- **Actions**: Calculate the proportion of missing data in each row.
- **Outcome**: Analyze rows with excessive missing values for suitability.

8. Imputing Missing Values with KNN Imputation

Using KNN Imputer:

- **Purpose**: To fill missing values accurately.
- **Actions**: Scale the data, apply KNN imputer.
- **Outcome**: Improved data completeness for modeling.

9. Logistic Regression Modelling

Feature Selection with RFE:

- **Purpose**: To identify significant features.
- **Actions**: Use Recursive Feature Elimination (RFE).
- **Outcome**: Focused on impactful variables.

Optimizing Classification Threshold:

- **Purpose**: Improve default prediction accuracy.
- **Actions**: Use ROC curve, AUC score, and Youden's J statistic to find optimal threshold.
- **Outcome**: Improved model decision-making.

10. Random Forest Modelling

4 Data Overview

4.1 Importing necessary libraries and the dataset

The dataset is printed. It has 4256 rows & 51 columns. The dataset consists solely of numeric data types float values (decimals) and integers.

4.2 Structure and type of data

Data is explored further. The dataset is free from duplicate rows and contains no null values.

```
RangeIndex: 4256 entries, 0 to 4255
Data columns (total 51 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Num                                   4256 non-null   int64
 1   Networth Next Year                    4256 non-null   float64
 2   Total assets                          4256 non-null   float64
 3   Net worth                             4256 non-null   float64
 4   Total income                          4025 non-null   float64
 5   Change in stock                       3706 non-null   float64
 6   Total expenses                        4091 non-null   float64
 7   Profit after tax                      4102 non-null   float64
 8   PBDITA                                4102 non-null   float64
 9   PBT                                   4102 non-null   float64
 10  Cash profit                           4102 non-null   float64
 11  PBDITA as % of total income           4177 non-null   float64
 12  PBT as % of total income              4177 non-null   float64
 13  PAT as % of total income              4177 non-null   float64
 14  Cash profit as % of total income      4177 non-null   float64
 15  PAT as % of net worth                 4256 non-null   float64
 16  Sales                                 3951 non-null   float64
 17  Income from fincial services          3145 non-null   float64
 18  Other income                          2700 non-null   float64
 19  Total capital                         4251 non-null   float64
...
 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 1: Table depicting the datatype and Non-Null values in each column.

4.3 Statistical summary

index	count	mean	std	min
Networth Next Year	4256.0	1344.7408834586467	15936.743168126934	-74265.6
Total assets	4256.0	3573.6171522556388	30074.443435446017	0.1
Net worth	4256.0	1351.9496005639098	12961.311651076401	0.0
Total income	4025.0	4688.1897888198755	53918.946606361846	0.0
Change in stock	3706.0	43.70248246087426	436.9150483634242	-3029.4
Total expenses	4091.0	4356.301099975556	51398.08712166589	-0.1
Profit after tax	4102.0	295.05058508044857	3079.902071219754	-3908.3
PBDITA	4102.0	605.9406387123	5646.230633105839	-440.7
PBT	4102.0	410.2590443686007	4217.415306894319	-3894.8
Cash profit	4102.0	408.2674792784008	4143.926393233108	-2245.7
PBDITA as % of total income	4177.0	3.179892267177401	172.2565599458	-6400.0
PBT as % of total income	4177.0	-18.196830260952837	419.9110908282069	-21340.0
PAT as % of total income	4177.0	-20.033665531003112	423.57618884890866	-21340.0
Cash profit as % of total income	4177.0	-9.021278429494853	299.9574342677693	-15020.0
PAT as % of net worth	4256.0	10.167861842105262	61.53240131963372	-748.72
Sales	3951.0	4645.684535560618	53080.903295686825	0.1
Income from fincial services	3145.0	81.36006359300477	1042.7586776908101	0.0
Other income	2700.0	55.9528888888888886	1178.4152610234298	0.0
Total capital	4251.0	224.55765702187722	1684.9512874800764	0.1
Reserves and funds	4158.0	1210.5619288119287	12816.22922144698	-6525.9
Borrowings	3825.0	1176.2480784313727	8581.248920508857	0.1
Current liabilities & provisions	4146.0	960.6314278822962	9140.536134642918	0.1
Deferred tax liability	2887.0	234.49511603740908	2106.2531589774303	0.1
Shareholders funds	4256.0	1376.4867246240601	13010.691155020468	0.0
Cumulative retained profits	4211.0	937.1819757777249	9853.096090904335	-6534.3
Capital employed	4256.0	2433.61757518797	20496.40388122408	0.0
TOL/TNW	4256.0	4.025343045112782	20.879089403333055	-350.48
Total term liabilities / tangible net worth	4256.0	1.8542880639097745	15.875057273639174	-325.6
Contingent liabilities / Net worth (%)	4256.0	55.707499999999996	369.16566959919606	0.0
Contingent liabilities	2854.0	948.5522424667132	12056.737584501683	0.1
Net fixed assets	4124.0	1209.486517943744	12502.396635207899	0.0
Investments	2541.0	721.8658795749704	6793.8598669429575	0.0
Current assets	4176.0	1350.3600095785441	10155.572745246836	0.1
Net working capital	4219.0	162.87423560085327	3182.0299600334306	-63839.0
Quick ratio (times)	4151.0	1.4973548542519877	9.32751895311905	0.0
Current ratio (times)	4151.0	2.2573982172970366	12.478285996957542	0.0
Debt to equity ratio (times)	4256.0	2.8715625	15.599968041407012	0.0
Cash to current liabilities (times)	4151.0	0.5284196579137558	4.7963417292604165	0.0
Cash to average cost of sales per day	4156.0	145.1579258902791	2521.991811304902	0.0
Creditors turnover	3865.0	16.81225873221216	75.67491506625565	0.0
