Abubacker Siddhique I
FRA
July 28, 2024

Finance & Risk Analytics

Table of Contents

List Of Figures

List Of Tables:

Problem 1

Define the problem and perform Exploratory Data Analysis:

- The aim of this project is to evaluate the financial well-being and creditworthiness of companies using the balance sheets of the respective companies and to identify potential cases of default.
- We need to build a predictive model that will help the organization anticipate potential challenges with the financial performance of the companies and enable proactive risk mitigation strategies.

Data Description:

- The dataset consists of 4256 rows and 51 attributes.
- There are null values present in almost all the columns
- All the columns are of float datatype.

```
 0   Networth_Next_Year                                 4256 non-null   float64
 1   Total_assets                                       4256 non-null   float64
 2   Net_worth                                          4256 non-null   float64
 3   Total_income                                       4025 non-null   float64
 4   Change_in_stock                                    3706 non-null   float64
 5   Total_expenses                                     4091 non-null   float64
 6   Profit_after_tax                                   4102 non-null   float64
 7   PBDITA                                             4102 non-null   float64
 8   PBT                                                4102 non-null   float64
 9   Cash_profit                                        4102 non-null   float64
 10  PBDITA_as_perc_of_total_income                     4177 non-null   float64
 11  PBT_as_perc_of_total_income                        4177 non-null   float64
 12  PAT_as_perc_of_total_income                        4177 non-null   float64
 13  Cash_profit_as_perc_of_total_income                4177 non-null   float64
 14  PAT_as_perc_of_net_worth                           4256 non-null   float64
 15  Sales                                              3951 non-null   float64
 16  Income_from_fincial_services                       3145 non-null   float64
 17  Other_income                                       2700 non-null   float64
 18  Total_capital                                      4251 non-null   float64
 19  Reserves_and_funds                                 4158 non-null   float64
 20  Borrowings                                         3825 non-null   float64
 21  Current_liabilities_and_provisions                 4146 non-null   float64
 22  Deferred_tax_liability                             2887 non-null   float64
 23  Shareholders_funds                                 4256 non-null   float64
 24  Cumulative_retained_profits                        4211 non-null   float64
 25  Capital_employed                                   4256 non-null   float64
 26  TOL_to_TNW                                         4256 non-null   float64
 27  Total_term_liabilities__to__tangible_net_worth     4256 non-null   float64
 28  Contingent_liabilities__to__Net_worth_perc         4256 non-null   float64
 29  Contingent_liabilities                             2854 non-null   float64
 30  Net_fixed_assets                                   4124 non-null   float64
 31  Investments                                        2541 non-null   float64
 32  Current_assets                                     4176 non-null   float64
 33  Net_working_capital                                4219 non-null   float64
 34  Quick_ratio_times                                  4151 non-null   float64
 35  Current_ratio_times                                4151 non-null   float64
 36  Debt_to_equity_ratio_times                         4256 non-null   float64
 37  Cash_to_current_liabilities_times                  4151 non-null   float64
 38  Cash_to_average_cost_of_sales_per_day              4156 non-null   float64
 39  Creditors_turnover                                 3865 non-null   float64
 40  Debtors_turnover                                   3871 non-null   float64
 41  Finished_goods_turnover                            3382 non-null   float64
 42  WIP_turnover                                       3492 non-null   float64
 43  Raw_material_turnover                              3828 non-null   float64
 44  Shares_outstanding                                 3446 non-null   float64
 45  Equity_face_value                                  3446 non-null   float64
 46  EPS                                                4256 non-null   float64
 47  Adjusted_EPS                                       4256 non-null   float64
 48  Total_liabilities                                  4256 non-null   float64
 49  PE_on_BSE                                          1629 non-null   float64
dtypes: float64(50)
```

Figure 1

Data Dictionary:

Column	Description
Networth Next Year	Net worth of the customer in the next year
Total assets	Total assets of the customer
Net worth	The 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 the 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 customers 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 assets 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	The 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	The sum of all types of liabilities
PE on BSE	The company's current stock price divided by its earnings per share

Table 1

Statistical summary:

	count	mean	std	min	25%	50%	75%	max
Networth_Next_Year	4256.00	1344.74	15936.74	-74265.60	3.98	72.10	330.82	805773.40
Total_assets	4256.00	3573.62	30074.44	0.10	91.30	315.50	1120.80	1176509.20
Net_worth	4256.00	1351.95	12961.31	0.00	31.48	104.80	389.85	613151.60
Total_income	4025.00	4688.19	53918.95	0.00	107.10	455.10	1485.00	2442828.20
Change_in_stock	3706.00	43.70	436.92	-3029.40	-1.80	1.60	18.40	14185.50
Total_expenses	4091.00	4356.30	51398.09	-0.10	96.80	426.80	1395.70	2366035.30
Profit_after_tax	4102.00	295.05	3079.90	-3908.30	0.50	9.00	53.30	119439.10
PBDITA	4102.00	605.94	5646.23	-440.70	6.93	36.90	158.70	208576.50
PBT	4102.00	410.26	4217.42	-3894.80	0.80	12.60	74.17	145292.60
Cash_profit	4102.00	408.27	4143.93	-2245.70	2.90	19.40	96.25	176911.80
PBDITA_as_perc_of_total_income	4177.00	3.18	172.26	-6400.00	4.97	9.68	16.47	100.00
PBT_as_perc_of_total_income	4177.00	-18.20	419.91	-21340.00	0.56	3.34	8.94	100.00
PAT_as_perc_of_total_income	4177.00	-20.03	423.58	-21340.00	0.35	2.37	6.42	150.00
Cash_profit_as_perc_of_total_income	4177.00	-9.02	299.96	-15020.00	2.00	5.66	10.73	100.00
PAT_as_perc_of_net_worth	4256.00	10.17	61.53	-748.72	0.00	8.04	20.20	2466.67
Sales	3951.00	4645.68	53080.90	0.10	113.35	468.60	1481.20	2384984.40
Income_from_fincial_services	3145.00	81.36	1042.76	0.00	0.50	1.90	9.80	51938.20
Other_income	2700.00	55.95	1178.42	0.00	0.40	1.50	6.20	42856.70
Total_capital	4251.00	224.56	1684.95	0.10	13.20	42.60	103.15	78273.20
Reserves_and_funds	4158.00	1210.56	12816.23	-6525.90	5.30	55.15	282.52	625137.80
Borrowings	3825.00	1176.25	8581.25	0.10	24.40	99.80	358.30	278257.30
Current_liabilities_and_provisions	4146.00	960.63	9140.54	0.10	17.50	70.30	265.92	352240.30
Deferred_tax_liability	2887.00	234.50	2106.25	0.10	3.20	13.50	51.30	72796.60
Shareholders_funds	4256.00	1376.49	13010.69	0.00	32.30	107.60	408.90	613151.60
Cumulative_retained_profits	4211.00	937.18	9853.10	-6534.30	1.10	37.40	206.20	390133.80
Capital_employed	4256.00	2433.62	20496.40	0.00	61.30	221.20	790.30	891408.90
TOL_to_TNW	4256.00	4.03	20.88	-350.48	0.60	1.42	2.83	473.00

Figure 2

Total_term_liabilities__to__tangible_net_worth	4256.00	1.85	15.88	-325.60	0.05	0.34	1.00	456.00
Contingent_liabilities__to__Net_worth_perc	4256.00	55.71	369.17	0.00	0.00	5.36	31.01	14704.27
Contingent_liabilities	2854.00	948.55	12056.74	0.10	6.00	37.85	195.32	559506.80
Net_fixed_assets	4124.00	1209.49	12502.40	0.00	26.20	93.85	352.82	636604.60
Investments	2541.00	721.87	6793.86	0.00	1.00	8.20	63.80	199978.60
Current_assets	4176.00	1350.36	10155.57	0.10	36.60	148.35	515.00	354815.20
Net_working_capital	4219.00	162.87	3182.03	-63839.00	-1.10	16.70	86.50	85782.80
Quick_ratio_times	4151.00	1.50	9.33	0.00	0.41	0.67	1.03	341.00
Current_ratio_times	4151.00	2.26	12.48	0.00	0.93	1.23	1.72	505.00
Debt_to_equity_ratio_times	4256.00	2.87	15.60	0.00	0.22	0.79	1.75	456.00
Cash_to_current_liabilities_times	4151.00	0.53	4.80	0.00	0.02	0.07	0.19	165.00
Cash_to_average_cost_of_sales_per_day	4156.00	145.16	2521.99	0.00	2.88	8.04	21.97	128040.76
Creditors_turnover	3865.00	16.81	75.67	0.00	3.72	6.17	11.69	2401.00
Debtors_turnover	3871.00	17.93	90.16	0.00	3.81	6.47	11.85	3135.20
Finished_goods_turnover	3382.00	84.37	562.64	-0.09	8.19	17.32	40.01	17947.60
WIP_turnover	3492.00	28.68	169.65	-0.18	5.10	9.86	20.24	5651.40
Raw_material_turnover	3828.00	17.73	343.13	-2.00	3.02	6.41	11.82	21092.00
Shares_outstanding	3446.00	23764909.56	170979041.33	-2147483647.00	1308382.50	4750000.00	10906020.00	4130400545.00
Equity_face_value	3446.00	-1094.83	34101.36	-999998.90	10.00	10.00	10.00	100000.00
EPS	4256.00	-196.22	13061.95	-843181.82	0.00	1.49	10.00	34522.53
Adjusted_EPS	4256.00	-197.53	13061.93	-843181.82	0.00	1.24	7.62	34522.53
Total_liabilities	4256.00	3573.62	30074.44	0.10	91.30	315.50	1120.80	1176509.20
PE_on_BSE	1629.00	55.46	1304.45	-1116.64	2.97	8.69	17.00	51002.74

Figure 3

- As discussed before, we can see that there are null values present in almost all the columns.
- The average of net worth next year is around 1351.
- The profit before tax (PBT) has a mean and standard deviation of 410 & 4217 respectively.
- Most of the variables have a higher standard deviation indicating a wide spread and potential presence of outliers.
- More than half of the values in PE on BSE column are nulls.
- From the data it is evident that the variables are of different scales and it is recommended to perform scaling before model building.

Univariate & Multivariate analysis:

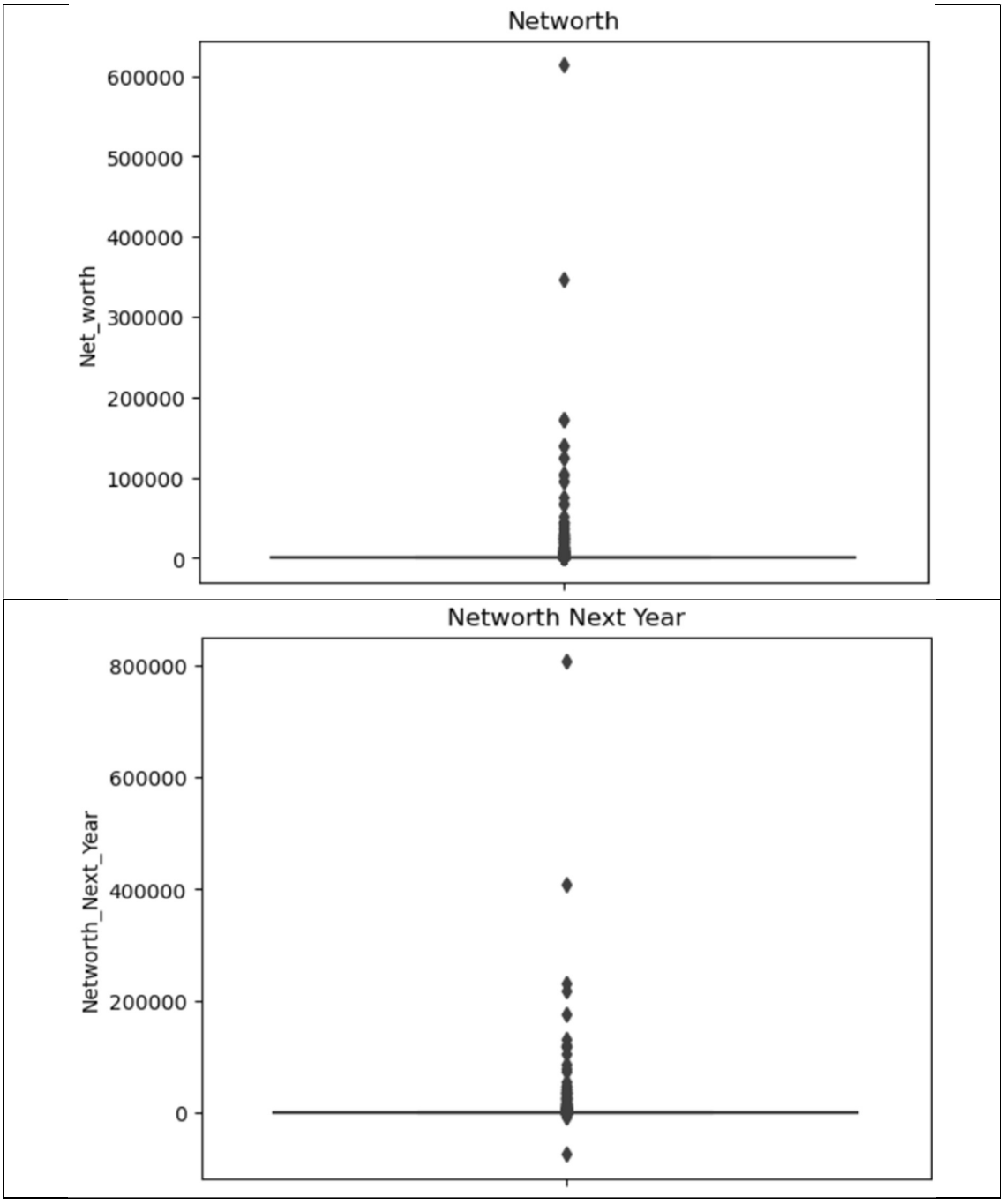

Figure 4

- For both net worth and net worth next year, the IQR, Q1, and Q3 are not visible due to the presence of extreme values.
- The distribution has high skewness.

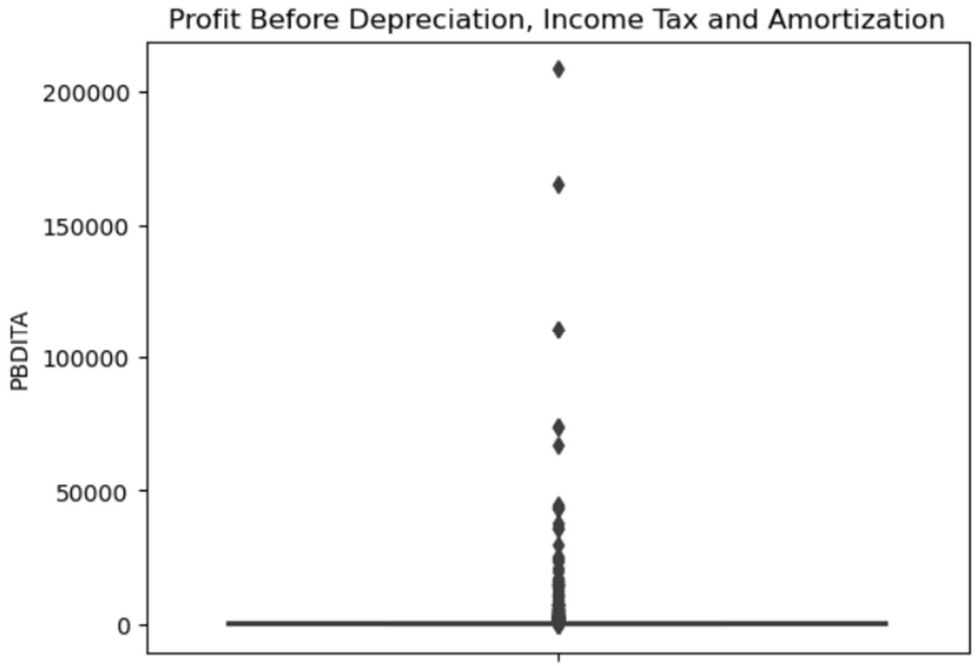

Figure 5

The PBDITA also follows a similar trend with the presence of outliers.

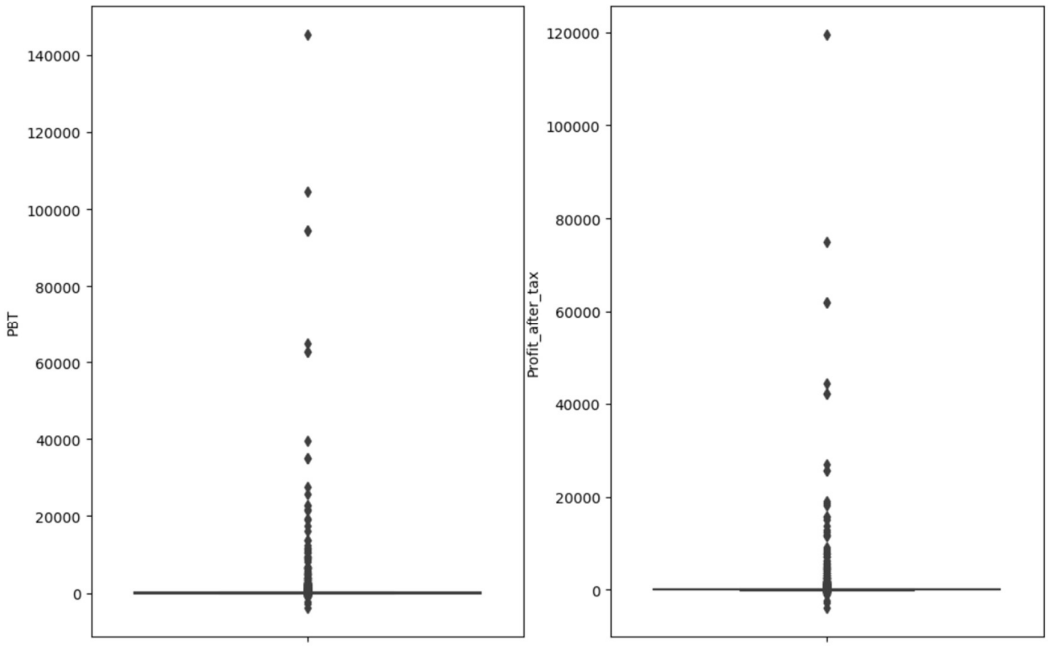

Figure 6

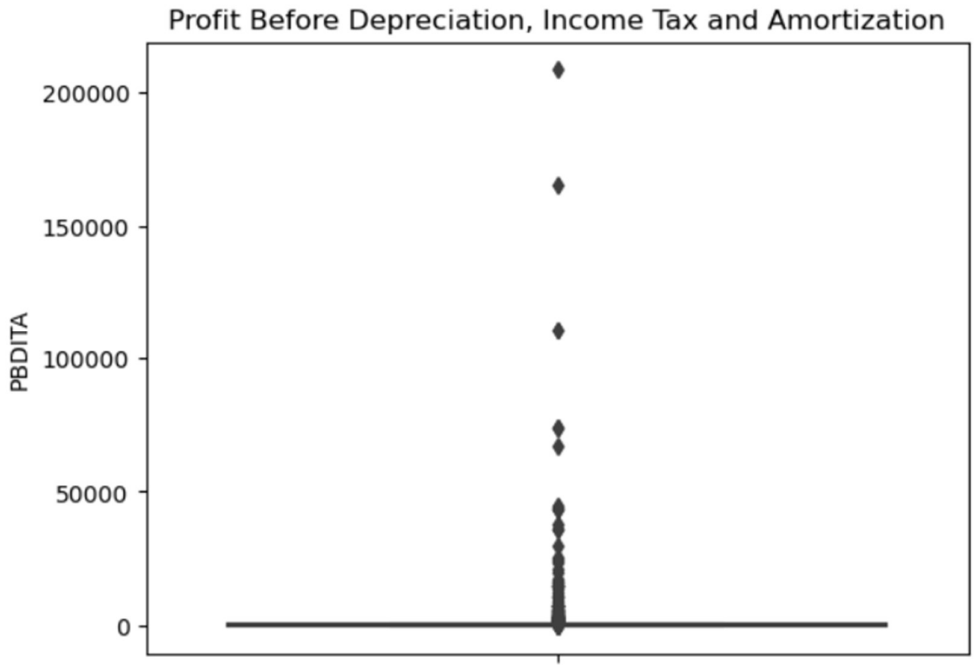

Figure 7

- PBDITA emphasizes operational efficiency, while PBT (Profit before tax) reflects overall profitability before tax deductions.
- Higher PBDITA indicates efficient management of operating costs.

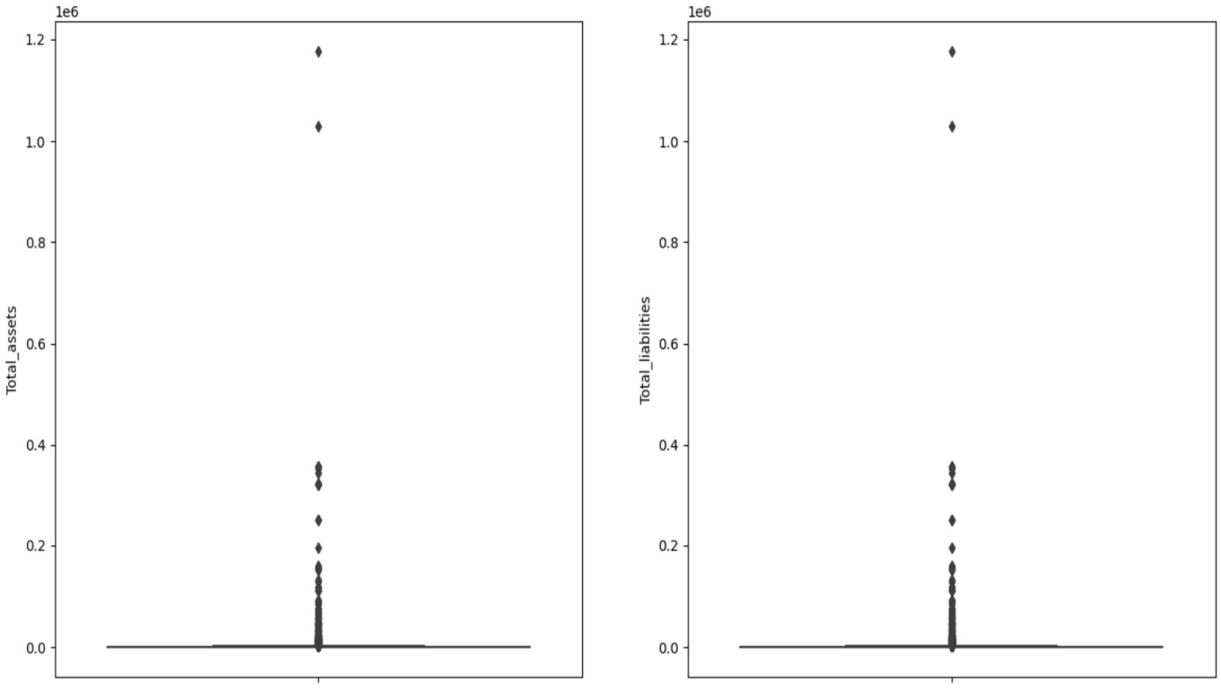

Figure 8

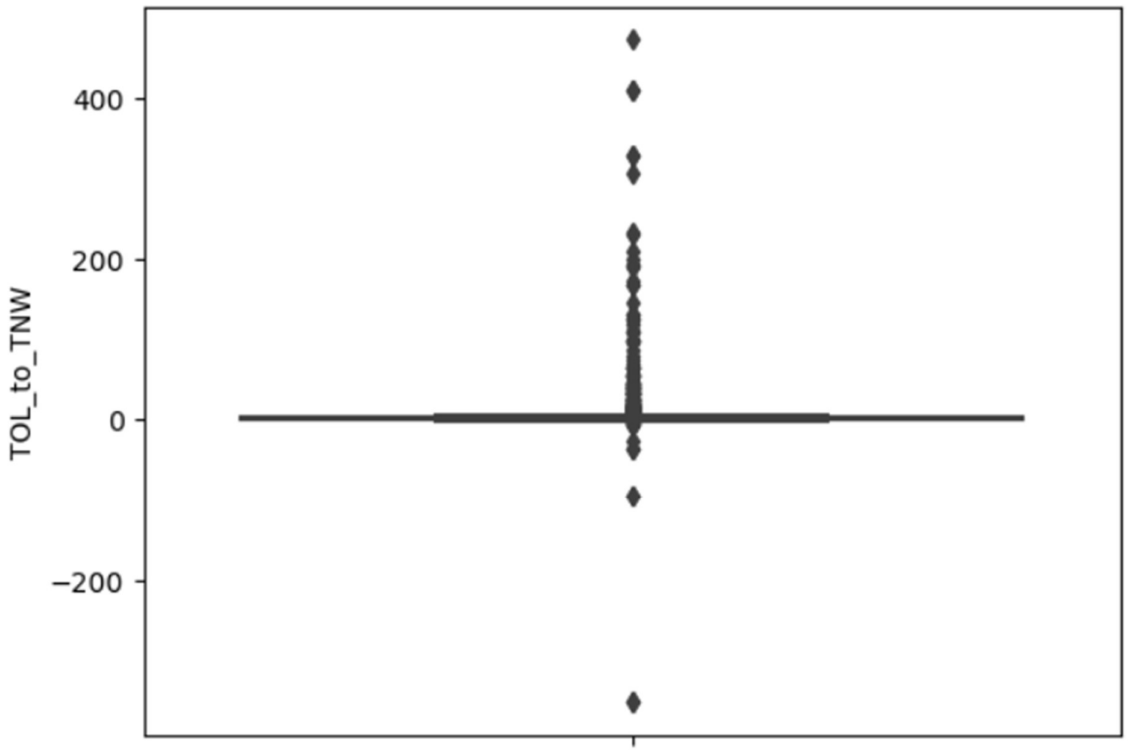

Figure 9

- Total liabilities of the customer divided by Total net worth
- A higher ratio suggests higher financial risk because the company relies more on borrowed funds.
- The distribution is right-skewed, indicating the presence of outliers above the right whiskers.

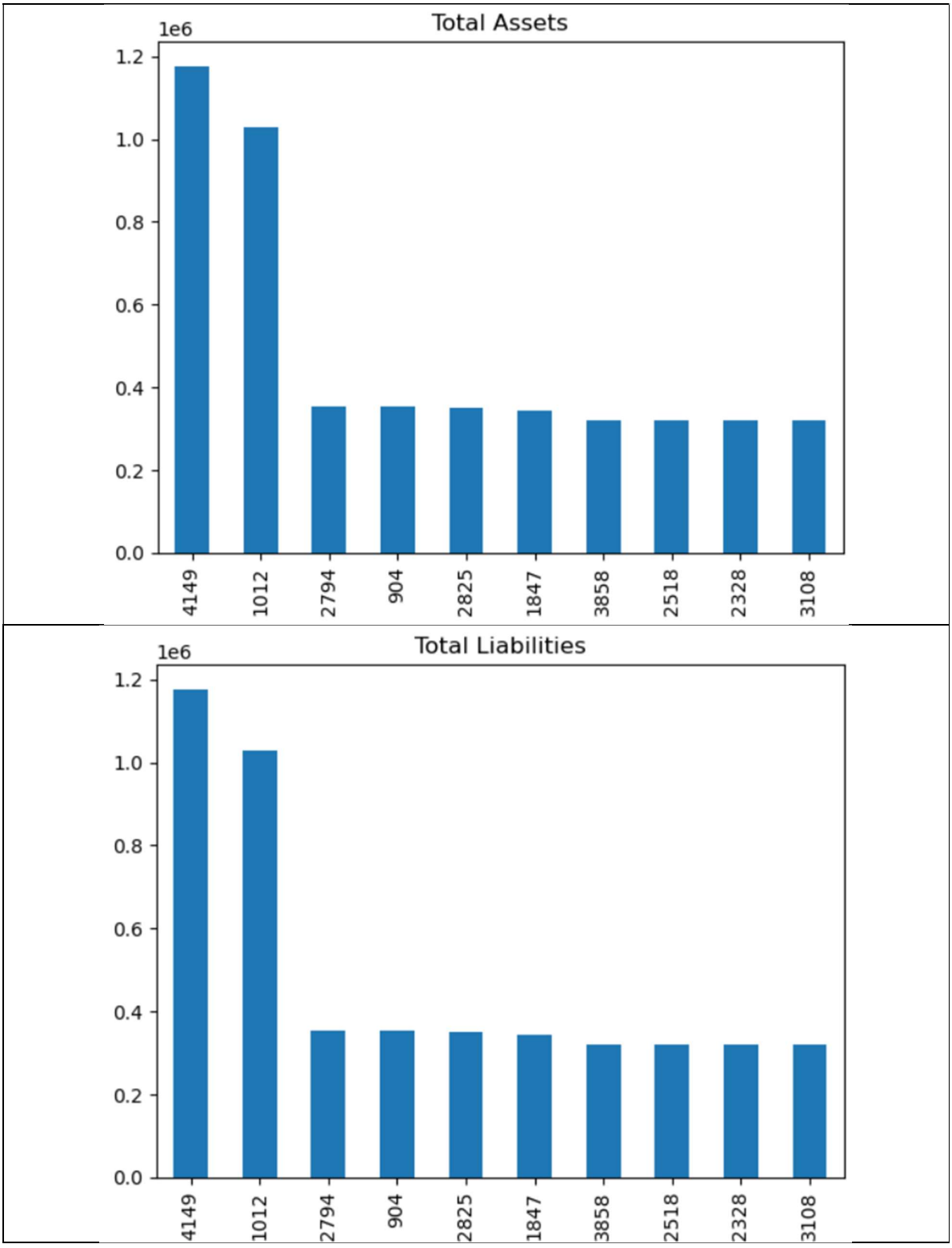

Figure 10

The above plot represents the top 10 companies by total assets and total liabilities.

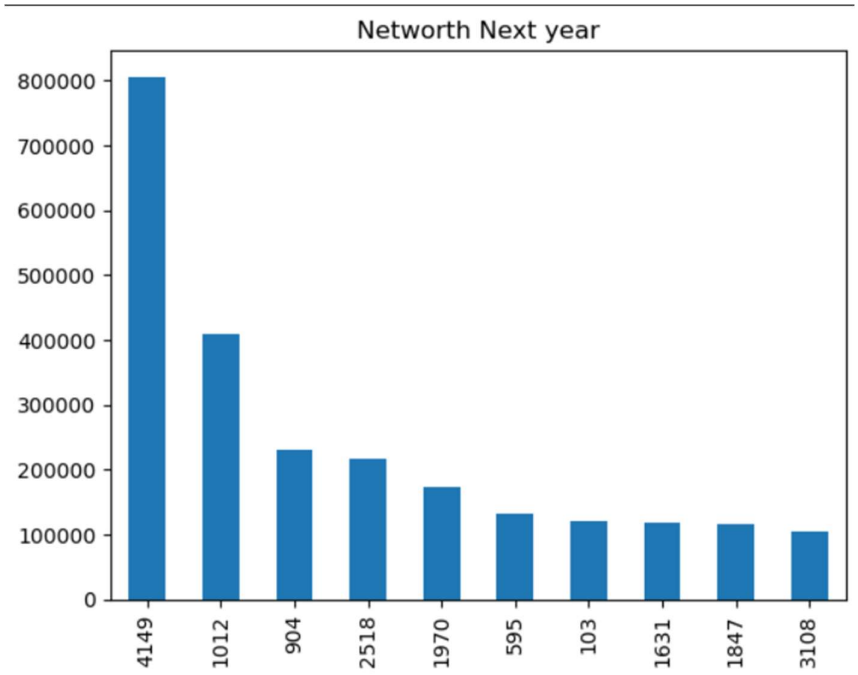

Figure 11

Given that the net worth for the next year is the main criterion for assessing potential defaulters, the box plot above showcases the top 10 companies based on their net worth for the upcoming year.