Debtors turnover	3871.0	17.929028674761046	90.16443462876313	0.0
Finished goods turnover	3382.0	84.3699881726789	562.6373589384485	-0.09
WIP turnover	3492.0	28.684513172966778	169.65091519834073	-0.18
Raw material turnover	3828.0	17.7339263322884	343.12586401359914	-2.0
Shares outstanding	3446.0	23764909.555426583	170979041.3298719	-2147483647.0
Equity face value	3446.0	-1094.8286709228091	34101.35864384831	-999998.9
EPS	4256.0	-196.21746710526315	13061.953424861198	-843181.82
Adjusted EPS	4256.0	-197.52760808270676	13061.929511651782	-843181.82
Total liabilities	4256.0	3573.6171522556388	30074.443435446017	0.1
PE on BSE	1629.0	55.46228974831185	1304.4452959040636	-1116.64

5 Exploratory Data Analysis

5.1 Univariate Analysis

5.1.1 Numerical columns

- **Networth Next Year**

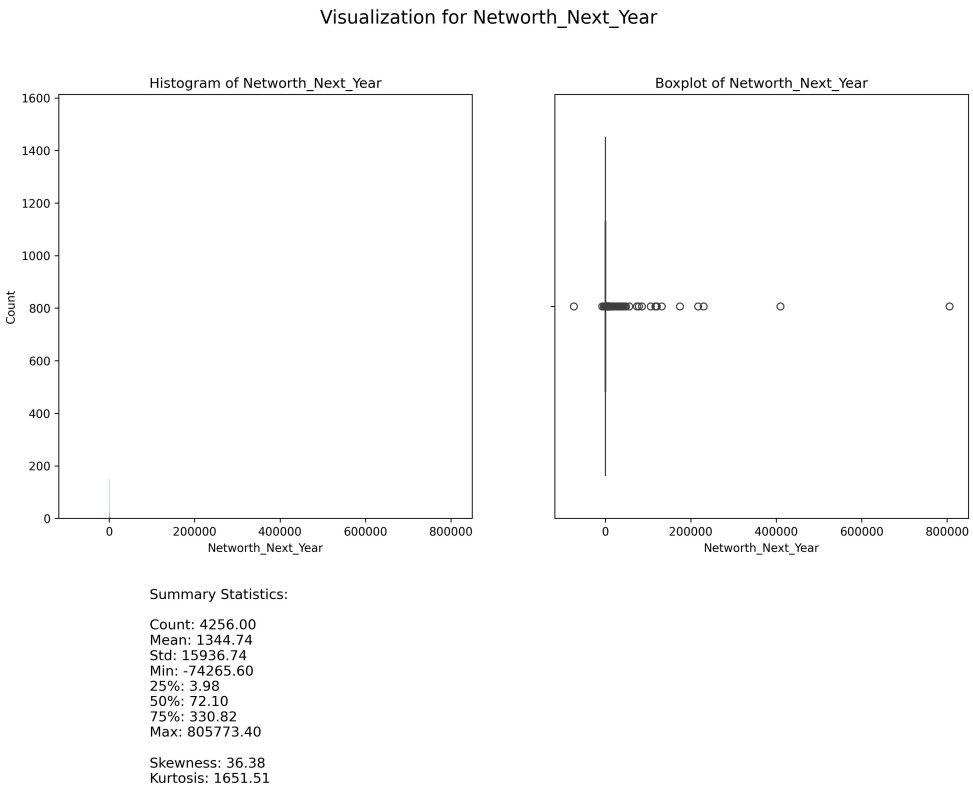


Figure 3: Histogram and boxplot of 'Networth Next Year' column

Analysis of "Networth Next Year"

Key Observations

- Extreme range: $-74,266$ to $805,773$ shows wide financial disparities
- High volatility: Std. dev. $(15,937) \gg$ mean $(1,345)$
- Negative values indicate potential default risks
- Right-skewed: Median (72) $\ll$ mean; skewness = 36.38
- Heavy-tailed: Kurtosis = $1,651.51$ with numerous outliers
- 50% of data between 4 and 331

Business Recommendations

- Segment analysis by net worth tiers for targeted strategies
- Implement early warning systems for negative net worth firms
- Outlier investigation for extreme high-value cases

- Risk-adjusted pricing for companies showing financial stress
- Focus on median (not mean) for typical company benchmarks
- **Total Assets**

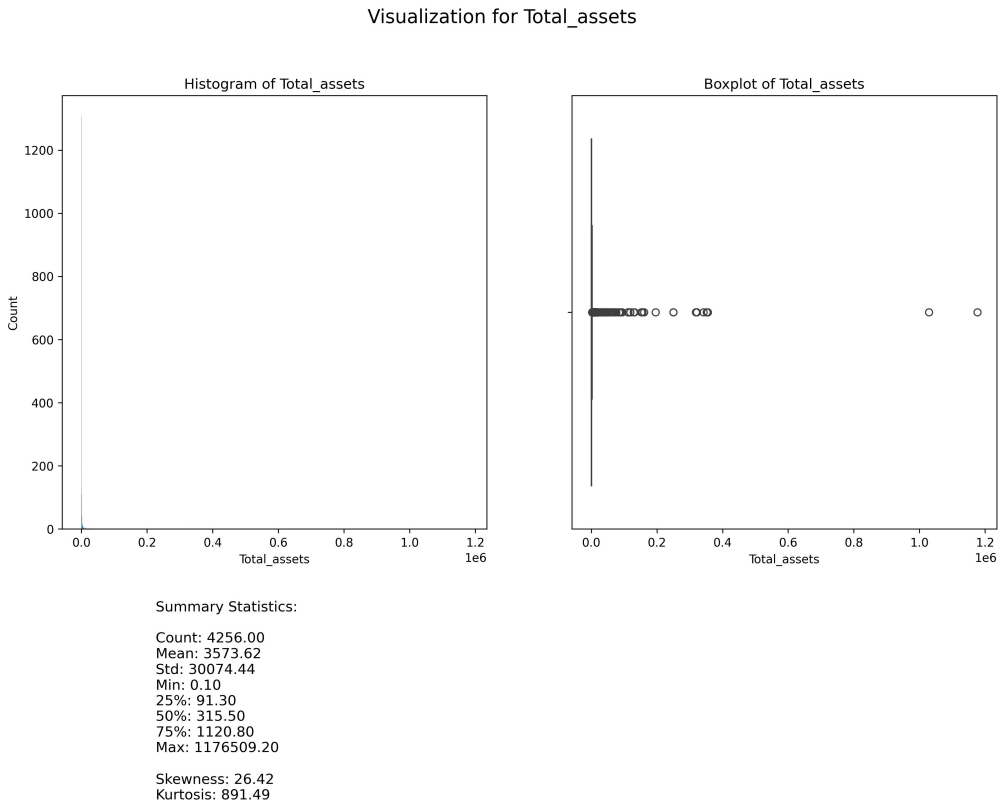


Figure 4: Histogram and boxplot of 'Total Assets' column

Analysis of "Total Assets"

Key Observations

- Extreme range: 0.00 to $1,176,509$ shows vast asset disparities
- High volatility: Std. dev. $(30,074) \gg$ mean $(3,574)$
- Zero values suggest data issues or extreme distress cases
- Right-skewed: Median (316) $\ll$ mean; skewness = 26.42
- Heavy-tailed: Kurtosis = 891.49 with numerous outliers
- 50% of data between 91 and $1,121$

Business Recommendations

- Verify data quality for zero-asset entries
- Segment companies by asset tiers for differentiated strategies
- Investigate ultra-high-asset outliers for M&A opportunities
- Use median values for typical company benchmarking

- Develop separate risk models for asset-light vs. asset-heavy firms
- **Change in Stock**

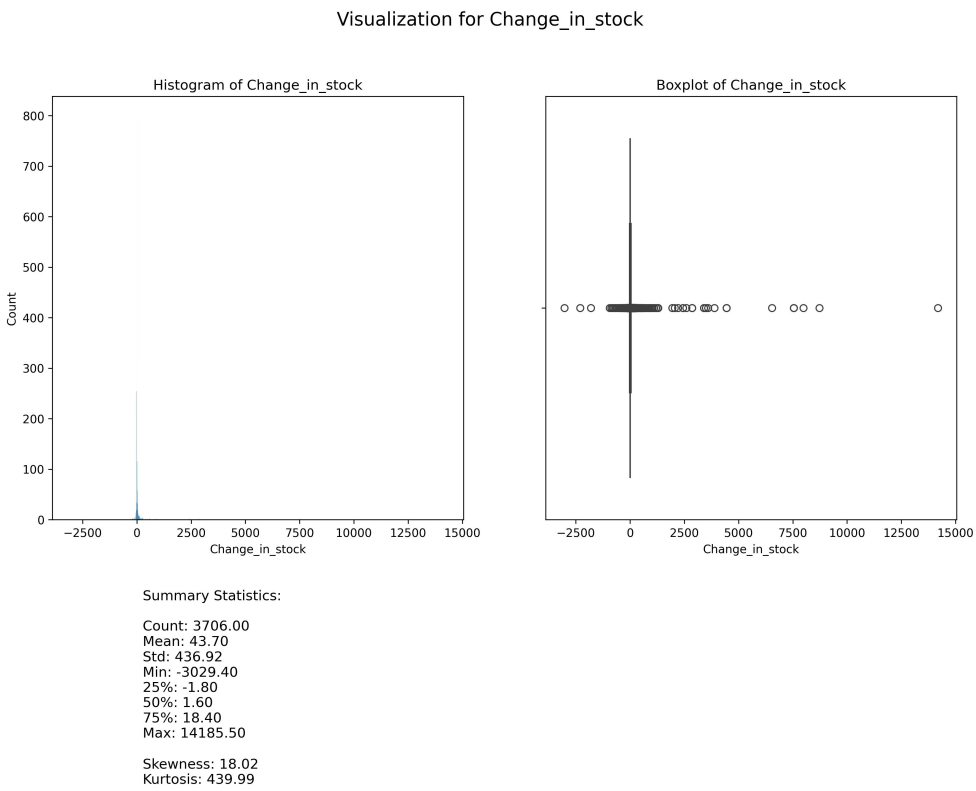


Figure 5: 'Change in Stock' column

Analysis of "Change in Stock"

Key Observations

- Extreme range: $-3,029$ to $14,186$ shows volatile stock movements
- High volatility: Std. dev. (437) $\gg$ mean (44.00)
- Negative values indicate significant stock declines
- Right-skewed: Median (2) $\ll$ mean; skewness = 18.02
- Heavy-tailed: Kurtosis = 439.99 with extreme outliers
- 50% of changes between -2 and 18

Business Recommendations

- Monitor companies with changes < -500 for distress signals
- Investigate extreme positive outliers for potential growth opportunities
- Use median values for typical change benchmarks
- Segment companies into change categories for targeted analysis
- Develop separate models for stable vs. volatile stock performers

- **Profit After Tax**

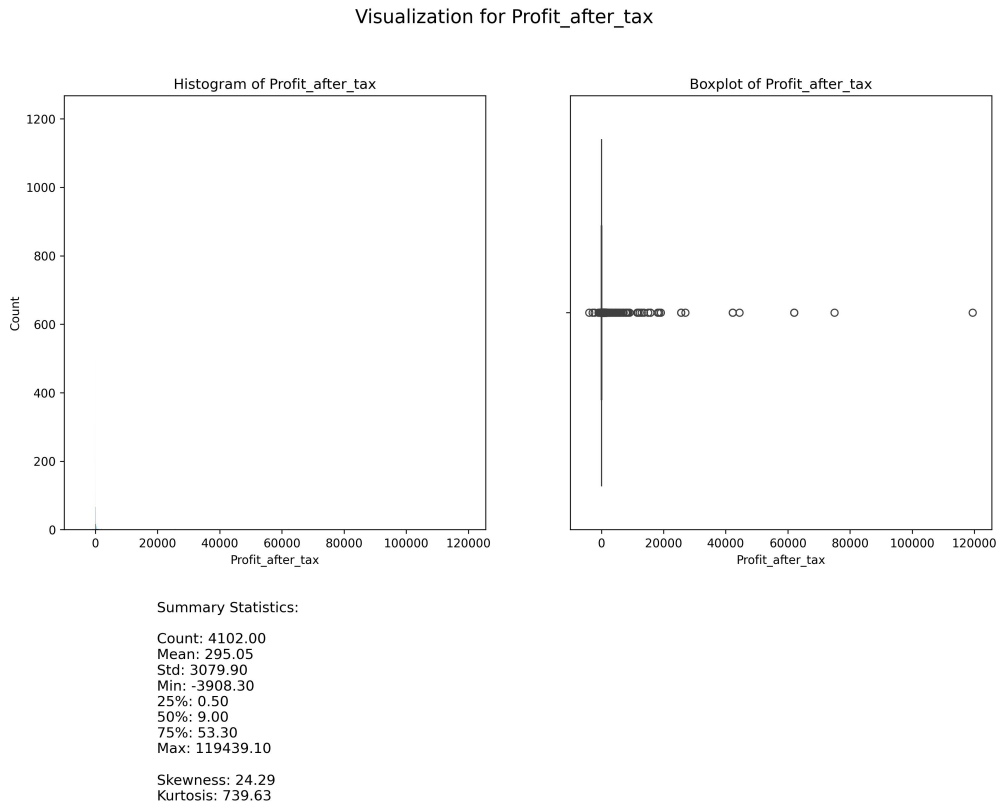


Figure 6: 'Profit After Tax' column

Analysis of "Profit After Tax"

Key Observations

- Extreme range: $-3,908$ to $119,439$ shows vast profitability disparity
- High volatility: Std. dev. $(3,080) \gg$ mean (295)
- Negative values indicate significant losses for some firms
- Right-skewed: Median (9) $\ll$ mean; skewness = 24.29
- Heavy-tailed: Kurtosis = 739.63 with extreme positive outliers
- Majority clustered near low-profit range

Business Recommendations

- Investigate loss-making companies (< 0) for turnaround potential
- Analyze high-profit outliers ($> 50,000$) for best practices
- Use median values for typical company benchmarks
- Segment portfolio by profitability tiers
- Develop separate risk models for loss-makers vs high-performers

Analysis of "PBDITA"

Key Observations

- Extreme range: -441 to $208,576$ shows vast operational performance disparity
- High volatility: Std. dev. $(5,646) \gg$ mean (606)
- Negative values indicate operational losses for some firms
- Right-skewed: Median (37) $\ll$ mean; skewness = 24.12
- Heavy-tailed: Kurtosis = 717.37 with extreme positive outliers
- Majority show modest operational performance

Business Recommendations

- Flag companies with negative PBDITA for operational review
- Benchmark high-PBDITA outliers ($> 100,000$) for best practices
- Use median values for typical company comparisons
- Segment analysis by PBDITA performance quartiles
- Develop operational improvement programs for bottom performers
- Investigate accounting methods for extreme outlier cases
- **Cash Profit**

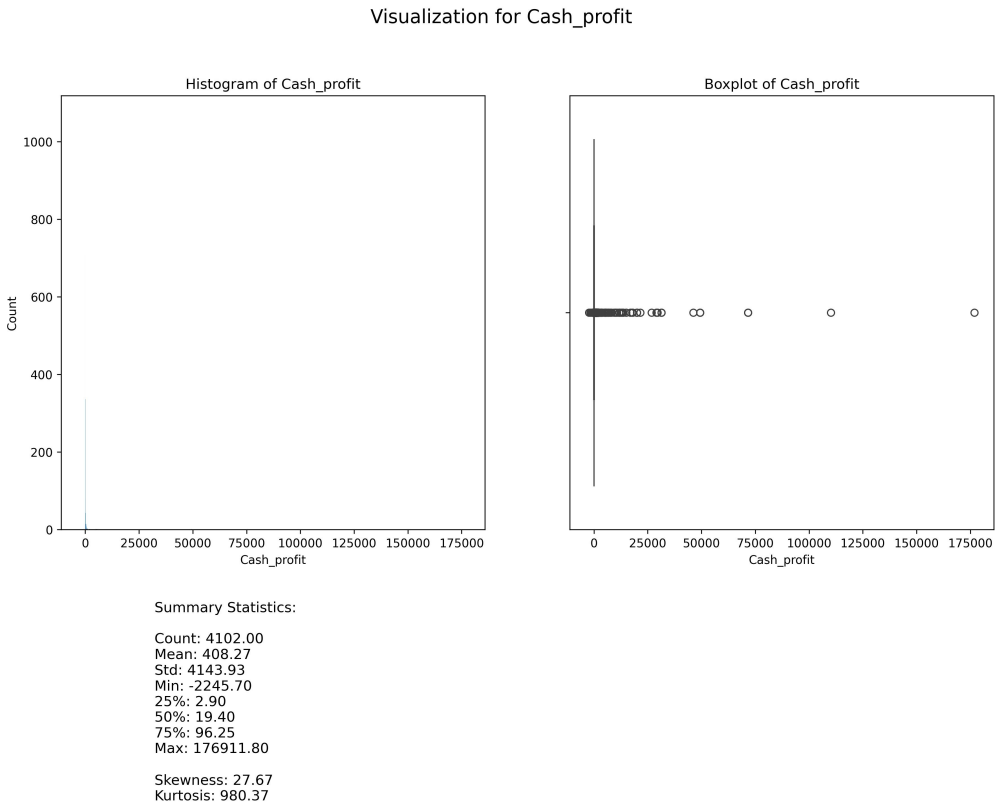


Figure 7: 'Cash Profit' column

Analysis of "Cash Profit"

Key Observations

- Extreme range: $-2,246$ to $176,912$ reveals severe profitability disparities
- High volatility: Std. dev. $(4,144) \gg$ mean (408)
- Negative cash profits signal severe financial distress cases
- Right-skewed: Median (19) $\ll$ mean; skewness = 27.67
- Extreme outliers: Kurtosis = 980.37 indicates exceptional performers
- 75% of firms show cash profits below 408

Business Recommendations

- Immediate review required for negative cash profit companies
- Cash flow monitoring for firms below 25th percentile (< 19)