Operational Efficiency:

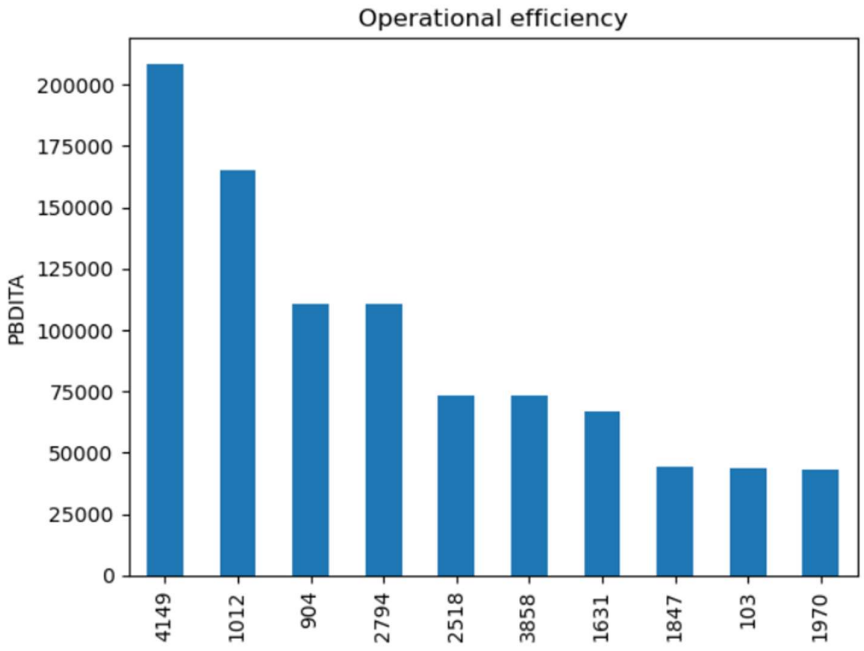

Figure 12

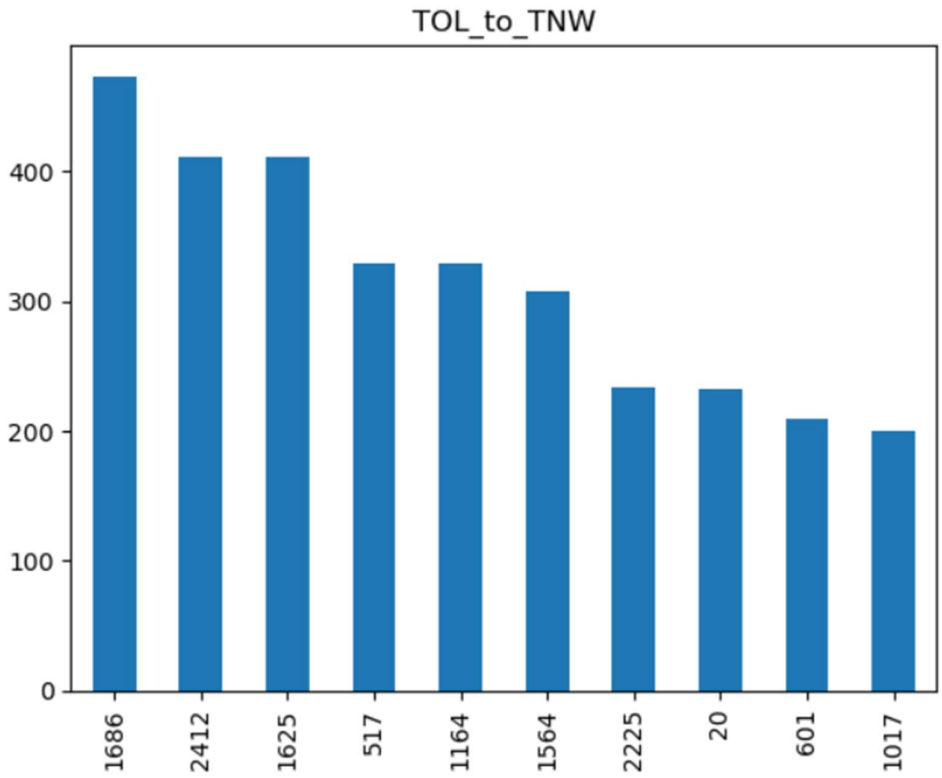

Figure 13

Companies with a higher TOL-TNW ratio suggest higher financial risk because the company relies more on borrowed funds.

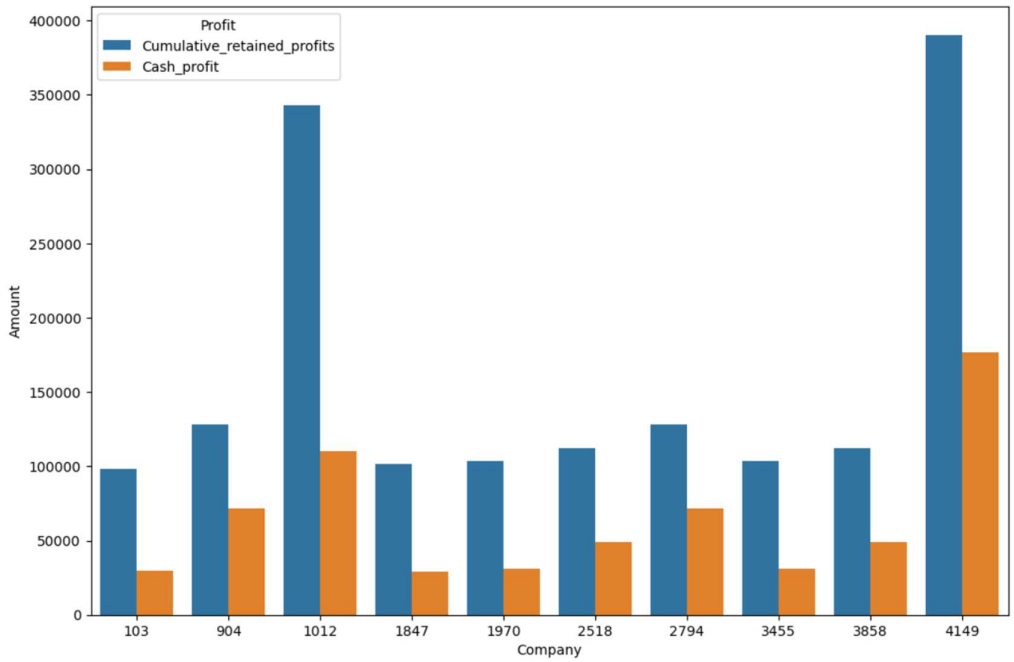

Figure 14

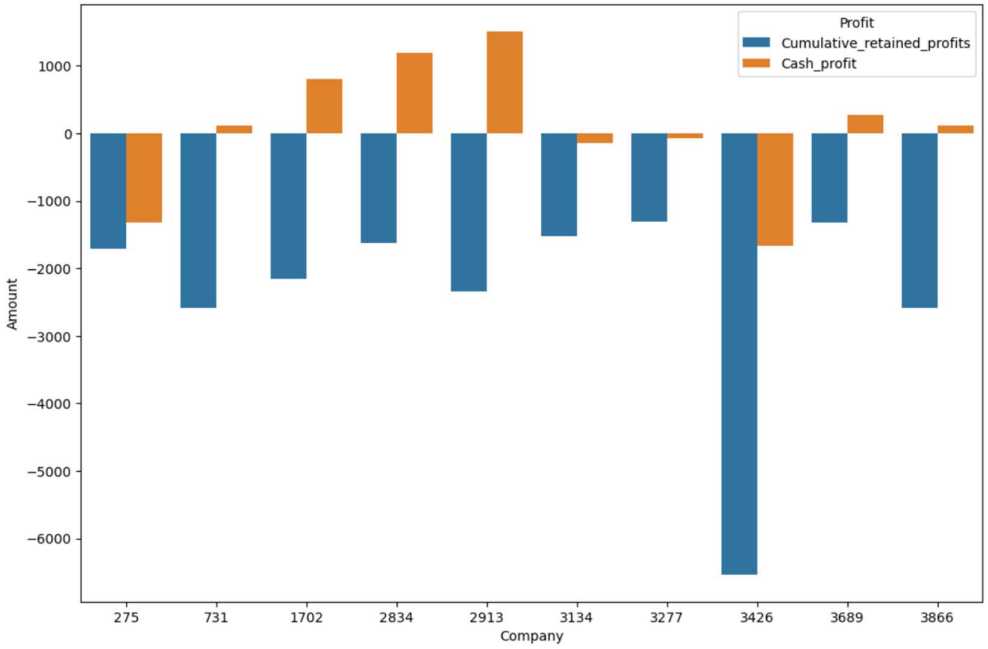

Figure 15

- Figure 14 shows that nearly half of the total profit for the top 10 companies is in the form of cash.
- It's a good indicator of the company's liquidity and financial health.
- Figure 15 shows the worst 10 companies in terms of cumulative profits.

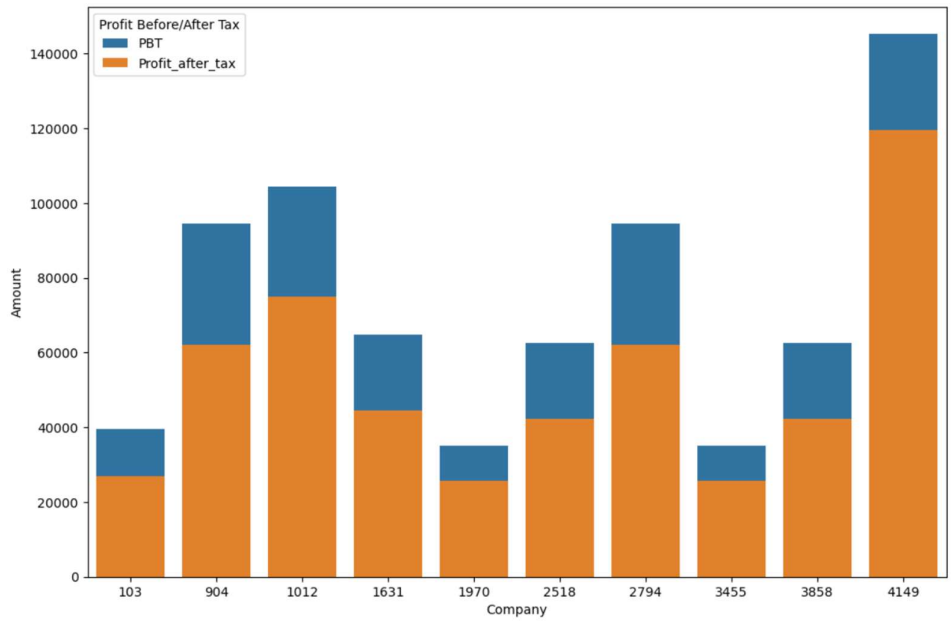

Figure 16

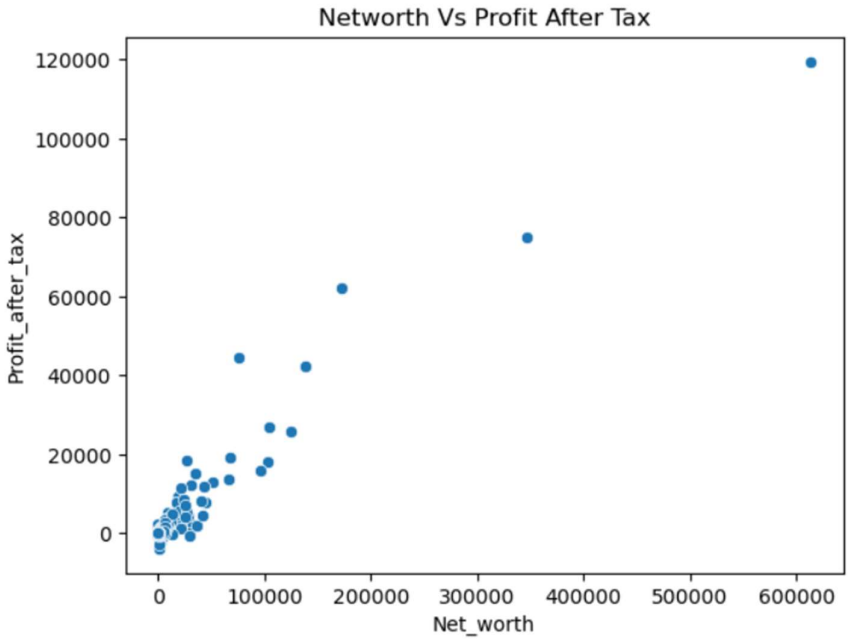

Figure 17

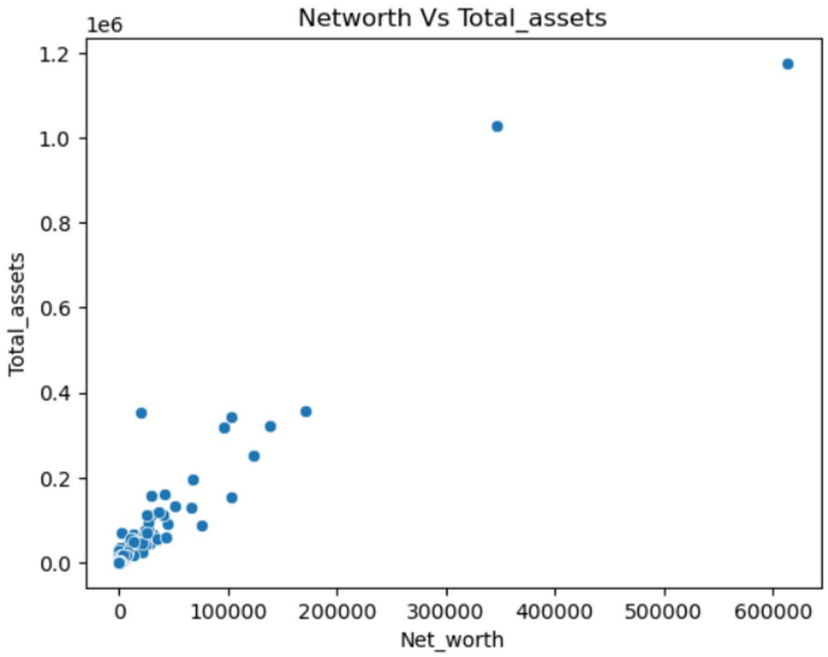

Figure 18

- There is a positive correlation between net worth and profits after tax.
- As the net worth increases, profits tend to increase.
- The same holds true for net worth and total assets.