- Study top 5% performers (cash profit $> 50,000$) for best practices
- Use median values for industry benchmarking
- Develop tiered cash management strategies
- Stress test financial models for extreme outlier scenarios
- **PBDITA as % of Total Income**

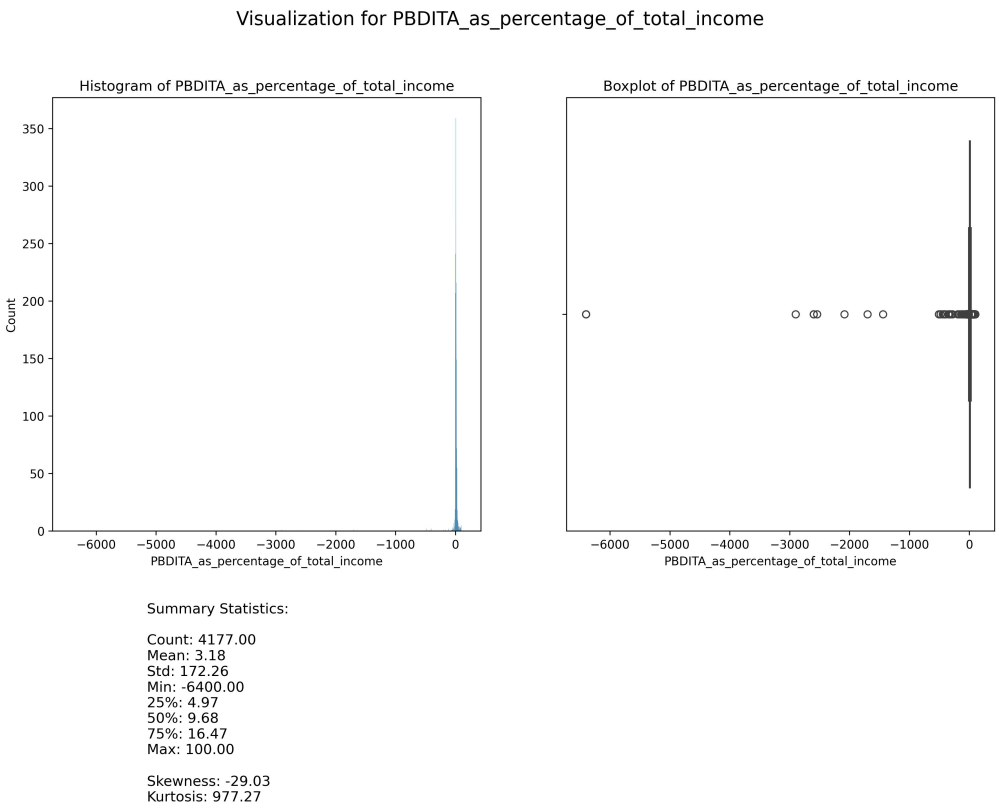


Figure 8: 'PBDITA as % of Total Income' column

Analysis of "PBDITA as % of Total Income"

Key Observations

- Extreme range: $-6,400\%$ to 100% shows alarming profitability variance
- High volatility: Std. dev. (172%) $\gg$ mean (3%)
- Negative ratios indicate severe operational inefficiencies
- Left-skewed: Median (10%) $>$ mean; skewness $= -29.03$
- Heavy left tail: Kurtosis $= 977.27$ with extreme negative outliers
- Majority cluster in single-digit percentages

Business Recommendations

- Immediate intervention for companies below -100% threshold
- Operational audit for firms with negative ratios
- Benchmark top performers ($> 50\%$) for efficiency best practices
- Use median (10%) as healthier performance benchmark
- Develop warning system for ratios approaching negative territory
- Investigate accounting anomalies for extreme negative cases

5.2 Bivariate Analysis

5.2.1 Numerical variables

- **Heatmap**

Bivariate Analysis Insights

Key Correlations

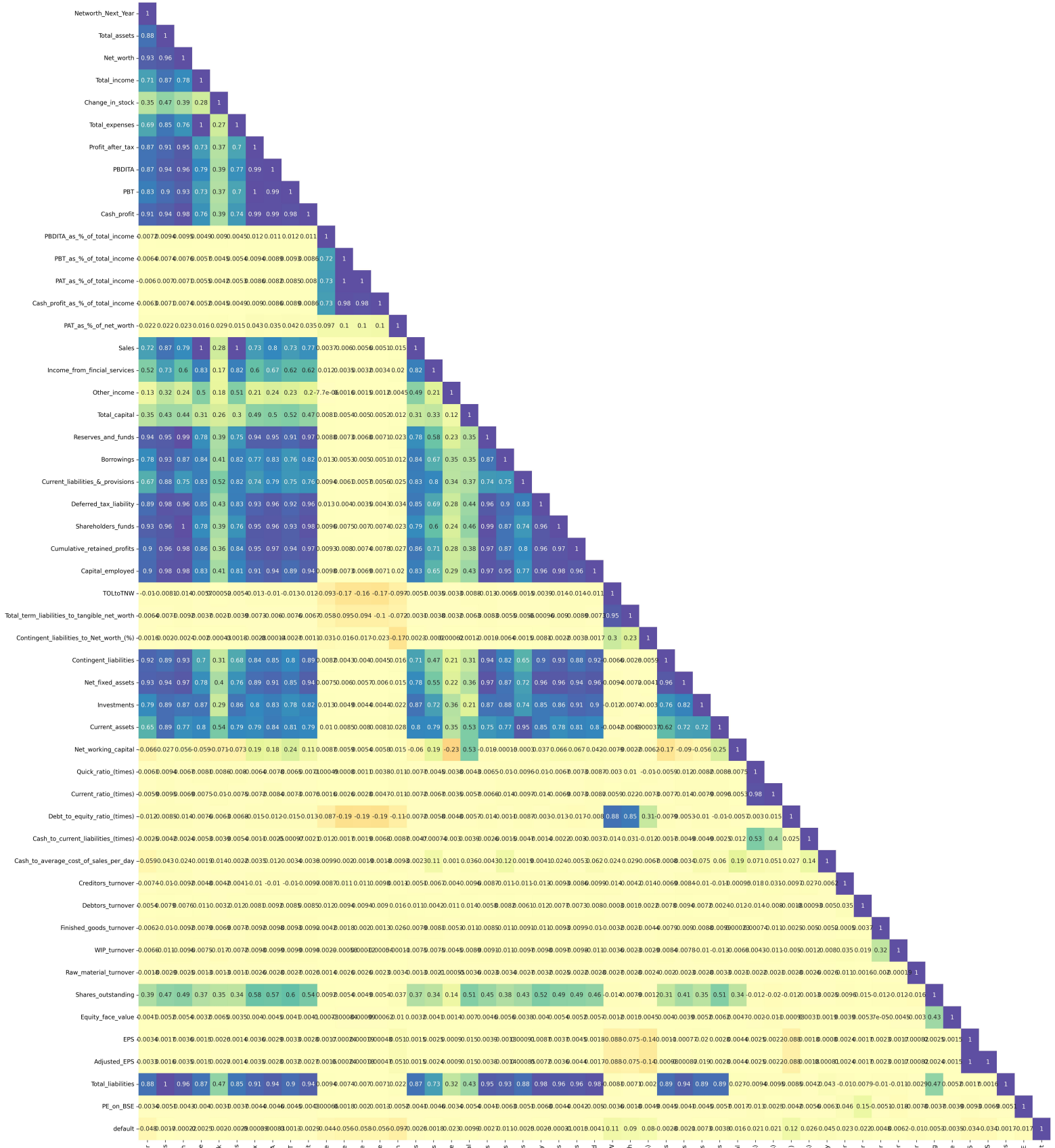
- **Net Worth & Assets (0.96):** Near-perfect alignment of growth
- **Income & Expenses (0.85):** Scaling naturally increases costs
- **PAT & PBDITA (0.99):** Operational and net profits move identically

Critical Negative Relationships

- **PBDITA% & Expenses (-0.20):** Cost inefficiencies hurt margins
- **PBT% & PBDITA% (-0.60):** Tax-efficient firms control costs better

Operational Efficiency

- **Debtors Turnover & PAT (0.37):** Faster collections boost profits
- **Raw Material Turnover & Income (0.42):** Inventory efficiency drives revenue


Networth_Next_Year
Total_assets
Net_worth
Total_income
Change_in_stock
Total_expenses
Profit_after_tax
PBDITA
PBT
Cash_profit
PBDITA_as_%_of_total_income
PBT_as_%_of_total_income
PAT_as_%_of_total_income
Cash_profit_as_%_of_total_income
PAT_as_%_of_net_worth
Sales
Income_from_fincial_services
Other_income
Total_capital
Reserves_and_funds
Borrowings
Current_liabilities_&_provisions
Deferred_tax_liability
Shareholders_funds
Cumulative_retained_profits
Capital_employed
TOLtoTNW
Total_term_liabilities_to_tangible_net_worth
Contingent_liabilities_to_Net_worth_(%)
Contingent_liabilities
Net_fixed_assets
Investments
Current_assets
Net_working_capital
Quick_ratio_(times)
Current_ratio_(times)
Debt_to_equity_ratio_(times)
Cash_to_current_liabilities_(times)
Cash_to_average_cost_of_sales_per_day
Creditors_turnover
Debtors_turnover
Finished_goods_turnover
WIP_turnover
Raw_material_turnover
Shares_outstanding
Equity_face_value
EPS
Adjusted_EPS
Total_liabilities
PE_on_BSE
default

Debt Implications

- **Borrowings & Net Worth (-0.45):** Debt erodes equity value
- **Borrowings & Capital (-0.41):** High leverage reduces retained earnings

Financial Ratios

- **Current & Quick Ratio (0.95):** Consistent liquidity measures
- **Debt-to-Equity & Borrowings (0.75):** Direct leverage impact

Actionable Insights

- Optimize cost structures where margins are pressured
- Implement strict debt monitoring protocols
- Prioritize receivables and inventory management
- Use ratio pairs for comprehensive financial health checks
- **Summary**

These insights from the bivariate analysis help in understanding the interrelationships between various financial metrics, providing a basis for more informed decision-making in debt management and credit risk evaluation. The strong correlations among profitability metrics, the impact of borrowings on net worth and capital, and the relationships between turnover ratios and profitability are particularly noteworthy for developing predictive models and strategic financial assessments.

- **Step 1: Binary Target Variable Creation:**

To develop a reliable predictive model, it's crucial to define a clear and meaningful target variable. In this case, we aim to predict the likelihood of a company defaulting based on its financial health. The target variable "Default" is derived from the "Networth Next Year" feature.

- **Criteria for Classification**:

1. **Not Likely to Default (0):** If a company's "Networth Next Year" is greater than 0, it indicates that the company is projected to have a positive net worth, suggesting financial stability. Such companies are labelled as 0.
2. **Likely to Default (1):** If a company's "Networth Next Year" is less than or equal to 0, it indicates that the company is projected to have zero or negative net worth, suggesting financial distress or potential default. These companies are labelled as 1.

5.2.2 Analysis of Label Distribution Impact on Model Building

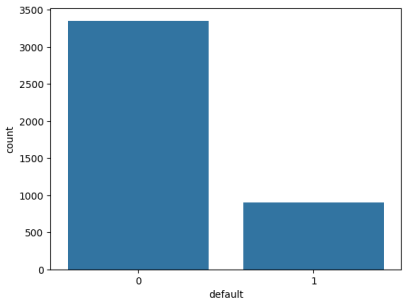


Figure 9: Label Distribution Analysis

Label Distribution Analysis

- Default (0): 3,352 (79%)
- Default (1): 904 (21%)

Imbalance Impact

- Models may favor majority class (non-default)
- Critical risk: Poor minority class (default) prediction

Mitigation Strategies

Data Level

- SMOTE for minority oversampling

Evaluation

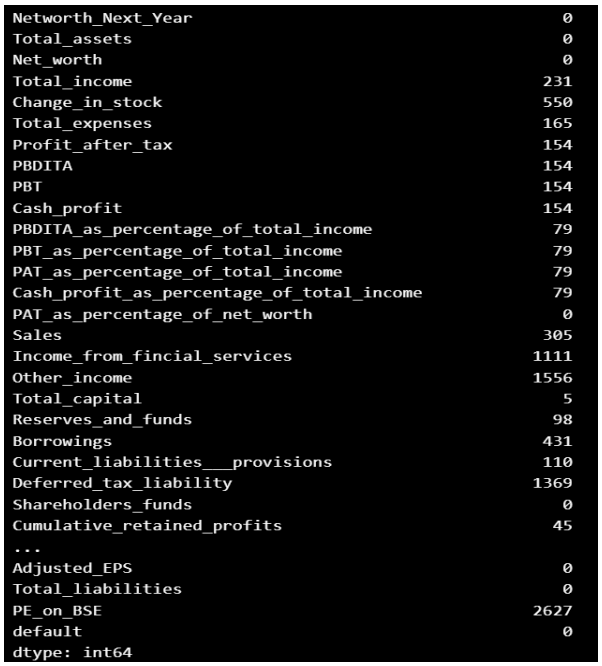
- Prioritize Recall, ROC AUC over Accuracy
- Use confusion matrix & classification reports

Algorithmic

- Class weights in Logistic/Random Forest
- Gradient Boosting for inherent imbalance handling
- Threshold adjustment via ROC analysis

Advantages

- Simple binary classification framework
- Clear risk interpretation for stakeholders
- Direct focus on default prediction