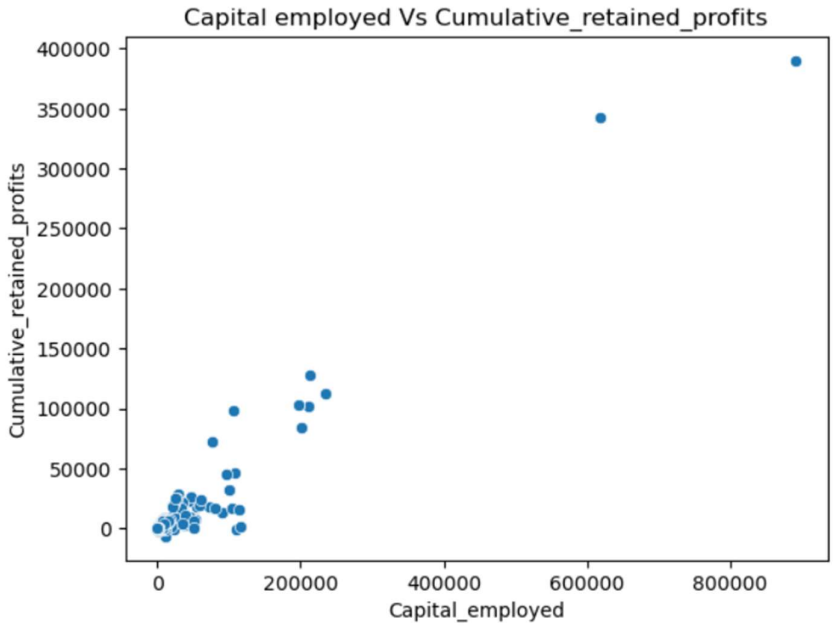

Figure 19

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.

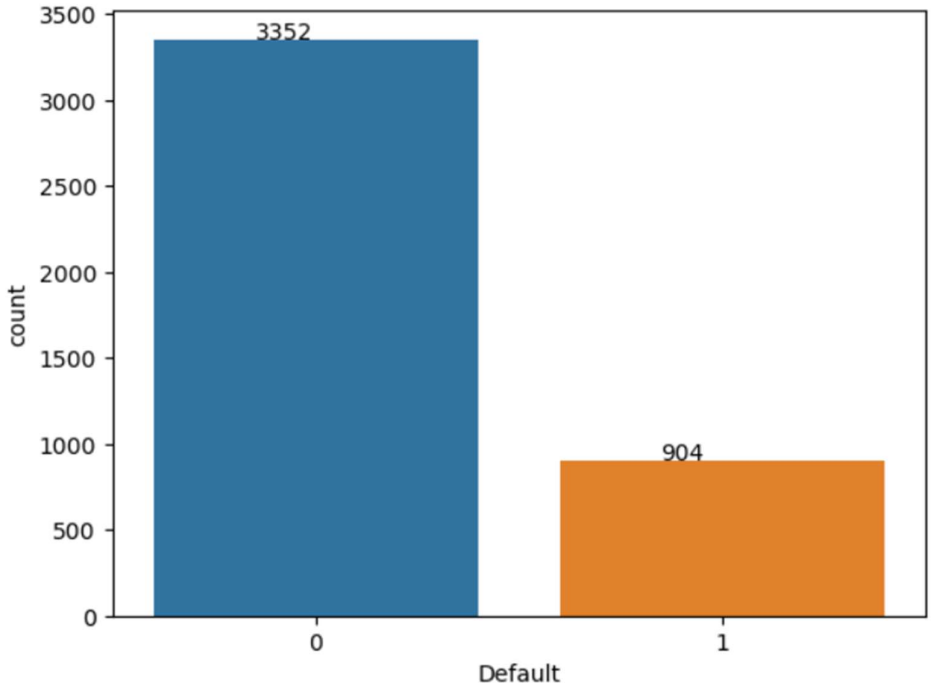

Figure 20

The number of companies that are considered as non-defaulters is 3 times higher than that of defaulters.

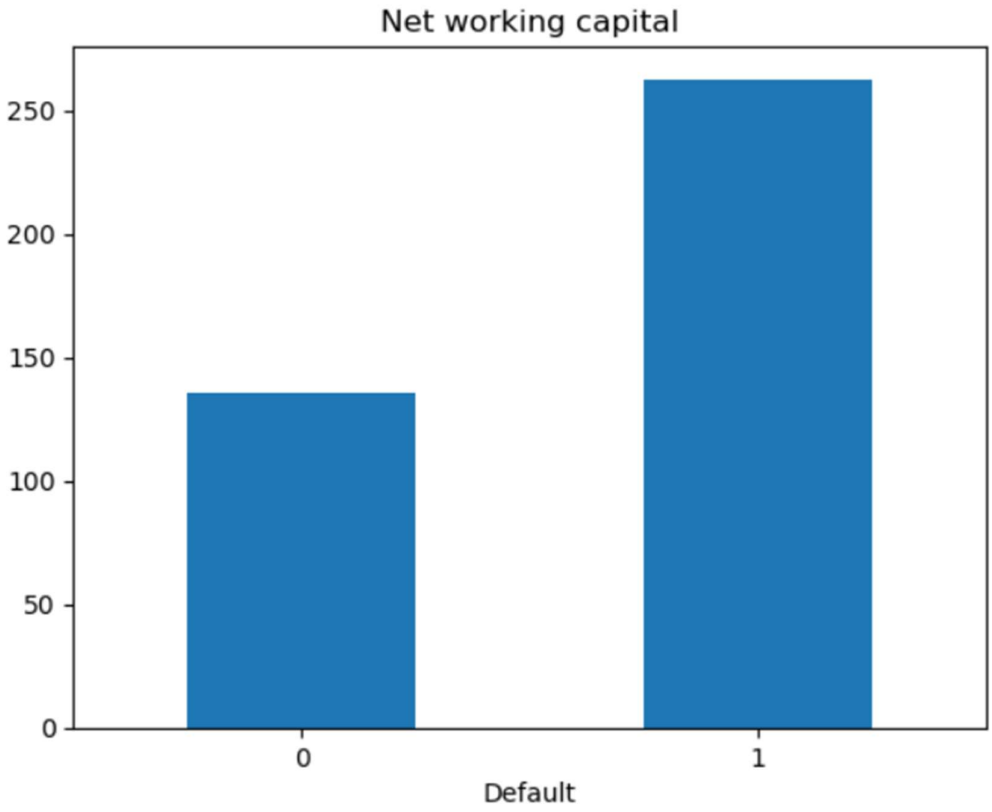

Figure 21

- Net working capital is the difference between the current liabilities and current assets.
- A higher net working indicates that the company has a higher share of liabilities.
- It is evident that those who tend to default have significantly higher average net working capital.

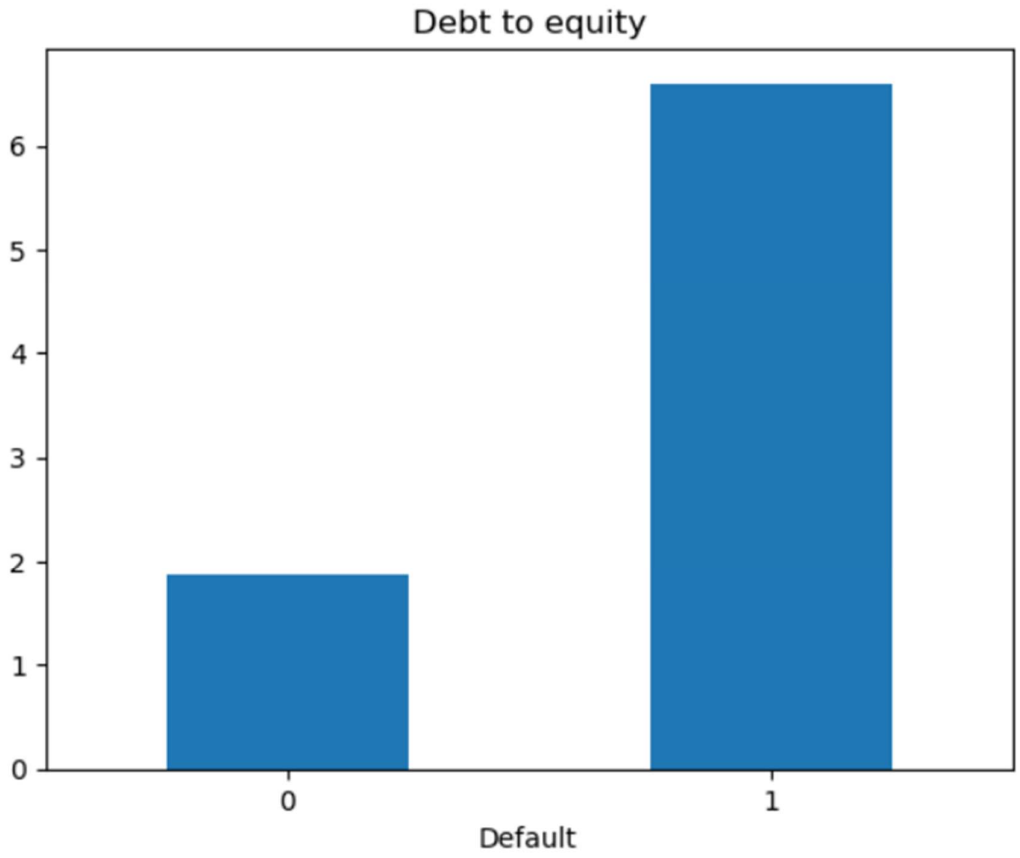

Figure 22

- A high debt-to-equity (D/E) ratio indicates that a company is heavily reliant on borrowed funds to finance its operations and growth.
- From the above plot, we can see that the defaulters have a 3 times higher D/E ratio as compared to the non-defaulters.

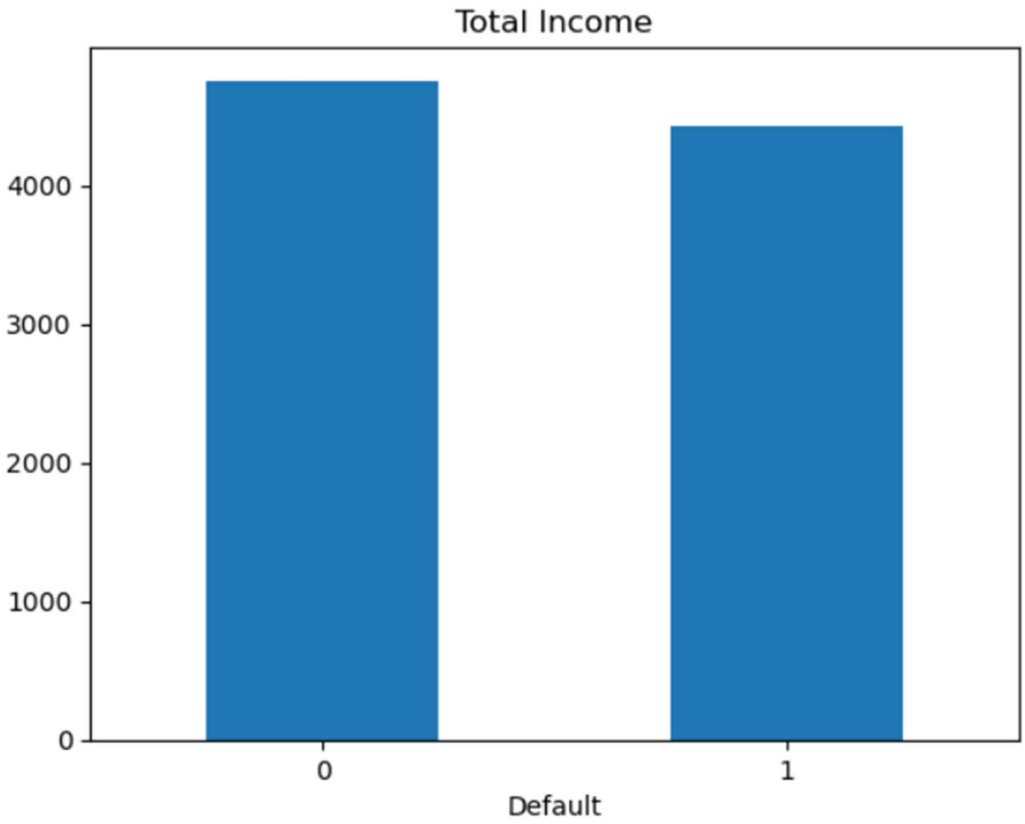

Figure 23

The non-defaulters have a slightly higher income comparatively.

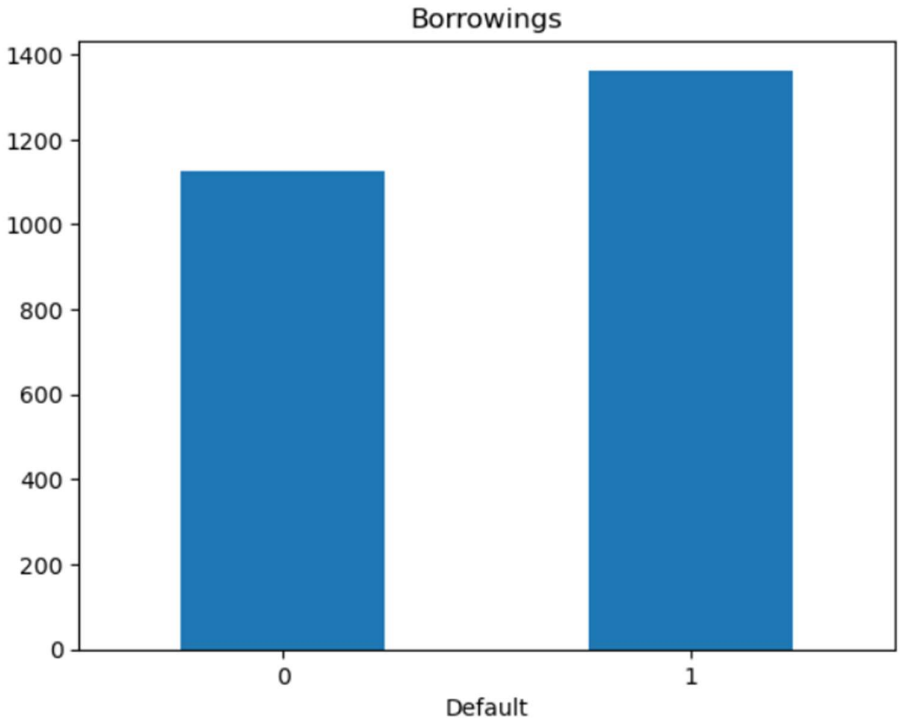

Figure 24

Defaulters have a higher average borrowing.

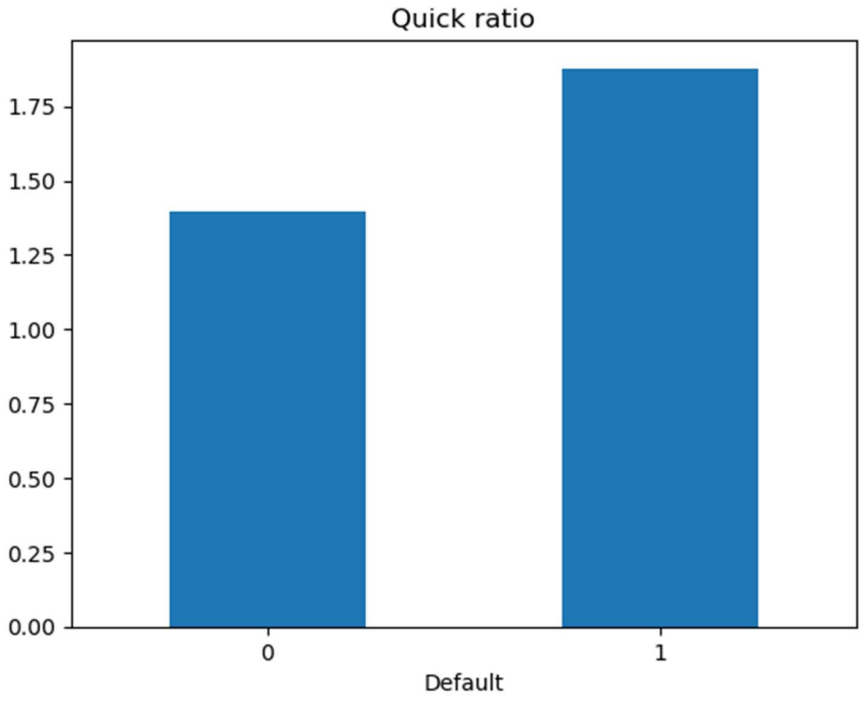

Figure 25

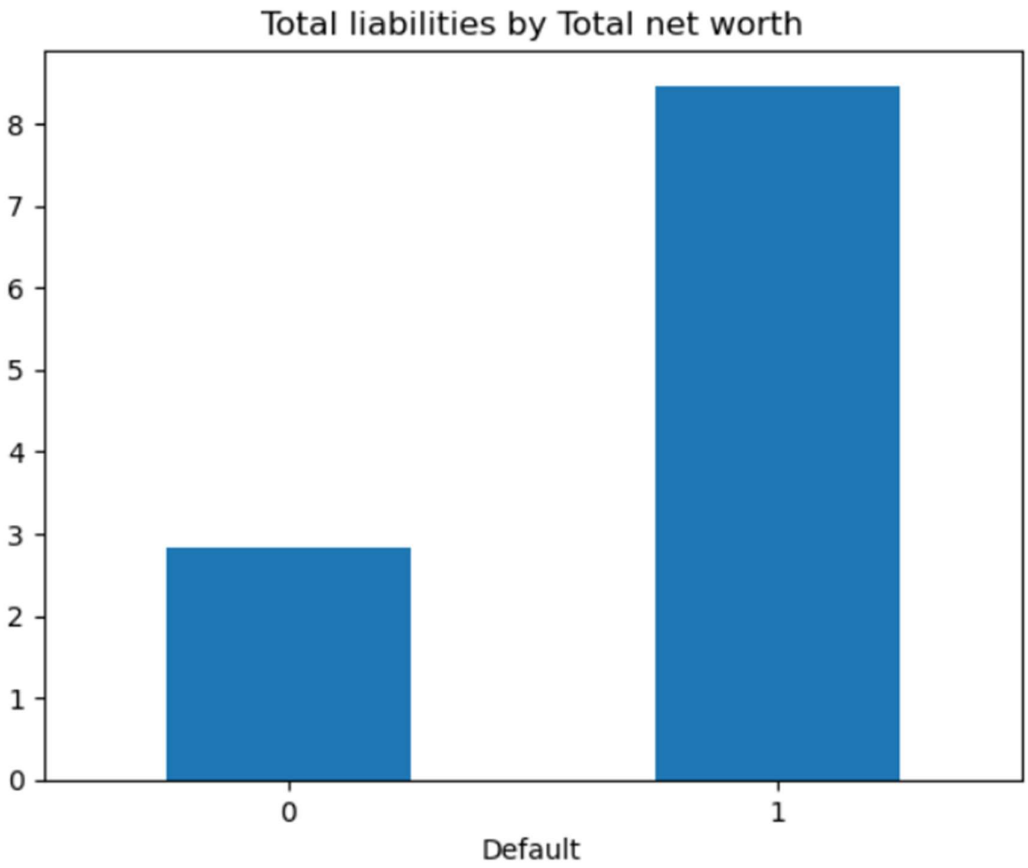

Figure 26

- High leverage means the company has significant debt obligations, which can be risky.
- The defaulters have a significantly higher liabilities to net worth ratio.

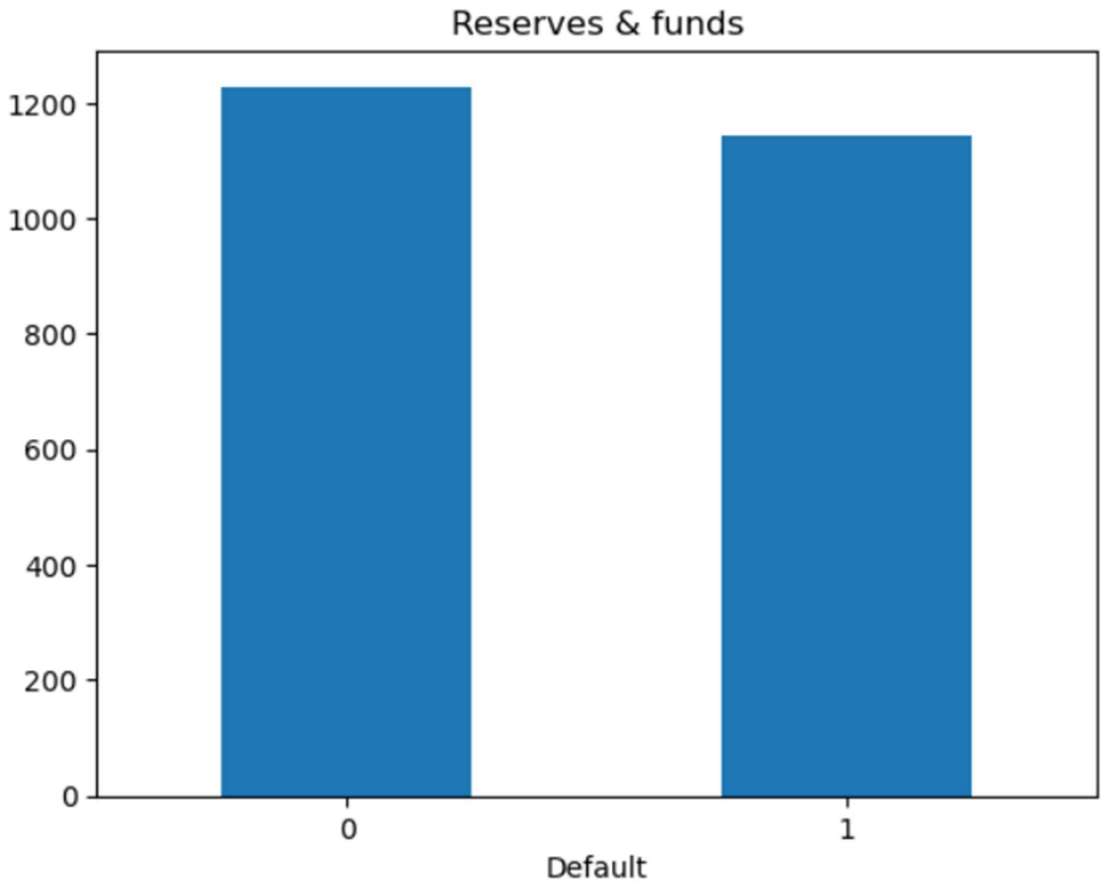

Figure 27

- High Reserves ensure that a company can continue operations smoothly during economic downturns or periods of low revenue.
- The companies that are tagged as non-default have a slightly higher reserve.

Figure 28

We can see that most of the variables are correlated to each other.

Let us filter the variables that have a correlation greater than 0.5.

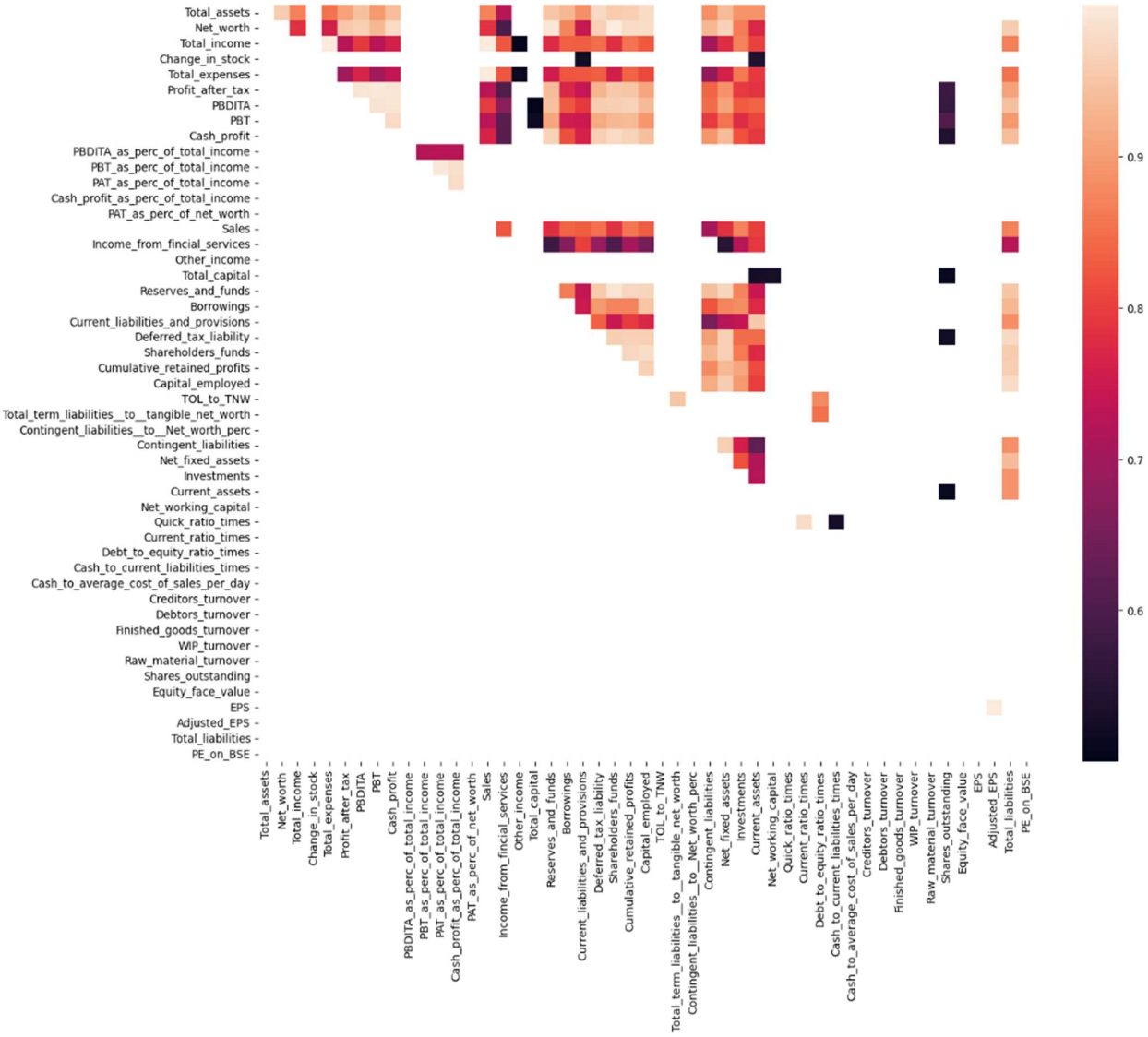

Figure 29

It is crucial to identify correlations and minimize them before model building, as highly correlated variables affect model performance.

Data Pre-processing:

Outlier detection and treatment:

- As discussed earlier, almost all the columns have outliers present.
- These outliers are expected to be present, especially in a dataset that contains financial information.
- Removing them will eliminate the chances of identifying potential default cases and the model will fail to predict real-world cases.
- Therefore, Instead of using traditional methods, we will impute the values above the 95th percentile and below the 5th percentile to nulls.
- Further it is recommended to drop columns with more than 30% of null values.
- From the figure below we can see that the columns Other_income,Deferred_tax_liability,Contingent_liabilities,Investments,PE_on_BSE have null values of more than 30% of the whole data. Therefore we are dropping them.
- After dropping we are left with 45 columns.
- After checking the null values by columns, we are filtering the rows that have non-null values for at least 90% of their attributes.

Total_assets	2.07
Net_worth	2.09
Total_income	7.42
Change_in_stock	14.97
Total_expenses	5.76
Profit_after_tax	5.69
PBDITA	5.64
PBT	5.76
Cash_profit	5.76
PBDITA_as_perc_of_total_income	3.76
PBT_as_perc_of_total_income	3.97
PAT_as_perc_of_total_income	4.32
Cash_profit_as_perc_of_total_income	3.92
PAT_as_perc_of_net_worth	1.15
Sales	9.12
Income_from_fincial_services	27.91
Other_income	38.18
Total_capital	2.00
Reserves_and_funds	4.39
Borrowings	12.01
Current_liabilities_and_provisions	4.70
Deferred_tax_liability	33.83
Shareholders_funds	2.04
Cumulative_retained_profits	3.15
Capital_employed	1.93
TOL_to_TNW	2.30
Total_term_liabilities__to__tangible_net_worth	2.28
Contingent_liabilities__to__Net_worth_perc	1.86
Contingent_liabilities	34.33
Net_fixed_assets	5.24
Investments	41.85
Current_assets	3.92
Net_working_capital	3.97
Quick_ratio_times	4.23
Current_ratio_times	4.18
Debt_to_equity_ratio_times	1.97
Cash_to_current_liabilities_times	4.44
Cash_to_average_cost_of_sales_per_day	4.75
Creditors_turnover	10.55
Debtors_turnover	10.88
Finished_goods_turnover	22.37
WIP_turnover	19.17
Raw_material_turnover	11.30
Shares_outstanding	20.65
Equity_face_value	20.14
EPS	3.29
Adjusted_EPS	3.43
Total_liabilities	2.07
PE_on_BSE	62.69
Default	0.00

Figure 30

Imputing null values:

MICE (Multiple Imputation by Chained Equation) impute missing values with reasonable estimates (e.g., mean, median, or regression predictions)

Before imputing

...	Debtors_turnover	Finished_goods_turnover	WIP_turnover	Raw_material_turnover	Shares_outstanding	Equity_face_value	EPS	Adjusted_EPS
...	5.65	3.99	3.37	14.87	8760056.00	10.00	4.44	4.44
...	NaN	NaN	NaN	NaN	NaN	NaN	0.00	0.00
...	2.51	17.67	8.76	8.35	NaN	NaN	0.00	0.00
...	1.91	18.14	18.62	11.11	10000000.00	10.00	17.60	17.60
...	68.00	45.87	28.67	19.93	107315.00	100.00	-6.52	-6.52

Figure 31

After imputing

Debtors_turnover	Finished_goods_turnover	WIP_turnover	Raw_material_turnover	Shares_outstanding	Equity_face_value	EPS	Adjusted_EPS
5.65	3.99	3.37	14.87	8760056.00	10.00	4.44	4.44
2.51	17.67	8.76	8.35	4055096.94	16.44	0.00	0.00
1.91	18.14	18.62	11.11	10000000.00	10.00	17.60	17.60
68.00	45.87	28.67	19.93	107315.00	100.00	-6.52	-6.52
7.25	5.73	4.62	3.72	3807100.00	10.00	12.69	0.63

Figure 32

We are also dropping the equity face value column, as its values remain almost constant for all the companies.