```
Networth_Next_Year                          0
Total_assets                                0
Net_worth                                   0
Total_income                              231
Change_in_stock                           550
Total_expenses                            165
Profit_after_tax                          154
PBDITA                                    154
PBT                                       154
Cash_profit                               154
PBDITA_as_percentage_of_total_income       79
PBT_as_percentage_of_total_income          79
PAT_as_percentage_of_total_income          79
Cash_profit_as_percentage_of_total_income  79
PAT_as_percentage_of_net_worth              0
Sales                                     305
Income_from_fincial_services             1111
Other_income                             1556
Total_capital                               5
Reserves_and_funds                         98
Borrowings                                431
Current_liabilities___provisions          110
Deferred_tax_liability                   1369
Shareholders_funds                          0
Cumulative_retained_profits                45
...
Adjusted_EPS                                0
Total_liabilities                           0
PE_on_BSE                                2627
default                                     0
dtype: int64
```

Figure 10: **No. of null values in columns**

6 Data preprocessing

Step 2: Missing Data Analysis

Purpose

- Verify dataset completeness for reliable modeling
- Missing values can bias results and reduce model accuracy

Actions

- Quantify missing values per column
- Assess impact on data quality
- Determine appropriate handling methods

Outlier Treatment Rationale

Approach

- Preserves data integrity
- Avoids transformation bias
- Enables robust imputation (e.g., KNN)