After imputing the null Values

```
Total_assets                                    0
Net_worth                                       0
Total_income                                    0
Change_in_stock                                 0
Total_expenses                                  0
Profit_after_tax                                0
PBDITA                                          0
PBT                                             0
Cash_profit                                     0
PBDITA_as_perc_of_total_income                  0
PBT_as_perc_of_total_income                     0
PAT_as_perc_of_total_income                     0
Cash_profit_as_perc_of_total_income             0
PAT_as_perc_of_net_worth                        0
Sales                                           0
Income_from_fincial_services                    0
Total_capital                                   0
Reserves_and_funds                              0
Borrowings                                      0
Current_liabilities_and_provisions              0
Shareholders_funds                              0
Cumulative_retained_profits                     0
Capital_employed                                0
TOL_to_TNW                                      0
Total_term_liabilities__to__tangible_net_worth  0
Contingent_liabilities__to__Net_worth_perc      0
Net_fixed_assets                                0
Current_assets                                  0
Net_working_capital                             0
Quick_ratio_times                               0
Current_ratio_times                             0
Debt_to_equity_ratio_times                      0
Cash_to_current_liabilities_times               0
Cash_to_average_cost_of_sales_per_day           0
Creditors_turnover                              0
Debtors_turnover                                0
Finished_goods_turnover                         0
WIP_turnover                                    0
Raw_material_turnover                           0
Shares_outstanding                              0
EPS                                             0
Adjusted_EPS                                    0
Total_liabilities                               0
dtype: int64
```

Figure 33

Target Creation:

The target variable is default and should take the value 1 when net worth next year is negative & 0 when net worth next year is positive.

Default	Count
0	3352
1	904

Table 2

Train-Test Split:

X-Train shape: 1819, 43

X-Test shape: 780, 43

y-Train proportion of default:

Default	Proportion
0	0.8
1	0.2

Table 3

y-Test proportion of default:

Default	Proportion
0	0.81
1	0.19

Table 4

Scaling:

Scaling ensures that features with larger ranges do not dominate, allowing all features to contribute equally to the model.

The Standard Scaler from scikit-learn is used for scaling the data. It standardizes features by removing the mean and changing the standard deviation to 1.

X-Train Scaled

	Total_assets	Net_worth	Total_income	Change_in_stock	Total_expenses	Profit_after_tax	PBDITA	PBT	Cash_profit	PBDITA_as_perc_of_total_income
0	-0.45	-0.29	-0.54	-0.22	-0.54	-0.29	-0.42	-0.30	-0.38	1.58
1	-0.48	-0.44	-0.52	-0.23	-0.52	-0.35	-0.44	-0.36	-0.41	0.38
2	-0.51	-0.50	-0.41	-0.24	-0.41	-0.37	-0.48	-0.37	-0.45	-0.93
3	0.19	0.40	0.56	2.30	0.65	-0.11	-0.15	-0.07	-0.25	-0.66
4	-0.39	-0.42	-0.47	-0.09	-0.49	-0.03	-0.22	-0.06	-0.14	2.33

Figure 34

X-Test Scaled

	Total_assets	Net_worth	Total_income	Change_in_stock	Total_expenses	Profit_after_tax	PBDITA	PBT	Cash_profit	PBDITA_as_perc_of_total_income
0	-0.51	-0.45	-0.53	-0.27	-0.53	-0.35	-0.48	-0.40	-0.42	0.17
1	-0.35	-0.38	-0.45	-0.14	-0.45	-0.29	-0.33	-0.29	-0.29	0.72
2	-0.48	-0.52	-0.49	-0.28	-0.48	-0.49	-0.49	-0.49	-0.46	-0.54
3	-0.45	-0.43	-0.49	-0.17	-0.49	-0.36	-0.46	-0.38	-0.42	-0.07
4	-0.47	-0.45	-0.50	-0.12	-0.50	-0.35	-0.45	-0.36	-0.43	0.09

Figure 35

Model Building:

Logistic Regression using Sklearn:

Classification report for Train data:

	precision	recall	f1-score	support
0	0.81	1.00	0.89	1463
1	0.67	0.02	0.04	356
accuracy			0.81	1819
macro avg	0.74	0.51	0.47	1819
weighted avg	0.78	0.81	0.73	1819

Figure 36

Classification report for Test Data:

	precision	recall	f1-score	support
0	0.80	0.99	0.89	628
1	0.11	0.01	0.01	152
accuracy			0.80	780
macro avg	0.46	0.50	0.45	780
weighted avg	0.67	0.80	0.72	780

Figure 37

- The model performs well for class 0 which is for non-defaulters.
- The model performance for class 1 is poor due to class imbalance.

Confusion matrix for Train data:

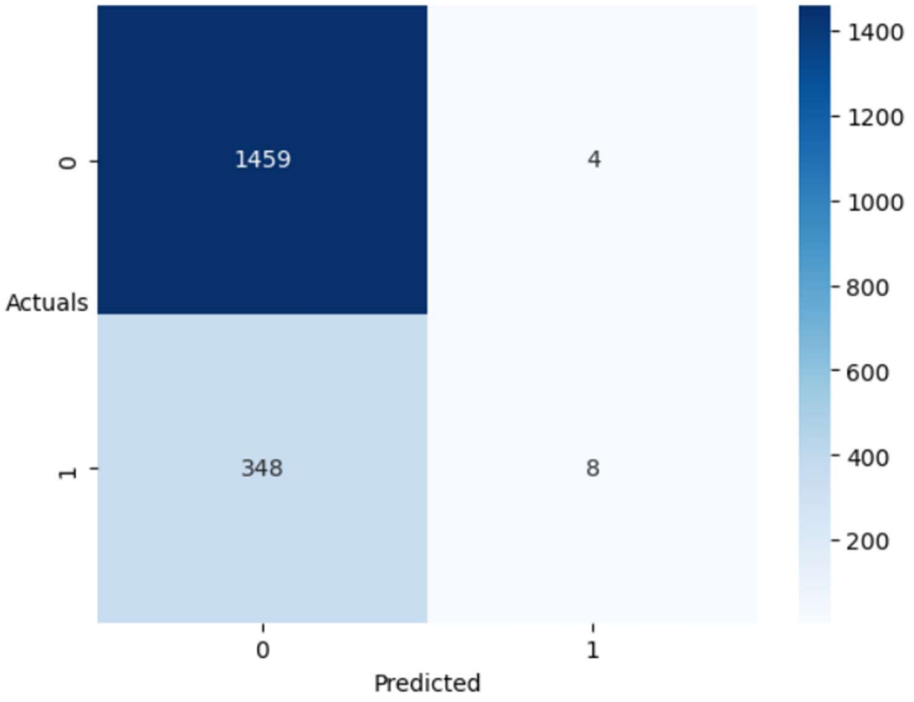

Figure 38

Confusion matrix for Test data:

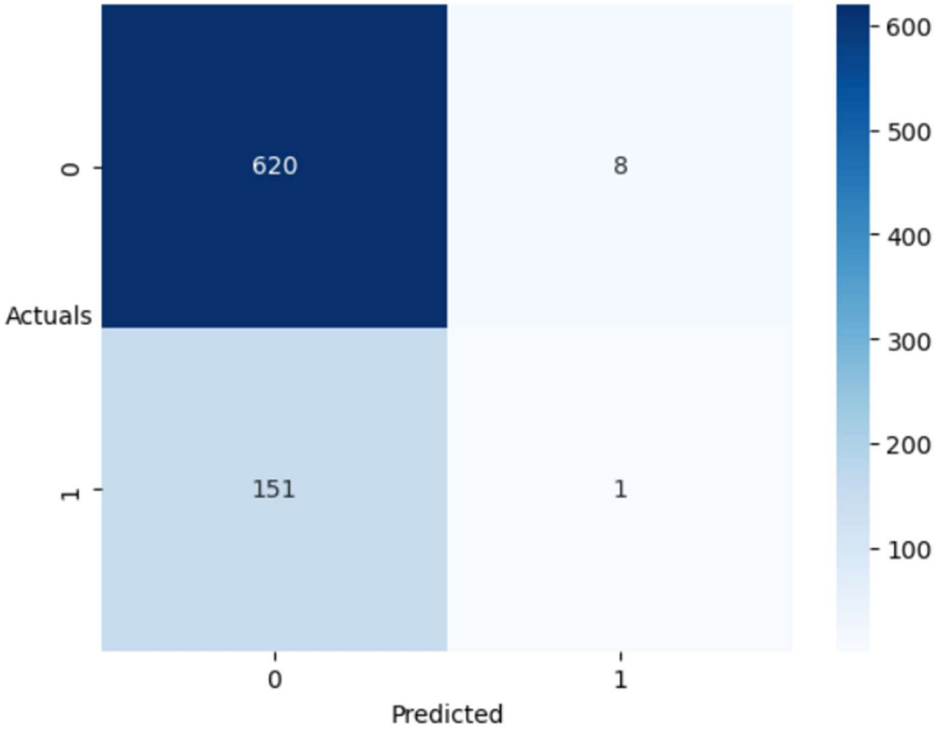

Figure 39

Random Forest:

Classification report for Train Data

	precision	recall	f1-score	support
0	0.97	0.99	0.98	1463
1	0.96	0.86	0.91	356
accuracy			0.97	1819
macro avg	0.96	0.93	0.94	1819
weighted avg	0.97	0.97	0.96	1819

Figure 40

Classification report for Test Data

	precision	recall	f1-score	support
0	0.80	0.86	0.83	628
1	0.19	0.13	0.15	152
accuracy			0.72	780
macro avg	0.49	0.50	0.49	780
weighted avg	0.68	0.72	0.70	780

Figure 41

- The model performance on test data is poor due to overfitting on train data which is a usual case in the random forest model.
- We cannot rely on accuracy due to class imbalance.
- For use recall for call 1 is crucial, as it identifies the true negatives which is a defaulter in this case.

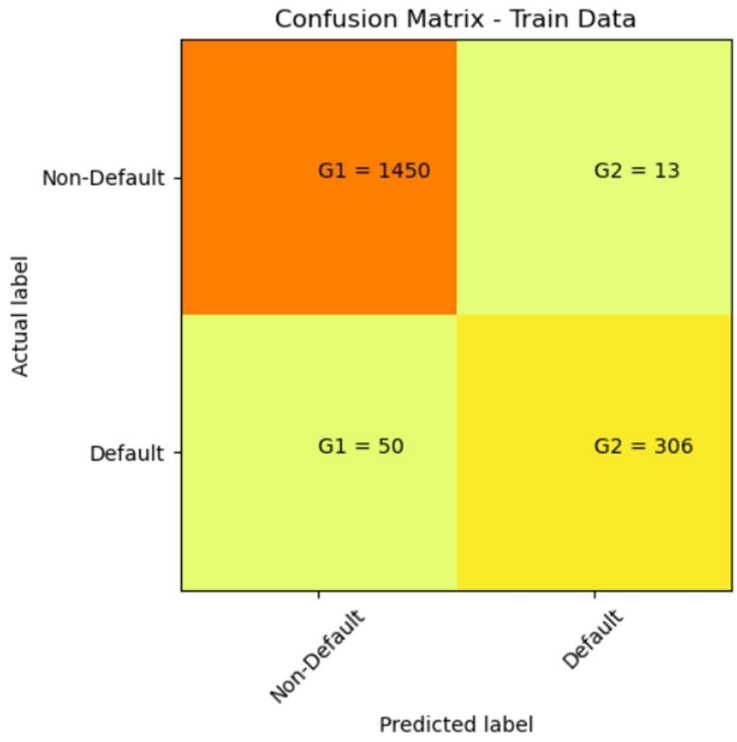

Figure 42

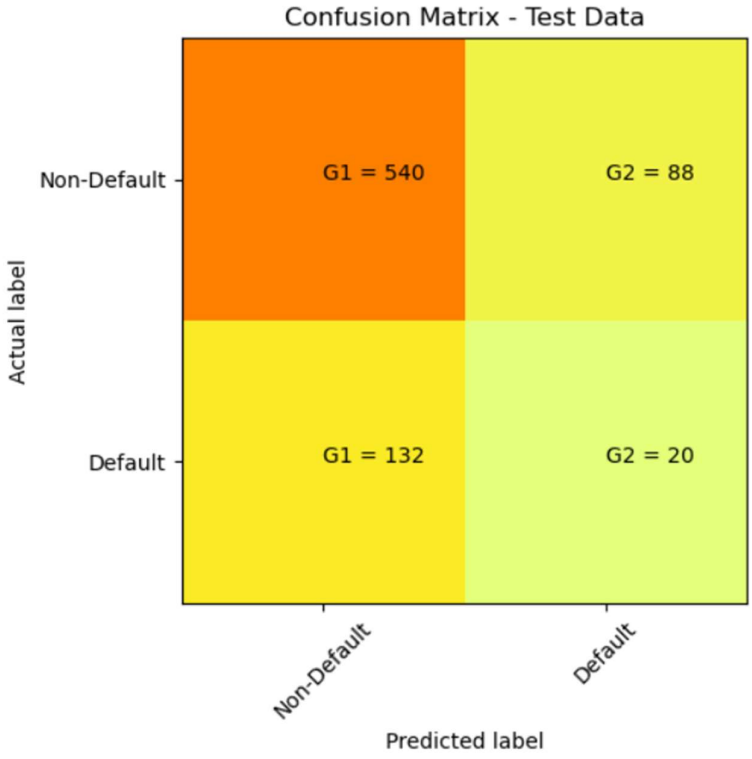

Figure 43

Model Performance Improvement:

Logistic Regression using Sklearn:

Feature selection using Recursive Feature Elimination:

- Recursive Feature Elimination (RFE) is a feature selection technique used to identify the most important features in a dataset.
- The least important features are removed from the dataset.
- Lower ranks indicate more important features.

Features with Rank 1

	Feature	Rank
0	Total_assets	1
6	PBDITA	1
7	PBT	1
8	Cash_profit	1
14	Sales	1
18	Borrowings	1
20	Shareholders_funds	1
21	Cumulative_retained_profits	1
26	Net_fixed_assets	1
42	Total_liabilities	1

Figure 44

Building a model using only these columns as the predictor.

Classification report for Train data:

	precision	recall	f1-score	support
0	0.81	1.00	0.89	1463
1	0.86	0.02	0.03	356
accuracy			0.81	1819
macro avg	0.83	0.51	0.46	1819
weighted avg	0.82	0.81	0.72	1819

Figure 45

Classification report for Test data

	precision	recall	f1-score	support
0	0.81	1.00	0.89	628
1	0.50	0.01	0.03	152
accuracy			0.81	780
macro avg	0.65	0.50	0.46	780
weighted avg	0.75	0.81	0.72	780

Figure 46

Still, the recall for class 1 is poor. Let us modify the class threshold from a default value of 0.5 and find an optimal value using the roc curve.

From the roc curve, a value of 0.198 found to be the optimal value.

Classification report for Train data:

	precision	recall	f1-score	support
0	0.83	0.72	0.77	1463
1	0.25	0.38	0.30	356
accuracy			0.66	1819
macro avg	0.54	0.55	0.54	1819
weighted avg	0.71	0.66	0.68	1819

Figure 47

Classification report for Test data:

	precision	recall	f1-score	support
0	0.83	0.39	0.53	628
1	0.21	0.66	0.32	152
accuracy			0.44	780
macro avg	0.52	0.53	0.42	780
weighted avg	0.71	0.44	0.49	780

Figure 48

A recall of 0.66 on test data is good. Let us try building a model using stats model.

Logistic Regression using stats model:

In order to use the stats model, we need to combine both the predictor and the response variable.

Train set Proportion

Default	Proportion
0	0.81
1	0.19

Table 5

Test set proportion

Default	Proportion
0	0.83
1	0.17

Table 6

Building a function called calculate VIF and drop which calculates the variance inflation factor for each column in the data set, filters columns having VIF greater than 5, drops the column with the maximum VIF, and reruns the loop until all the columns have VIF less than 5.

After running the loop for all the 44 columns, the selected columns with VIF less than 5 are

Final VIF values:

	variables	VIF
0	Change_in_stock	1.19
1	Profit_after_tax	2.88
2	PBDITA_as_perc_of_total_income	2.07
3	PAT_as_perc_of_total_income	2.16
4	PAT_as_perc_of_net_worth	1.79
5	Income_from_fincial_services	2.14
6	Total_capital	2.76
7	Reserves_and_funds	3.59
8	Borrowings	2.37
9	Current_liabilities_and_provisions	3.41
10	TOL_to_TNW	3.59
11	Total_term_liabilities__to__tangible_net_worth	3.50
12	Contingent_liabilities__to__Net_worth_perc	1.13
13	Net_working_capital	1.72
14	Current_ratio_times	1.81
15	Cash_to_current_liabilities_times	3.89
16	Cash_to_average_cost_of_sales_per_day	3.30
17	Creditors_turnover	1.23
18	Debtors_turnover	1.16
19	Finished_goods_turnover	1.59
20	WIP_turnover	1.66
21	Raw_material_turnover	1.13
22	Shares_outstanding	2.86
23	Adjusted_EPS	1.15

Figure 49

Building a model using the formula below

Default~Change_in_stock+Profit_after_tax+PBDITA_as_perc_of_total_income+PAT_as_perc_of_total_income+PAT_as_perc_of_net_worth+Income_from_fincial_services+Total_capital+Reserves_and_funds+Borrowings+Current_liabilities_and_provisions+TOL_to_TNW+Total_term_liabilities__to__tangible_net_worth+Contingent_liabilities__to__Net_worth_perc+Net_working_capital+Current_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+Adjusted_EPS

This creates a relationship between the predictor and the response variable.

Model summary:

Dep. Variable:	Default	**No. Observations:**	1003
Model:	Logit	**Df Residuals:**	978
Method:	MLE	**Df Model:**	24
Date:	Fri, 26 Jul 2024	**Pseudo R-squ.:**	0.04604
Time:	20:38:09	**Log-Likelihood:**	-461.65
converged:	True	**LL-Null:**	-483.93
Covariance Type:	nonrobust	**LLR p-value:**	0.006558

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.5260	0.085	-17.901	0.000	-1.693	-1.359
Change_in_stock	-0.0531	0.082	-0.651	0.515	-0.213	0.107
Profit_after_tax	0.1201	0.128	0.941	0.347	-0.130	0.370
PBDITA_as_perc_of_total_income	0.0208	0.117	0.177	0.859	-0.209	0.250
PAT_as_perc_of_total_income	0.0550	0.130	0.423	0.672	-0.200	0.310
PAT_as_perc_of_net_worth	-0.0717	0.099	-0.725	0.469	-0.266	0.122
Income_from_fincial_services	-0.2615	0.148	-1.772	0.076	-0.551	0.028
Total_capital	0.0250	0.139	0.180	0.857	-0.247	0.297
Reserves_and_funds	0.1597	0.131	1.223	0.221	-0.096	0.415
Borrowings	0.1429	0.106	1.354	0.176	-0.064	0.350
Current_liabilities_and_provisions	-0.0775	0.137	-0.567	0.570	-0.345	0.190
TOL_to_TNW	0.2270	0.132	1.714	0.087	-0.033	0.487
Total_term_liabilities__to__tangible_net_worth	-0.0917	0.133	-0.687	0.492	-0.353	0.170
Contingent_liabilities__to__Net_worth_perc	0.0580	0.082	0.705	0.481	-0.103	0.219
Net_working_capital	-0.0585	0.097	-0.606	0.544	-0.248	0.131
Current_ratio_times	0.1120	0.113	0.991	0.322	-0.110	0.334
Cash_to_current_liabilities_times	-0.0287	0.159	-0.181	0.857	-0.340	0.283
Cash_to_average_cost_of_sales_per_day	0.1738	0.129	1.351	0.177	-0.078	0.426
Creditors_turnover	0.0170	0.093	0.183	0.855	-0.165	0.199
Debtors_turnover	0.0083	0.090	0.093	0.926	-0.167	0.184
Finished_goods_turnover	-0.0976	0.120	-0.815	0.415	-0.332	0.137
WIP_turnover	0.0861	0.115	0.749	0.454	-0.139	0.311
Raw_material_turnover	-0.0311	0.096	-0.324	0.746	-0.219	0.157
Shares_outstanding	0.3683	0.186	1.980	0.048	0.004	0.733
Adjusted_EPS	0.0661	0.080	0.826	0.409	-0.091	0.223

Figure 50

A variable with a p-value greater than the significance level or alpha indicates a non-zero correlation between the independent and dependent variables at the population level. The variable is not statistically significant, and including it in the model may reduce precision.

Dropping the variables one by one until all the variables have p-values less than 0.05.

After repeating the same process for 20 times, the final formula obtained is

F=Default~Income_from_fincial_services+Reserves_and_funds+

TOL_to_TNW+Cash_to_average_cost_of_sales_per_day+

Shares_outstanding

Final model Summary:

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	1003
Model:	Logit	**Df Residuals:**	997
Method:	MLE	**Df Model:**	5
Date:	Fri, 26 Jul 2024	**Pseudo R-squ.:**	0.03789
Time:	21:09:09	**Log-Likelihood:**	-465.60
converged:	True	**LL-Null:**	-483.93
Covariance Type:	nonrobust	**LLR p-value:**	6.974e-07

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.5195	0.084	-18.006	0.000	-1.685	-1.354
Income_from_fincial_services	-0.2447	0.132	-1.849	0.064	-0.504	0.015
Reserves_and_funds	0.2200	0.088	2.487	0.013	0.047	0.393
TOL_to_TNW	0.1603	0.071	2.269	0.023	0.022	0.299
Cash_to_average_cost_of_sales_per_day	0.1858	0.069	2.691	0.007	0.051	0.321
Shares_outstanding	0.4249	0.127	3.333	0.001	0.175	0.675

Figure 51

Classification report for Train data:

	precision	recall	f1-score	support
0	0.82	1.00	0.90	815
1	0.64	0.04	0.07	188
accuracy			0.82	1003
macro avg	0.73	0.52	0.48	1003
weighted avg	0.78	0.82	0.74	1003

Figure 52

Classification report for Test data:

	precision	recall	f1-score	support
0	0.82	0.94	0.88	141
1	0.10	0.03	0.05	29
accuracy			0.78	170
macro avg	0.46	0.49	0.46	170
weighted avg	0.70	0.78	0.74	170

Figure 53

The model performance on predicting the default case is not so good due to class imbalance.

From the roc curve, the optimal threshold is 0.17.

Final Model

Classification report for Train data:

	precision	recall	f1-score	support
0	0.86	0.70	0.77	815
1	0.28	0.51	0.36	188
accuracy			0.66	1003
macro avg	0.57	0.60	0.56	1003
weighted avg	0.75	0.66	0.69	1003

Figure 54

Classification report for Test data:

	precision	recall	f1-score	support
0	0.83	0.66	0.74	141
1	0.17	0.34	0.23	29
accuracy			0.61	170
macro avg	0.50	0.50	0.48	170
weighted avg	0.72	0.61	0.65	170

Figure 55

The recall for class 1 on test data has increased from 0.03 to 30,which is good but not sufficient enough.

Confusion Matrix for Train data:

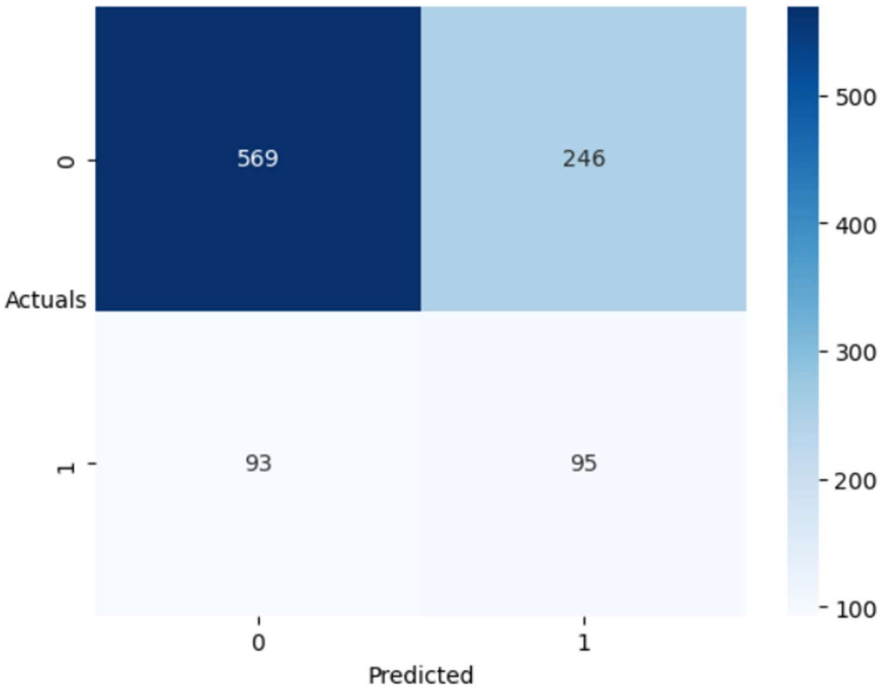

Figure 56

Confusion Matrix for Test data:

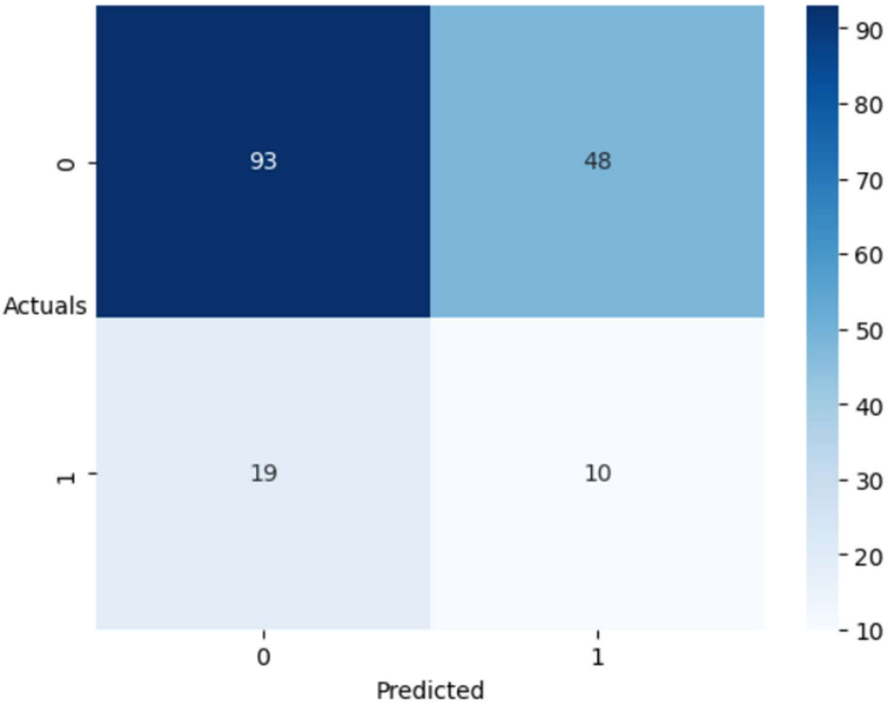

Figure 57

Hyperparameter Tuning for Random Forest:

We can find the optimal parameters for a random forest model by passing the required parameter choice through Grid Search Cv.

Parameter passed:

- 'max_depth': [5, 10, 15],
- 'min_samples_split': [15, 20, 30],
- 'n_estimators':[100,150,200],
- 'criterion': ['gini','entropy']

Building the final model using the below parameters:

- class_weight={0: 0.1, 1: 5}
- max_depth=10
- max_features=15
- min_samples_leaf=15
- min_samples_split=20
- n_estimators=400

Classification report on Train data:

	precision	recall	f1-score	support
0	1.00	0.06	0.12	1463
1	0.21	1.00	0.34	356
accuracy			0.25	1819
macro avg	0.60	0.53	0.23	1819
weighted avg	0.84	0.25	0.16	1819

Figure 58

Classification report for Test data:

	precision	recall	f1-score	support
0	0.75	0.09	0.16	628
1	0.19	0.88	0.31	152
accuracy			0.24	780
macro avg	0.47	0.48	0.24	780
weighted avg	0.64	0.24	0.19	780

Figure 59

Confusion Matrix for Train data:

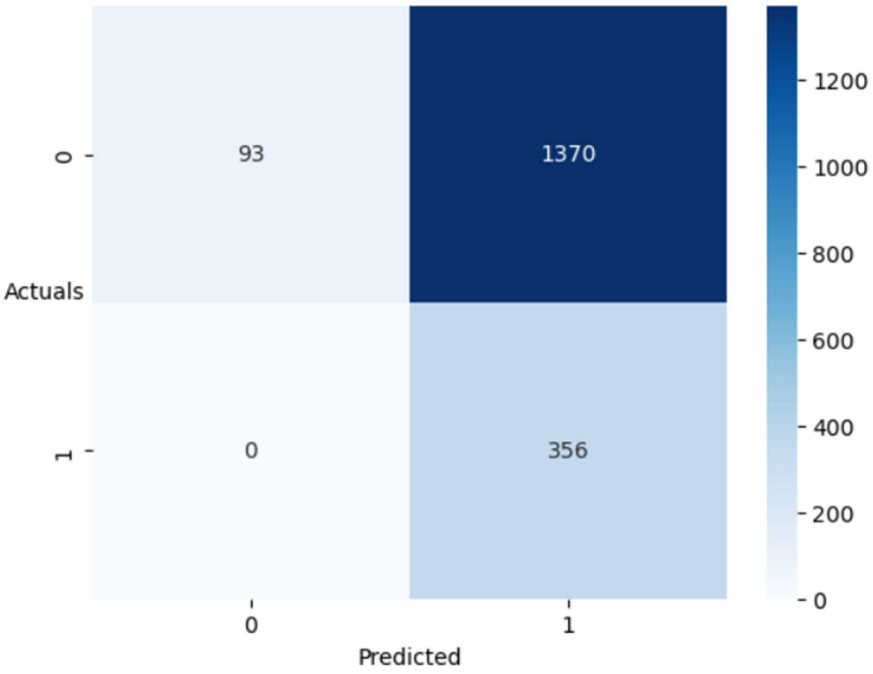

Figure 60

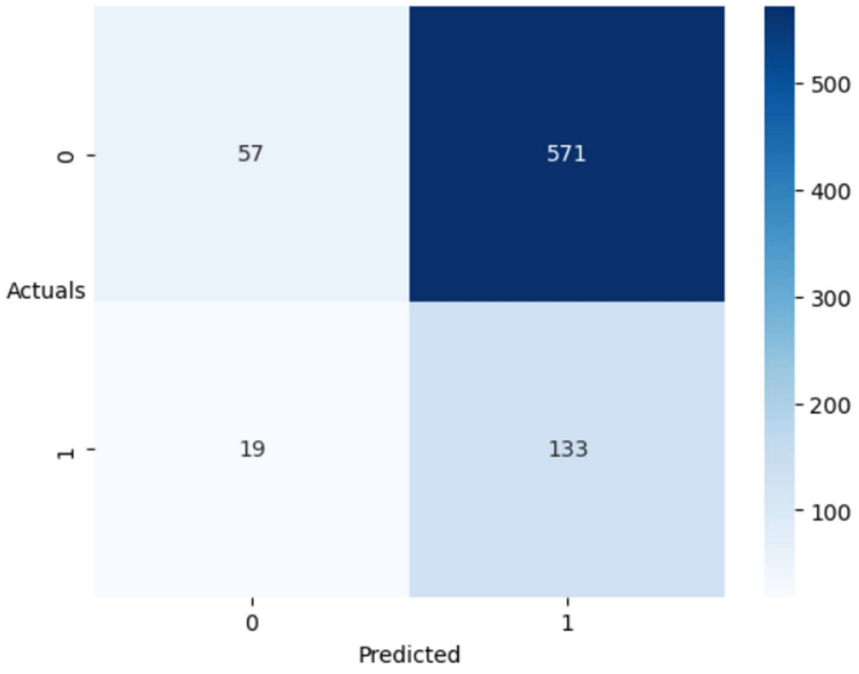

Figure 61

Model Performance Comparison and Final Model Selection:

Train Data

Model	Auc Score	Precision for class 1	Recall for class 1	Accuracy
Logistic Regression using Sklearn	0.55	0.25	0.38	0.66
Logistic Regression Stats model	0.6	0.28	0.51	0.66
Random forest	0.53	0.21	1	0.25

Table 7

Test Data

Model	Auc Score	Precision for class 1	Recall for class 1	Accuracy
Logistic Regression using Sklearn	0.52	0.21	0.66	0.44
Logistic Regression Stats model	0.5	0.17	0.34	0.61
Random forest	0.48	0.19	0.88	0.24

- The primary objective of this project is to reduce risk for financial institutions by identifying companies with a high likelihood of defaults.
- Since our goal is to accurately predict the true negatives (i.e., default cases), recall will be the most critical metric.
- Considering this the random forest model has the highest recall for both train and test data.
- Further random forest model can handle multi-collinearity, and outliers better as compared to the logistic regression model.
- Therefore we are proceeding with the random forest as our final model.
- Choosing accuracy and AUC score is not recommended due to class imbalance.