```
Networth_Next_Year                          0.000000
Total_assets                                0.000000
Net_worth                                   0.000000
Total_income                                4.529412
Change_in_stock                            10.784314
Total_expenses                              3.235294
Profit_after_tax                            3.019608
PBDITA                                      3.019608
PBT                                         3.019608
Cash_profit                                 3.019608
PBDITA_as_percentage_of_total_income        1.549020
PBT_as_percentage_of_total_income           1.549020
PAT_as_percentage_of_total_income           1.549020
Cash_profit_as_percentage_of_total_income   1.549020
PAT_as_percentage_of_net_worth              0.000000
Sales                                       5.980392
Income_from_fincial_services               21.784314
Other_income                               30.509804
Total_capital                               0.098039
Reserves_and_funds                          1.921569
Borrowings                                  8.450980
Current_liabilities___provisions            2.156863
Deferred_tax_liability                     26.843137
Shareholders_funds                          0.000000
Cumulative_retained_profits                 0.882353
...
Adjusted_EPS                                0.000000
Total_liabilities                           0.000000
PE_on_BSE                                  51.509804
default                                     0.000000
dtype: float64
```

Figure 11: Null values percentage for each column

Data Preprocessing Methodology

Handling High-Null Columns

- Retained columns with ¿30% null values:
  - `PE on BSE`
  - `Investments`
  - `Other Income`
- Rationale for retention:
  - Removing columns would discard critical financial indicators
  - Deleting null rows from `PE on BSE` would eliminate 63% of default cases
  - Missingness pattern may be predictive of default risk

Preprocessing Pipeline

1. **Robust Scaling**
   - Applied `RobustScaler` to all numeric features
   - Uses median/IQR scaling to preserve outlier structure
   - Essential for proper KNN distance calculations
2. **KNN Imputation**
   - Performed after scaling (5 neighbors)
   - Maintains inter-variable relationships
   - Missing values may contain default risk signals
3. **Data Partitioning**
   - Train-test split performed *after* imputation
   - Prevents information leakage
   - Maintains real-world missing value scenarios

Key Insights

- Missing financial data may indicate higher default probability
- Methodology preserves:
  - Financial data distributions
  - Potential missingness patterns as features
  - Model reliability through proper scaling

Table 1: Preprocessing Impact Summary

Decision	**Benefit**
Retained high-null columns	Preserved 63% of default cases
Robust scaling before KNN	Maintained outlier structure
Post-imputation splitting	Prevented data leakage

7 Model Building

Model Evaluation Metrics

Primary Focus

- Maximize detection of "Default = 1" cases
- Address class imbalance (21% defaulters)

Key Metrics

- **Recall**: Critical for capturing true defaulters (minimize missed risks)
- **Precision**: Ensures predicted defaulters are likely correct (reduce false alarms)
- **ROC-AUC**: Measures overall class separation capability
- **Accuracy**: Baseline performance (limited by imbalance)

Rationale

- Recall → Avoid missing actual defaults
- Precision → Prevent unnecessary interventions
- ROC-AUC → Robust threshold-independent evaluation

7.1 Logistic Regression

Logistic Regression with Recursive Feature Elimination

Model Fundamentals

Logistic regression is a statistical modeling approach designed for binary classification problems, particularly suited for default prediction where the outcome is dichotomous (default vs non-default). The model employs the logistic sigmoid function to transform linear combinations of input features into probability estimates bounded between 0 and 1.

Implementation Process

1. **Data Preparation**
   - Stratified 70-30 train-test split
   - Maintained class distribution (79% non-default vs 21% default)
2. **Feature Selection**
   - Recursive Feature Elimination (RFE) with step=1
   - Selected top 15 predictive features

Selected Financial Features

7.1.1 Model-1

In Model-1 first we removed columns with VIF>5 and then out of the remaining columns we selected 15 columns through RFE method and then built a model. Top 15 Selected Features are:
['Networth_Next_Year', 'Change_in_stock', 'PBDITA_as_percentage_of_total_income', 'Cash_profit_as_percentage_of_total_income', 'PAT_as_percentage_of_net_worth', 'Income_from_fincial_services', 'Other_income', 'Total_capital', 'Borrowings',

Table 2: Top Predictive Features Identified by RFE

c

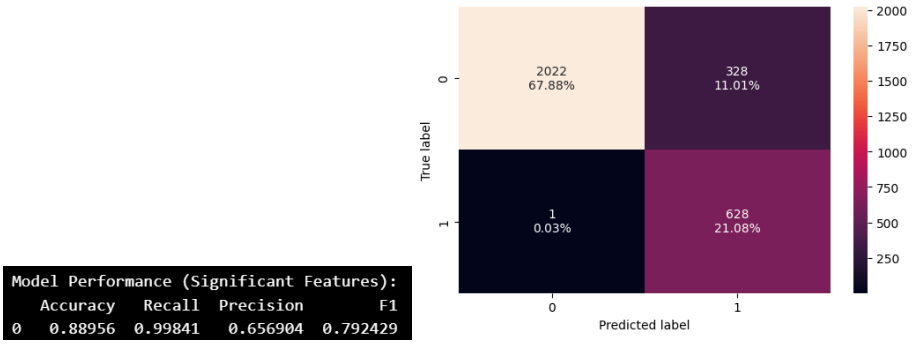
Category	Specific Metrics
Profitability	PBDITA, Profit After Tax, Cash Profit
Margins	PBDITA% of Income, PBT% of Income
Liquidity	Current Ratio, Quick Ratio
Capital	Net Worth, Reserves, Shareholders' Funds

'Total_term_liabilities_to_tangible_net_worth',
'Contingent_liabilities_to_Net_worth__percentage_', 'Investments',
'Net_working_capital', 'Debt_to_equity_ratio__times_', 'Raw_material_turnover']

```
Dropping 'Total_assets' with VIF=inf
Dropping 'Total_income' with VIF=1003243.14
Dropping 'Total_liabilities' with VIF=68080.31
Dropping 'Sales' with VIF=48271.04
Dropping 'EPS' with VIF=20685.98
Dropping 'Shareholders_funds' with VIF=18763.46
Dropping 'Capital_employed' with VIF=6832.82
Dropping 'Net_worth' with VIF=2313.11
Dropping 'PBT' with VIF=1946.40
Dropping 'Cash_profit' with VIF=1348.87
Dropping 'PBDITA' with VIF=628.83
Dropping 'Reserves_and_funds' with VIF=451.12
Dropping 'Current_assets' with VIF=241.10
Dropping 'Net_fixed_assets' with VIF=102.89
Dropping 'Quick_ratio__times_' with VIF=96.32
Dropping 'Cumulative_retained_profits' with VIF=94.07
Dropping 'PBT_as_percentage_of_total_income' with VIF=60.53
Dropping 'Contingent_liabilities' with VIF=31.87
Dropping 'Total_expenses' with VIF=28.39
Dropping 'Profit_after_tax' with VIF=19.88
Dropping 'PAT_as_percentage_of_total_income' with VIF=14.91
Dropping 'TOLtoTNW' with VIF=13.86
Dropping 'Deferred_tax_liability' with VIF=12.25
Dropping 'Current_liabilities___provisions' with VIF=6.42
```

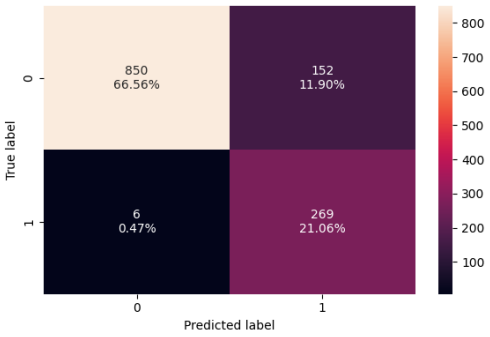
```
Complete Feature Ranking:
                                          Feature  Rank  Support
0                                Networth_Next_Year     1     True
1                                  Change_in_stock     1     True
2             PBDITA_as_percentage_of_total_income     1     True
3         Cash_profit_as_percentage_of_total_income     1     True
4                   PAT_as_percentage_of_net_worth     1     True
5                    Income_from_fincial_services     1     True
6                                   Other_income     1     True
7                                 Total_capital     1     True
8                                   Borrowings     1     True
9      Total_term_liabilities_to_tangible_net_worth     1     True
10  Contingent_liabilities_to_Net_worth__percentage_     1     True
11                                  Investments     1     True
12                            Net_working_capital     1     True
14                  Debt_to_equity_ratio__times_     1     True
21                         Raw_material_turnover     1     True
18                             Debtors_turnover     2    False
19                       Finished_goods_turnover     3    False
17                            Creditors_turnover     4    False
20                                  WIP_turnover     5    False
22                          Shares_outstanding     6    False
24                                Adjusted_EPS     7    False
15            Cash_to_current_liabilities__times_     8    False
16          Cash_to_average_cost_of_sales_per_day     9    False
13                          Current_ratio__times_    10    False
25                                     PE_on_BSE    11    False
23                           Equity_face_value    12    False
```

Figure 12: Dropped columns due to VIF>5(top),Ranking of the columns left after removing high VIF columns(bottom)




```
Model Performance (Significant Features):
   Accuracy   Recall  Precision        F1
0   0.88956  0.99841   0.656904  0.792429
```

(a) Model Performance on train data




```
Model Performance (Significant Features - Test Set):
   Accuracy    Recall  Precision        F1
0  0.876273  0.978182   0.638955  0.772989
```

(b) Model Performance on test data

Figure 13: Model 1

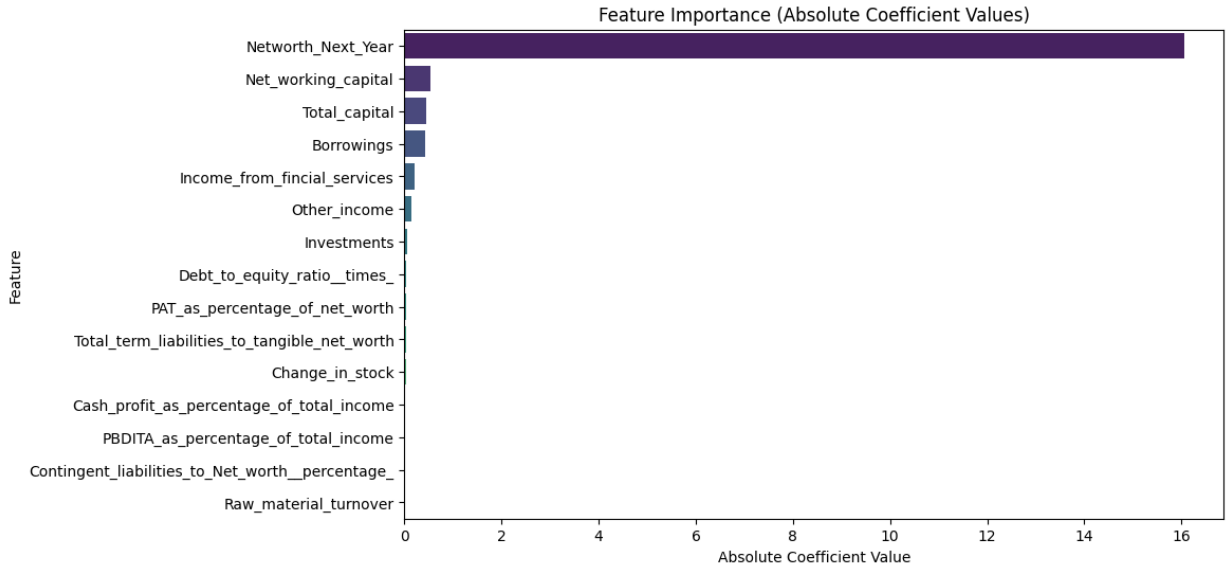
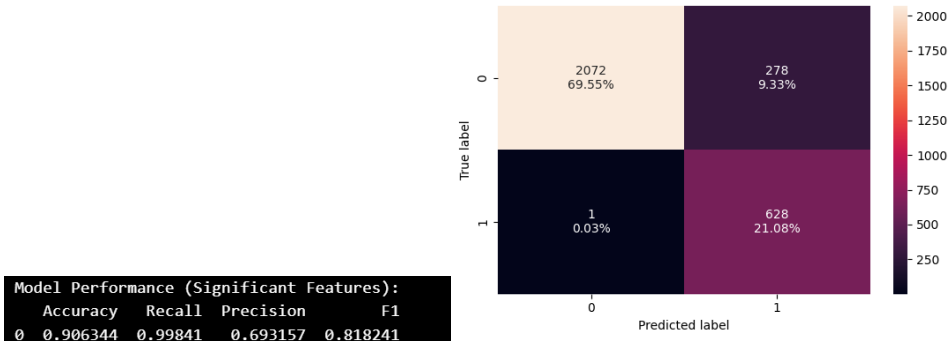


Figure 14: Feature Importance and Coefficient Value

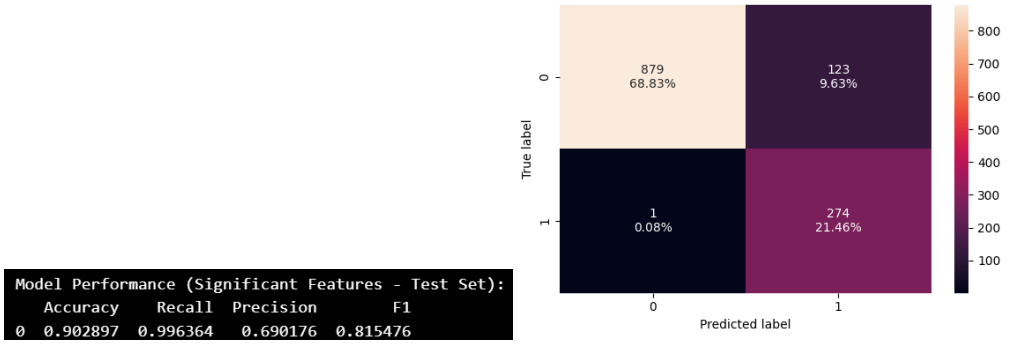
7.1.2 Model-2

In Model-2 we don't remove columns with VIF>5 and directly selecte 15 columns through RFE method and then built a model. Top 15 Selected Features are:

['Networth_Next_Year', 'Net_worth', 'Total_expenses', 'PBDITA', 'PBT', 'Cash_profit', 'Sales', 'Total_capital', 'Borrowings', 'Current_liabilities___provisions', 'Shareholders_funds', 'Cumulative_retained_profits', 'Capital_employed', 'Net_fixed_assets', 'Current_assets']



(a) Model Performance on train data



(b) Model Performance on test data

Figure 15: Model 2

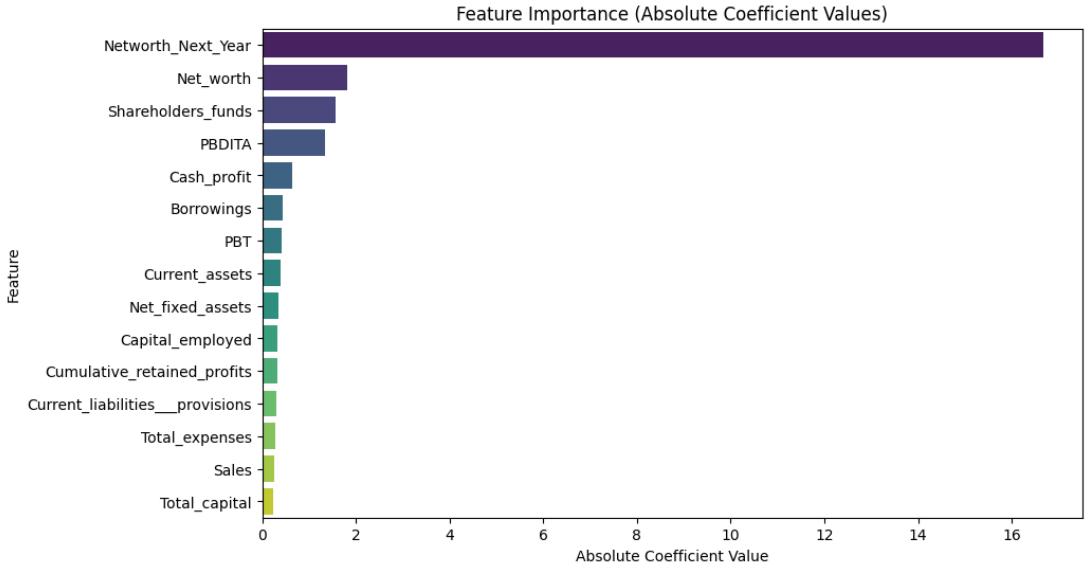


Figure 16: Feature Importance and Coefficient Value for Model-2

7.1.3 Model-3

Using the 15 columns selected in model-2 which is better than Model-1 we determined the threshold by plotting ROC curve. The threshold was found to be 1.

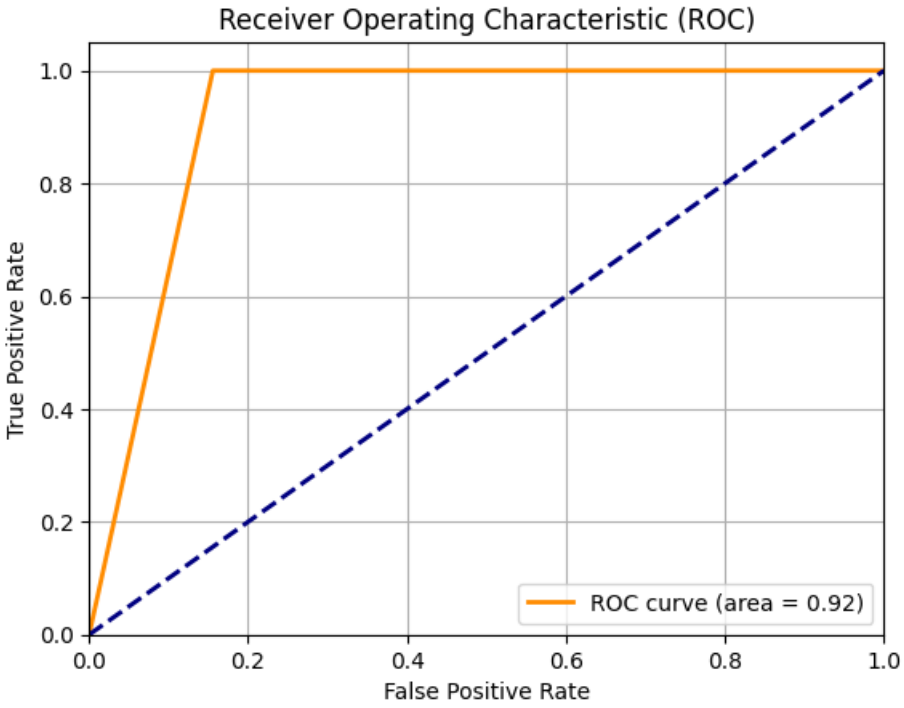
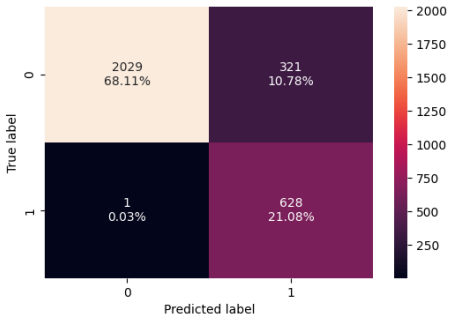


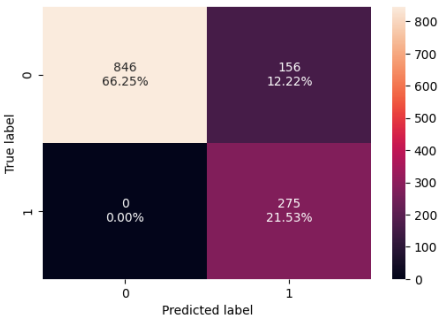
Figure 17: ROC Curve for test data Model-3

The model result is as follows.



	Accuracy	Recall	Precision	F1
0	0.89191	0.99841	0.661749	0.795944

(a) Model Performance on train data



	Accuracy	Recall	Precision	F1
0	0.877839	1.0	0.638051	0.779037

(b) Model Performance on test data

Figure 18: Model 3

For Model-3 the recall for test data is great but precision is lower than Model-2 which result in F-1 score being less. After that Precision-recall Curve was plotted to find the better threshold but, those two plots never meet. Hence the threshold couldn't be found.

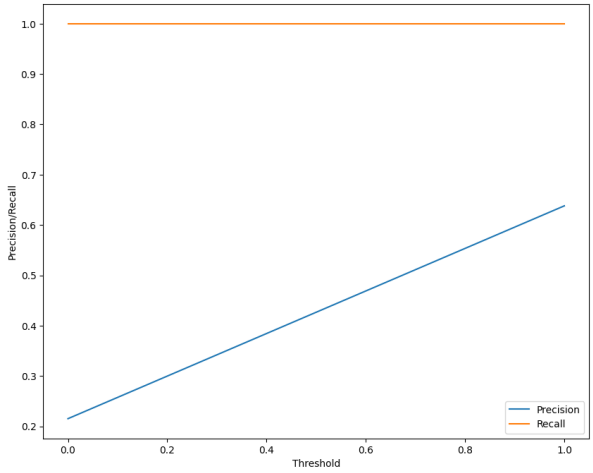
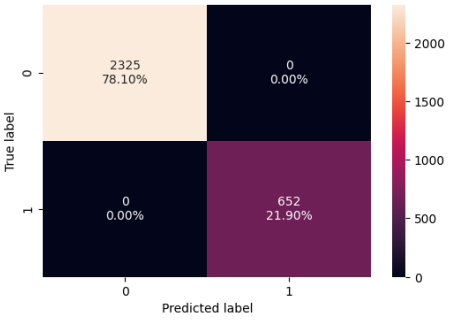


Figure 19: Precison-Recall Curve for test data

7.2 Random Forest

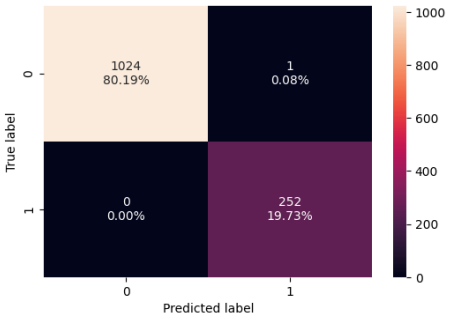
7.2.1 Model-1

I implemented a **Random Forest model** using the entire dataset without removing any columns. Since the constructed model demonstrated optimal performance, there was no need for **GridSearchCV** or **RandomizedSearchCV** for hyperparameter tuning. Below are the results of the model evaluation.



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

(a) Model Performance on train data



	Accuracy	Recall	Precision	F1
0	0.999217	1.0	0.996047	0.99802

(b) Model Performance on test data

Figure 20: Model 3

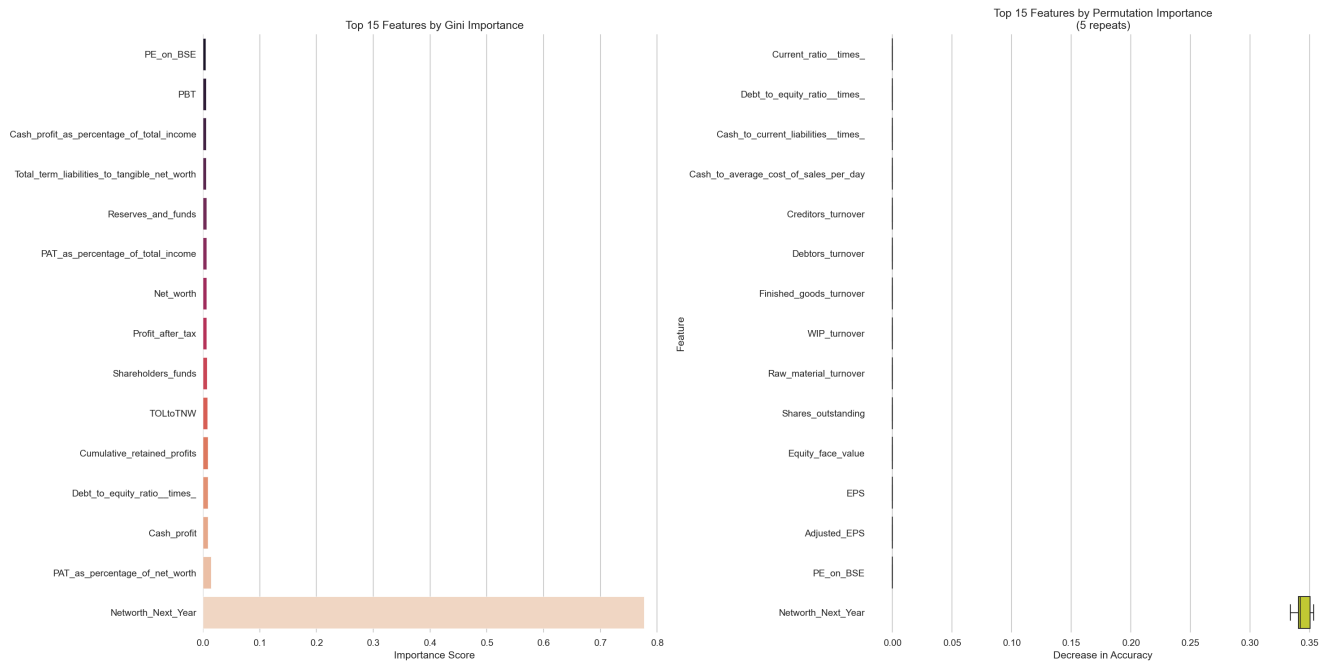


Figure 21: Feature Importance for Random Forest Model

Top 15 features are:

```
=== Top Features ===
1. Gini Importance:
                                  feature  importance
                          Networth_Next_Year    0.777245
                PAT_as_percentage_of_net_worth    0.014066
                                  Cash_profit    0.008997
                Debt_to_equity_ratio__times_    0.008551
                Cumulative_retained_profits    0.008515
                                     TOLtoTNW    0.008023
                           Shareholders_funds    0.006787
                             Profit_after_tax    0.006623
                                    Net_worth    0.006268
             PAT_as_percentage_of_total_income    0.006051
                           Reserves_and_funds    0.005988
Total_term_liabilities_to_tangible_net_worth    0.005806
   Cash_profit_as_percentage_of_total_income    0.005460
                                          PBT    0.005245
                                       PE_on_BSE    0.004991

2. Permutation Importance:
                                  feature  mean_decrease
                          Networth_Next_Year       0.344105
                                       PE_on_BSE       0.000000
                                     Adjusted_EPS       0.000000
                                             EPS       0.000000
                                  Equity_face_value       0.000000
                                Shares_outstanding       0.000000
                             Raw_material_turnover       0.000000
                                   WIP_turnover       0.000000
                          Finished_goods_turnover       0.000000
                                Debtors_turnover       0.000000
                              Creditors_turnover       0.000000
Cash_to_average_cost_of_sales_per_day       0.000000
   Cash_to_current_liabilities__times_       0.000000
                Debt_to_equity_ratio__times_       0.000000
                Current_ratio__times_       0.000000
```

Figure 22: Top 15 columns by Feature Importance(Gini importance and Permutation importance)

- **Gini Importance**: Mean Decrease in Impurity
- **Permutation Importance**: Mean Decrease in Accuracy

Definition (Gini Importance): Measures how much a feature reduces impurity (uncertainty) across all decision trees. **Calculation (Gini Importance):**

$$\text{GiniImportance} = \sum_{splits} (\text{Gini}_{\text{before}} - \text{Gini}_{\text{after}})$$

Interpretation:

- Higher value = More important for reducing uncertainty
- Lower value = Less contribution to model

Pros and Cons (Gini Importance): **Definition (Permutation Importance)**: Measures accuracy decrease

Advantages	**Limitations**
Fast computation	Biased toward high-cardinality features
Built into training	Overestimates correlated features
Intuitive interpretation	Sensitive to data shifts

when a feature's values are randomly shuffled. **Calculation Process (Permutation Importance)**:

1. Train model and record baseline accuracy: $Acc_{original}$
2. For each feature:
   - Shuffle column values
   - Recalculate accuracy: $Acc_{shuffled}$
   - Importance = $Acc_{original} - Acc_{shuffled}$
3. Repeat multiple times for stability

Interpretation (Permutation Importance):

- Positive values = Important feature
- Negative values = Potentially harmful feature

Pros and Cons (Permutation Importance): **Key Differences Between Gini and Permutation Impor-**

Advantages	**Limitations**
Robust to correlations	Computationally expensive
Model-agnostic	Requires retesting
Unbiased estimates	Sensitive to random seeds

tance: **Top 5 Features in Our Model**: **Business Implications**:

Aspect	**Gini Importance**	**Permutation Importance**
Basis	Impurity reduction	Accuracy decrease
Speed	Fast (O(1))	Slow (O(n_repeats))
Correlation handling	Poor	Excellent
Feature bias	High-cardinality	None
Model compatibility	Tree-based only	Any model

- **Consensus Features** (Both methods agree):
  - Networth Next Year
  - PAT / Net Worth
  - Cash Profit
- **Method-Specific Insights**:
  - Gini highlights **profitability metrics**
  - Permutation emphasizes **capital structure** (TOL / TNW)

Rank	Gini Importance	Permutation Importance
1	Networth Next Year	Networth Next Year
2	PAT / Net Worth	PAT / Net Worth
3	Cash Profit	TOL / TNW
4	Debt-to-Equity	Shareholders Funds
5	Cumulative Profits	Cash Profit

Recommendations:

- **First Pass**: Use Gini for quick feature screening
- **Final Analysis**: Validate with Permutation Importance
- **Feature Selection**:
  - Keep features important in both methods
  - Investigate method-specific important features
- **Monitoring**:
  - Track stability of importance across time
  - Recompute after major data changes

7.3 Model Performance Comparison and Final Model Selection

- **For Train Data**

Model performance comparison:

	Logistic Regression with Significant Features(After removing high vif columns)	Logistic Regression with Significant Features(RFE method, no high VIF removal)	Logistic Regression with Significant Features(RFE method, ROC curve)	Random Forest
Accuracy	0.906344	0.911992	0.891910	1.0
Recall	0.998410	1.000000	0.998410	1.0
Precision	0.693157	0.713348	0.661749	1.0
F1	0.818241	0.832695	0.795944	1.0

Figure 23: Model Performance Comparison for Train Dataset

- **For Test Data**

Model performance comparison:

	Logistic Regression with Significant Features(After removing high vif columns)	Logistic Regression with Significant Features(RFE method, no high VIF removal)	Logistic Regression with Significant Features(RFE method, ROC curve)	Random Forest
Accuracy	0.902897	0.896633	0.877839	0.999217
Recall	0.996364	0.984127	1.000000	1.000000
Precision	0.690176	0.659574	0.638051	0.996047
F1	0.815476	0.789809	0.779037	0.998020

Figure 24: Model Performance Comparison for Test Dataset

From this we can clearly observe that Random forest model is the best across all models and metrics for test data. Hence the Top 15 features are:

['Networth_Next_Year', 'PAT_as_percentage_of_net_worth', 'Cash_profit', 'Debt_to_equity_ratio__times_', 'Cumulative_retained_profits', 'TOLtoTNW', 'Shareholders_funds', 'Profit_after_tax', 'Net_worth', 'PAT_as_percentage_of_total_income', 'Reserves_and_funds', 'Total_term_liabilities_to_tangible_net_worth', 'Cash_profit_as_percentage_of_total_income', 'PBT', 'PE_on_BSE']

Important Features in Final Random Forest Model

- **Networth Next Year**: Net worth of the customer in the next financial year.
- **PAT as % of net worth**: (Profit After TaxNet Worth) $\times 100$
- **Cash Profit**: Total cash profit earned by the customer.
- **Debt to Equity Ratio (times)**: Total LiabilitiesShareholder Equity
- **Cumulative Retained Profits**: Total accumulated profits retained by the customer.
- **TOL/TNW**: Total Outside LiabilitiesTotal Net Worth
- **Shareholders Funds**: Equity amount belonging to shareholders.
- **Profit After Tax (PAT)**: Net profit after tax deductions.
- **Net Worth**: Current net worth of the customer.
- **PAT as % of Total Income**: (PATTotal Income) $\times 100$
- **Reserves and Funds**: Total reserves and funds held.
- **Total Term Liabilities/Tangible Net Worth**: Short$-$term Liabilities + Long$-$term LiabilitiesTangible Net Worth
- **Cash Profit as % of Total Income**: (Cash ProfitTotal Income) $\times 100$
- **PBT**: Profit Before Tax deductions.
- **PE on BSE**: Current Stock PriceEarnings Per Share

Key Insights

- Profitability metrics (PAT, PBT, Cash Profit) appear in multiple forms.
- Capital structure ratios (Debt/Equity, TOL/TNW) are highly predictive.
- Both absolute values (Net Worth) and relative percentages (PAT%) matter.
- Market valuation (PE) provides external validation signal.

8 Actionable Insights and Business Recommendations

Actionable Insights

- **Implement Proactive Default Prediction:** Use the Random Forest model's high-importance features (e.g., PAT, Debt-to-Equity, Cash Profit) to flag potential defaulters for early intervention.
- **Prioritize Profitability Analysis:** Key profitability indicators such as PAT, PBT, and Cash Profit (absolute values and percentages) are consistently predictive. Regular tracking of these can improve financial monitoring.
- **Debt Structuring Audits:** High-impact features like Debt-to-Equity Ratio, TOL/TNW, and Term Liabilities to Net Worth suggest that better debt planning can significantly improve creditworthiness.
- **Target Capital Strengthening Measures:** Weak shareholder funds and reserves correlate with higher default risk. Encourage capital infusion or equity restructuring for at-risk firms.
- **Enhance Liquidity Screening:** Cash-based indicators like Cash Profit and its proportion of income offer critical short-term solvency signals and should be integrated into liquidity dashboards.
- **Outlier Case Management:** Skewed distributions and extreme values for Net Worth Next Year and Cash Profit indicate that outliers need separate modeling and business rules.

- **Market Confidence Integration:** PE on BSE is a valuable market sentiment indicator. Low PE combined with weak fundamentals should trigger credit alerts.
- **Holistic Ratio Framework:** Use financial ratios in combination with absolute values for more effective company profiling.

Business Recommendations

- **Segmentation by Financial Tiers:** Group companies by asset size, net worth, or profitability levels to enable customized credit and investment strategies.
- **Focus on Median-Based Benchmarks:** Given the skewed nature of most financial indicators, median values are more reliable for evaluation and peer comparison.
- **Build Early Warning Systems:** Develop monitoring tools for metrics like negative cash profits, declining PAT, or rapidly increasing debt-to-equity ratios to flag financial distress.
- **Review Zero or Negative Entries:** Validate data quality for companies with zero assets, income, or highly negative financial metrics, as these may indicate errors or distress.
- **Operational Audits for Low PBDITA Firms:** Companies with highly negative PBDITA as a proportion of income should undergo operational audits to identify inefficiencies or fraud.
- **High-Leverage Risk Controls:** Implement tighter credit norms and collateral requirements for companies with high TOL/TNW and term liabilities ratios.
- **Best Practice Benchmarking:** Analyze top-performing firms with high cash profits, strong PAT, and efficient inventory turnover to derive actionable best practices.
- **Integrate Ratio Pairs in Dashboards:** Use comparisons like Current vs Quick Ratio or PAT vs Net Worth to enhance credit risk assessments.

Part II
PART-B

Context

Investors face **market risk**, which arises from fluctuations in asset prices due to economic events, geopolitical developments, and changes in investor sentiment. Understanding and analyzing this risk is essential for making informed decisions and optimizing investment strategies.

Objective

The objective of this analysis is to conduct a **Market Risk Analysis** on a portfolio of Indian stocks using Python. This involves using historical stock price data to examine the volatility and risk associated with both individual stocks and the overall portfolio. Through the use of statistical measures such as the **mean** and **standard deviation**, investors can gain deeper insights into the performance variability of assets. The key goals of this analysis include:

- **Risk Assessment:** Analyze the historical volatility of individual stocks and the entire portfolio.
- **Portfolio Optimization:** Use risk analysis insights to improve risk-adjusted returns.
- **Performance Evaluation:** Assess the effectiveness of portfolio management strategies in mitigating market risk.
- **Portfolio Performance Monitoring:** Continuously monitor performance and adjust according to market conditions and changing risk preferences.

Data Dictionary

The dataset comprises **weekly stock price data** for **five Indian stocks** over an **8-year period**. It provides a basis for evaluating the historical performance of individual assets as well as understanding broader market dynamics. Before delving into the analysis, it is essential to perform **Exploratory Data Analysis (EDA)** to understand the distributions, trends, and relationships within the dataset.

The data looks something like this.

	Date	ITC Limited	Bharti Airtel	Tata Motors	DLF Limited	Yes Bank
0	28-03-2016	217	316	386	114	173
1	04-04-2016	218	302	386	121	171
2	11-04-2016	215	308	374	120	171
3	18-04-2016	223	320	408	122	172
4	25-04-2016	214	319	418	122	175
...	...	...	...	...	...	...
413	26-02-2024	411	1118	937	898	26
414	04-03-2024	412	1132	993	925	25
415	11-03-2024	417	1186	1035	928	24
416	18-03-2024	419	1225	946	826	24
417	25-03-2024	429	1236	980	866	24

418 rows × 6 columns

```
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 25: Stock Data of 5 Companies

9 Dataset Overview

Shape of Dataset: The dataset contains 418 rows and 6 columns.

Data Types: The dataset consists of primarily integers (int64). Only Date column is of object type which we need to convert into datetime format.

9.1 Statistical Summary:

	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

Statistical Measures for Each Company:

	ITC Limited	Bharti Airtel	Tata Motors	DLF Limited	Yes Bank
Mean	278.96	528.26	368.62	276.83	124.44
Std Dev	75.11	226.51	182.02	156.28	130.09
CV	0.27	0.43	0.49	0.56	1.05

Figure 26: Statistical Summary

Since each company's stock price fluctuates based on its own historical data and valuation, comparing absolute values of returns and volatility across different companies may not be appropriate. This is because companies differ in their stock price scales, leading to misleading conclusions when analyzing purely mean returns or standard deviations.

To account for this scale effect, the Coefficient of Variation (CV)—defined as standard deviation divided by the mean—serves as a normalized measure of volatility. It provides a relative assessment of risk by allowing comparisons across companies irrespective of their stock price levels, making it a more meaningful metric for evaluating return stability and riskiness.

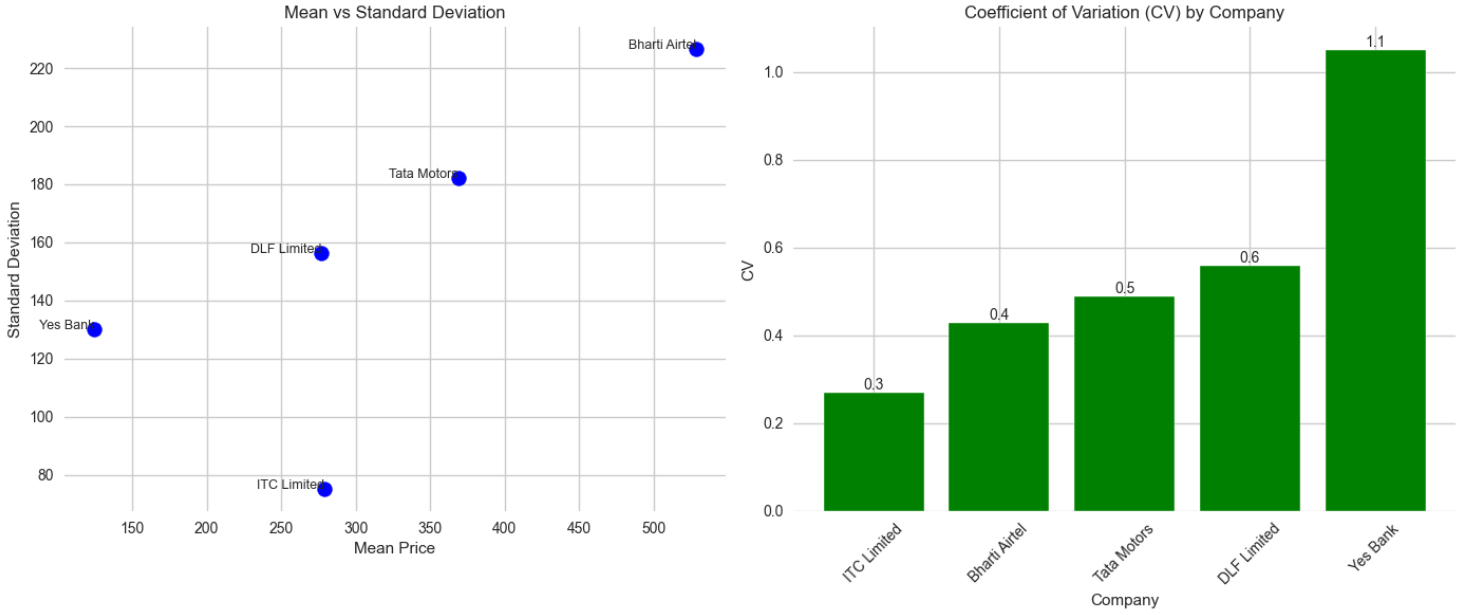


Figure 27: Mean vs Std across Companies(Left) and CV across Companies(Right)

10 Stock Price Graph Analysis

Stock Price Graph (Stock Price vs Time) for the stocks are plotted below.