Feature Importance:

	Imp
Cash_to_average_cost_of_sales_per_day	0.05
Borrowings	0.05
Net_fixed_assets	0.05
Debtors_turnover	0.04
Quick_ratio_times	0.04
Contingent_liabilities__to__Net_worth_perc	0.03
Finished_goods_turnover	0.03
Creditors_turnover	0.03
Raw_material_turnover	0.03
Current_ratio_times	0.03
Total_capital	0.03
Shares_outstanding	0.03
TOL_to_TNW	0.03
Change_in_stock	0.03
Net_working_capital	0.02
PBDITA_as_perc_of_total_income	0.02
Sales	0.02
WIP_turnover	0.02
Cash_to_current_liabilities_times	0.02
Total_expenses	0.02
Current_liabilities_and_provisions	0.02
Debt_to_equity_ratio_times	0.02
Total_term_liabilities__to__tangible_net_worth	0.02
PBT_as_perc_of_total_income	0.02
Income_from_fincial_services	0.02
Cumulative_retained_profits	0.02
Cash_profit	0.02
Cash_profit_as_perc_of_total_income	0.02
EPS	0.02
Profit_after_tax	0.02
Adjusted_EPS	0.02
Total_income	0.02
Capital_employed	0.02
Reserves_and_funds	0.01
PAT_as_perc_of_net_worth	0.01
PAT_as_perc_of_total_income	0.01
Shareholders_funds	0.01
Current_assets	0.01
PBDITA	0.01
Total_liabilities	0.01
PBT	0.01
Net_worth	0.01
Total_assets	0.01

Figure 62

Inference:

Cash to average cost of sales per day, borrowings and the net fixed assets are the top 3 columns in differentiating the two classes.

Random Forest after smote:

Classification report for Train data:

	precision	recall	f1-score	support
0	0.99	0.14	0.25	1473
1	0.54	1.00	0.70	1454
accuracy			0.57	2927
macro avg	0.76	0.57	0.47	2927
weighted avg	0.76	0.57	0.47	2927

Figure 63

	precision	recall	f1-score	support
0	0.89	0.11	0.19	618
1	0.53	0.99	0.69	637
accuracy			0.55	1255
macro avg	0.71	0.55	0.44	1255
weighted avg	0.71	0.55	0.44	1255

Figure 64

Actionable Insights & Recommendations:

- It is evident that those who tend to default have significantly higher average net working capital.
- Higher net working capital would mean the company has comparatively higher liabilities.
- The defaulters have a 3 times higher D/E ratio as compared to the non-defaulters.
- On average, the defaulters tend to borrow more money as compared to the non-defaulters.
- High reserves indicate that the company is financially stable and can help companies sustain financial uncertainties.
- Prioritize assessing the D/E ratio of potential borrowers. A high D/E ratio indicates higher financial risk. Set a threshold for acceptable D/E ratios to filter out high-risk applicants.
- Implement a risk-based pricing model where interest rates are adjusted based on the risk profile of the companies.
- Set borrowing limits for those with high existing liabilities.
- Set up alerts when key financial metrics of a borrower reach concerning levels.
- Introduce interest incentives for those who repay the loan.
- Give preference to companies with high reserves and funds.

Problem 2

Define the problem statement:

Problem Statement:

Market Risk:

It is the risk associated with uncertainty in the market such as the price change risk in a stock market. Investors face market risk, arising from asset price fluctuations due to economic events, geopolitical developments, and investor sentiment changes. Understanding and analysing this risk is crucial for informed decision-making and optimizing investment strategies.

The objective of this analysis is to conduct a Market Risk Analysis on a portfolio of Indian stocks using Python.

Data Description:

- The dataset consists of the historical price of 5 Indian stocks in the market over an 8-year time period.
- The dataset enables us to analyse the historical performance of individual stocks and the overall market dynamics.

Data Dictionary:

```
---  ------         --------------  -----
 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)
```

Figure 65

- The dataset consists of 6 columns. Out of these, 5 is of the int data type, and the date column is of the datatype object.
- There are 418 rows and 6 columns.

Statistical Summary:

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

Figure 66

- There are no null values present in the dataset.
- Bharti Airtel has the highest average price followed by Tata Motors.
- Yes bank has the lowest mean price.
- Even though Airtel has the highest mean, it also has the widest spread in terms of standard deviation, which is around 227.

Stock Price Graph Analysis:

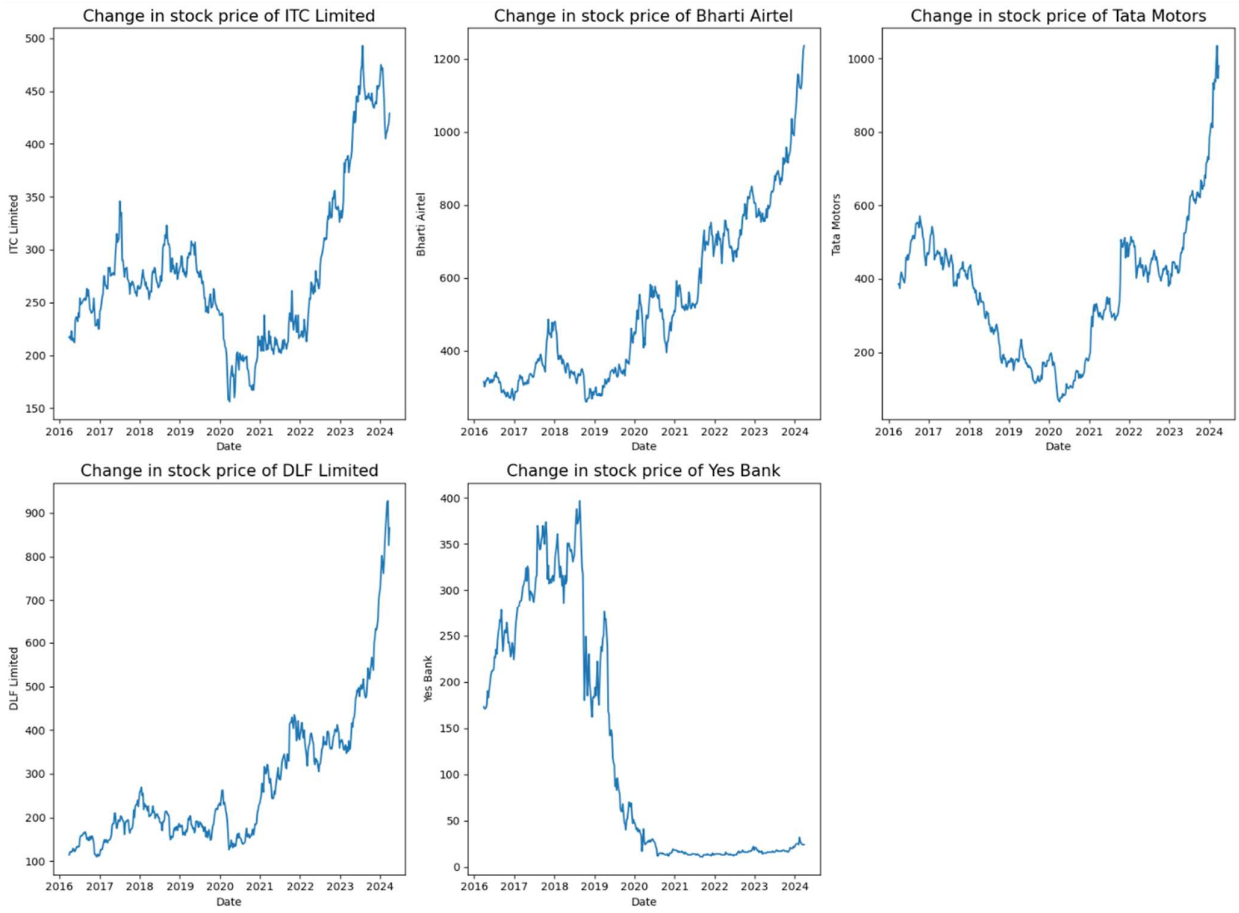

Figure 67

- From the plot above, it is evident that both Bharti Airtel and DLF Limited have experienced a steady increase in price over the years.
- Both ITC Limited and Tata Motors have also seen a considerable rise, but their share prices plunged during 2020.
- This could be due to the fact that the entire nation was quarantined during that time because of COVID-19.
- Yes Bank shares reached their highest point in 2018, and since then, the share price has been declining.

Stock Returns Calculation and Analysis

Logarithmic Returns:

- The dataset consists of the weekly price changes of the stocks. Logarithmic return is the log of difference between the two consecutive week's prices.
- The average return can be calculated by taking the mean of the output
- Volatility can be measured by calculating the standard deviation of the returns.

The first few rows

	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

Figure 68

Average Returns:

Stocks	Average Returns
ITC Limited	0.001634
Bharti Airtel	0.003271
Tata Motors	0.002234
DLF Limited	0.004863
Yes Bank	-0.004737

Table 8

Volatility:

Stocks	Volatility
ITC Limited	0.035904
Bharti Airtel	0.038728
Tata Motors	0.060484
DLF Limited	0.057785
Yes Bank	0.093879

Table 9

Average returns Vs volatility:

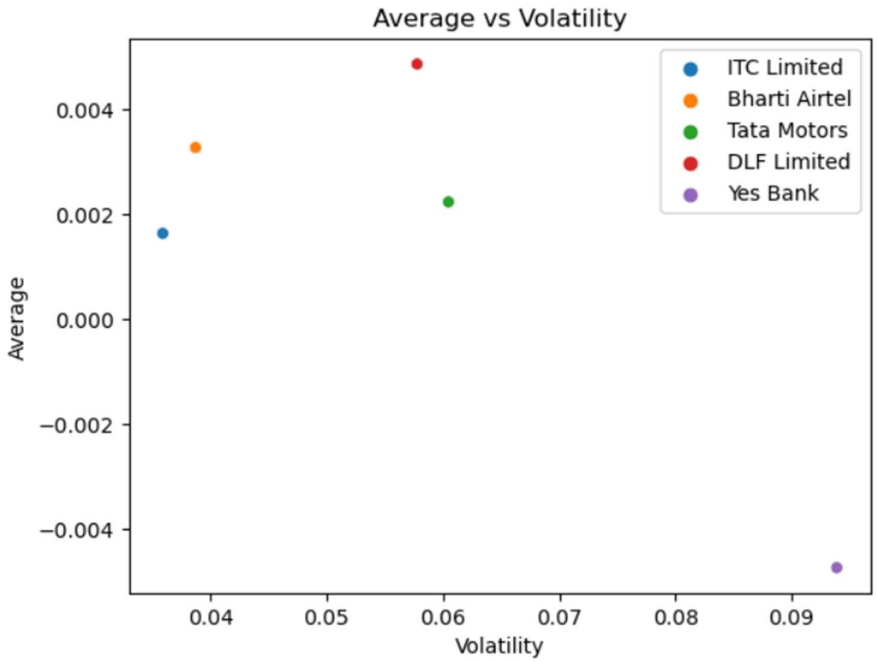

Figure 69

For comparison purposes, let us take the mean of the average returns and mean volatility of all stocks.

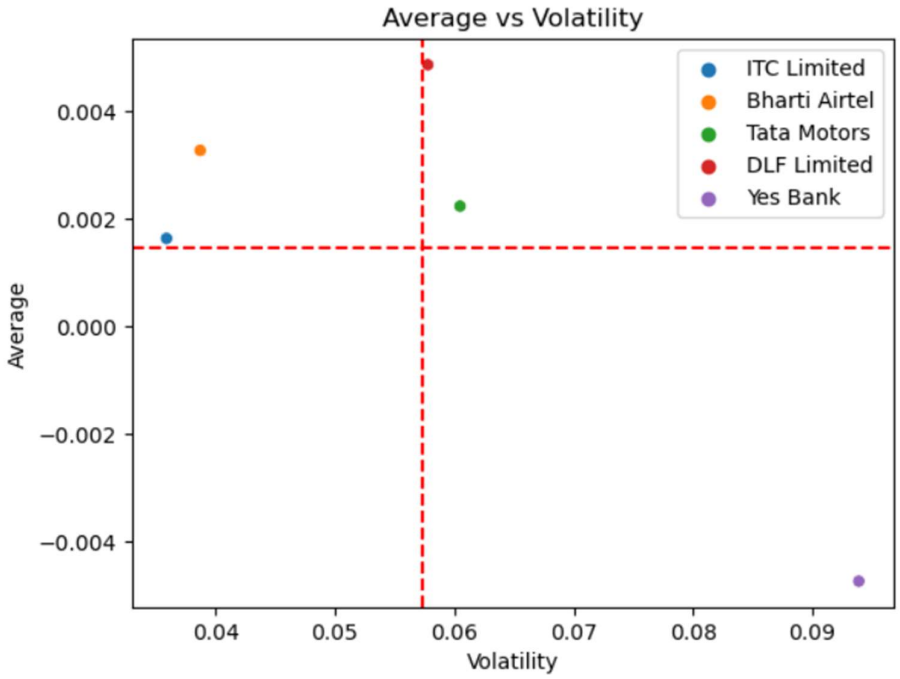

Figure 70

- We can see that, yes bank has the lowest average returns and the highest volatility.
- Both ITC Limited and Bharti Airtel have delivered impressive returns while maintaining lower volatility.
- DLF Limited has the highest average return compared to others, but its volatility is slightly above the group average.
- Tata Motors has the third highest returns overall, accompanied by somewhat elevated volatility.

Actionable Insights & Recommendations:

- Invest across different sectors to spread risk. Avoid over-concentration in any single industry.
- Since ITC Limited and Bharti Airtel have delivered impressive returns with lower volatility, try to invest a larger proportion in those shares.
- Regularly assess the volatility of your holdings and rebalance the portfolio to reduce when necessary.
- Ensure that your investment strategy aligns with your risk tolerance and financial goals.
- If you have a higher risk tolerance, try investing in DLF, which has delivered the highest returns with a slightly higher volatility. Start your investments with a smaller capital and increase it systematically.
- Distribute 70% of your capital in stable stocks such as ITC & Airtel and the remaining 25% in DLF & Tata Motors.
- Avoid investing in Yes Bank as it has the lowest returns with the highest volatility.