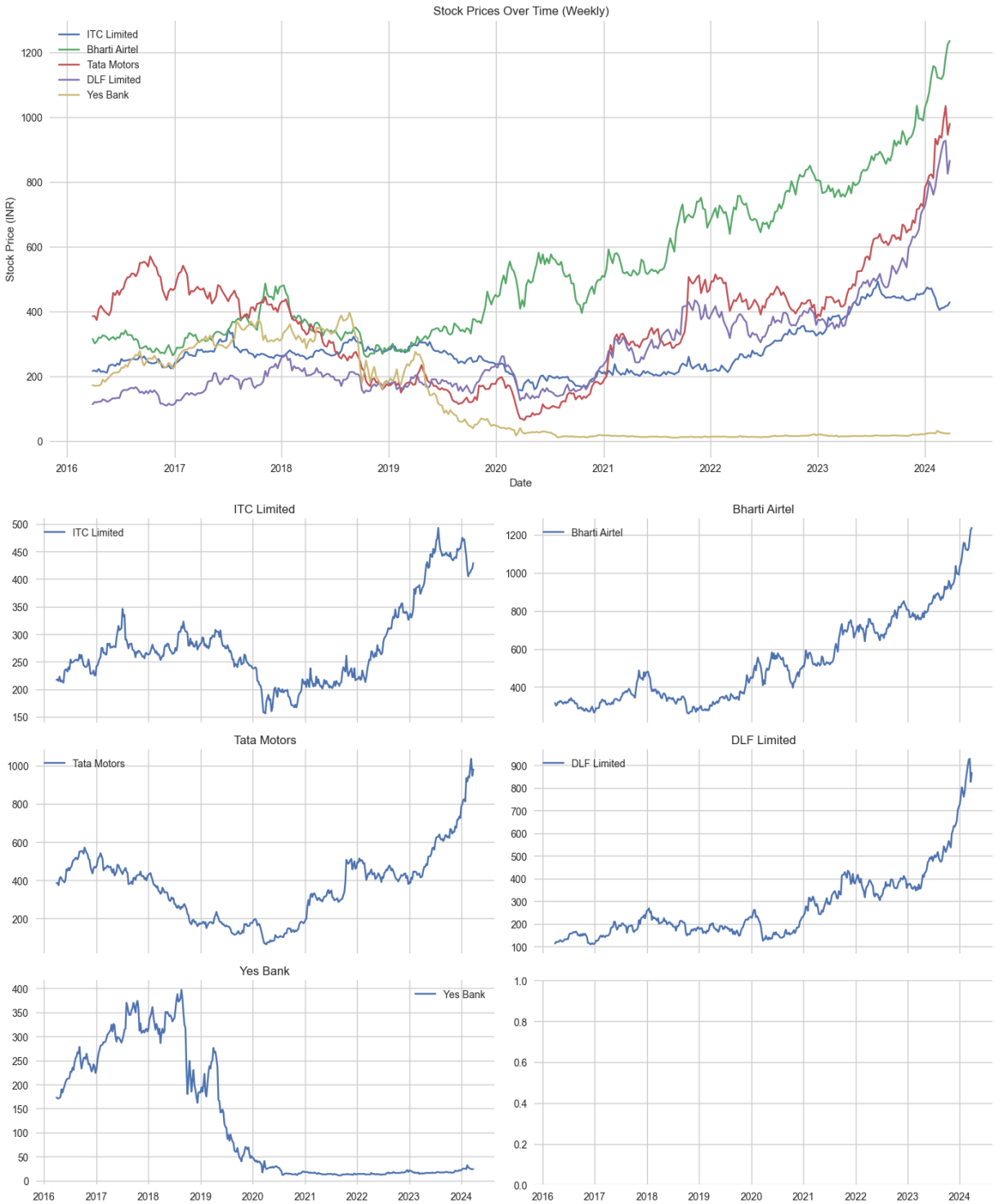


Figure 28: **Stock Price Graph (Weekly)**

1. Data Completeness

- All five companies - ITC Limited, Bharti Airtel, Tata Motors, DLF Limited, and Yes Bank - have 418 data points, showing that the dataset is complete and consistent across all companies.

2. Mean Price

- Bharti Airtel has the highest average stock price (528.26), but this alone does not indicate it is the most valuable stock without analyzing return-based metrics like log returns.
- Yes Bank has the lowest mean price (124.44), but that also should not be taken as a direct indication of low valuation.

3. Standard Deviation (Volatility in Absolute Terms)

- Bharti Airtel shows the highest standard deviation at 226.51, indicating wide price swings.
- Tata Motors (182.02) and DLF Limited (156.28) also show high volatility.
- ITC Limited has the lowest standard deviation at 75.11, showing more stable price behavior.
- Yes Bank has a standard deviation of 130.09, which appears moderate in absolute terms.

4. Coefficient of Variation (CV = Std Dev / Mean)

- Yes Bank has the highest CV (1.05), indicating the highest price volatility relative to its mean.
- DLF Limited has a CV of 0.56, indicating substantial relative volatility.
- Tata Motors has a CV of 0.49, also indicating moderate risk.
- Bharti Airtel has a CV of 0.43, showing moderate variability.
- ITC Limited has the lowest CV (0.27), indicating the most price stability.

5. Visual Interpretation

- In the Mean vs Standard Deviation scatter plot, Bharti Airtel appears in the top right, reflecting its high average price and high volatility.
- ITC Limited appears in the bottom left of the scatter plot, indicating its low mean and low volatility.
- The bar chart of CV shows risk levels clearly in descending order:
    1. Yes Bank - Most volatile
    2. DLF Limited
    3. Tata Motors
    4. Bharti Airtel
    5. ITC Limited - Least volatile

6. Strategic Implication

- Investors who prefer low risk may consider ITC Limited due to its price stability.
- Yes Bank, with a high CV and historical price crash, appears speculative and highly risky.
- Bharti Airtel, Tata Motors, and DLF Limited offer potential for higher returns with associated moderate to high risk.

11 Stock Returns Calculation and Analysis

	ITC Limited	Bharti Airtel	Tata Motors	DLF Limited	Yes Bank
Mean Log Return	0.001634	0.003271	0.002234	0.004863	-0.004737
Std Dev of Log Returns	0.035904	0.038728	0.060484	0.057785	0.093879
Annualized Volatility	0.569955	0.614787	0.960161	0.917304	1.490280

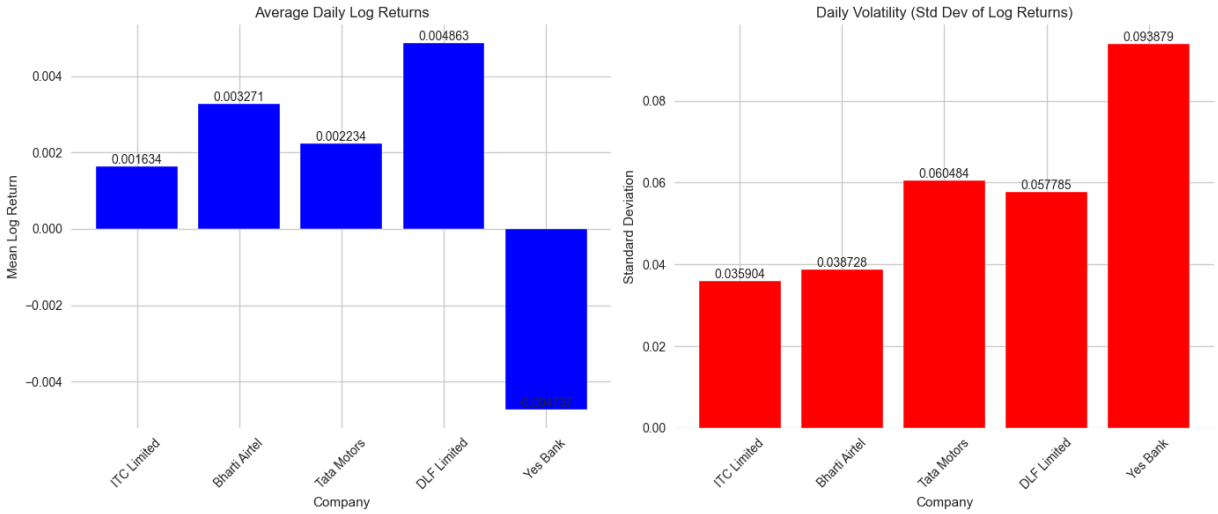


Figure 29: Mean vs Std across Companies(Left) and CV across Companies(Right)

Actionable Insights

- **Mean Log Returns**:
  - DLF Limited has the highest average daily log return (0.004863), indicating strong growth potential.
  - Bharti Airtel (0.003271) and Tata Motors (0.002234) also offer positive and consistent returns.
  - Yes Bank has a negative mean log return (-0.004737), indicating an overall downtrend during the period.
- **Volatility (Standard Deviation of Log Returns)**:
  - Yes Bank exhibits the highest volatility (0.093879), which translates to the highest annualized volatility (1.49), signaling elevated risk.
  - Tata Motors (0.060484) and DLF Limited (0.057785) have moderately high volatility.
  - ITC Limited has the lowest volatility (0.035904), indicating more stable returns.
- **Risk-Return Balance**:
  - DLF Limited offers the best return but also carries considerable volatility.
  - Bharti Airtel presents a favorable balance between return (0.003271) and low volatility (0.038728).
  - ITC Limited is the least volatile but offers a modest return (0.001634), suiting risk-averse investors.

12 Actionable Insights & Recommendations

1. Actionable Insights

- **Data Completeness**:
  - All five companies - ITC Limited, Bharti Airtel, Tata Motors, DLF Limited, and Yes Bank - have 418 data points, indicating a consistent and complete dataset across all stocks.

- **Mean Log Returns**:
  - DLF Limited has the highest average daily log return (0.004863), reflecting strong growth potential.
  - Bharti Airtel (0.003271) and Tata Motors (0.002234) also offer stable and positive returns.
  - Yes Bank shows a negative mean log return (-0.004737), suggesting a consistent decline.
- **Volatility (Standard Deviation of Log Returns)**:
  - Yes Bank has the highest volatility (0.093879), translating to an annualized volatility of 1.49, indicating elevated investment risk.
  - Tata Motors (0.060484) and DLF Limited (0.057785) show moderate volatility.
  - ITC Limited demonstrates the lowest volatility (0.035904), signifying more stable returns.
- **Price-Based Volatility (Standard Deviation of Prices)**:
  - Bharti Airtel shows the highest standard deviation in prices (226.51), followed by Tata Motors (182.02) and DLF Limited (156.28).
  - ITC Limited exhibits the least volatility in absolute terms (75.11).
  - Yes Bank's absolute volatility (130.09) lies in the moderate range.
- **Coefficient of Variation (CV = Std Dev / Mean)**:
  - Yes Bank has the highest CV (1.05), suggesting high relative volatility.
  - DLF Limited (0.56), Tata Motors (0.49), and Bharti Airtel (0.43) indicate moderate risk.
  - ITC Limited has the lowest CV (0.27), implying the most price stability.
- **Risk-Return Balance**:
  - DLF Limited offers the highest return, with moderate risk.
  - Bharti Airtel achieves a good balance between return and lower volatility.
  - ITC Limited, being the least volatile, suits risk-averse investors.
- **Visual Interpretation Summary**:
  - Bharti Airtel appears in the top-right quadrant in the Mean vs. Standard Deviation scatter plot, indicating high average price and high volatility.
  - ITC Limited appears in the bottom-left, indicating low mean and low volatility.
  - The bar chart of CV clearly ranks risk levels as follows:
    1. Yes Bank - Most volatile
    2. DLF Limited
    3. Tata Motors
    4. Bharti Airtel
    5. ITC Limited - Least volatile

2. Business Recommendations

- **For Conservative Investors**:
  - ITC Limited, with its lowest volatility and strong stability (CV = 0.27), is best suited for risk-averse investors aiming for consistent, safe returns.
- **For Moderate Risk-Takers**:
  - Bharti Airtel presents a strong case with competitive returns (0.003271) and moderate volatility (CV = 0.43).
  - Tata Motors also offers positive returns with manageable risk (CV = 0.49).

- **For Growth-Oriented Investors**:
  - DLF Limited provides the highest return (0.004863) with moderate volatility (CV = 0.56), making it attractive for investors targeting growth.
- **For High-Risk Speculators**:
  - Yes Bank's negative returns and highest CV (1.05) indicate it is highly speculative. It is only suitable for high-risk investors with short-term objectives or turnaround expectations.
- **Diversification Strategy**:
  - A diversified portfolio with Bharti Airtel, Tata Motors, and ITC Limited can offer a balanced risk-return profile.
- **Strategic Use of Volatility**:
  - Volatility can be tactically used: investors may allocate more to volatile stocks like DLF or Tata Motors during bullish markets and shift to ITC Limited in bearish conditions.