Financial & Risk Analytics – Credit risk – Default Companies

Part A- Coded Project

Business Report

DSBA – Course

Created by – Rishabh Gupta

Foreword

Business Context:

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

Objective:

A group of venture capitalists want to develop a Financial Health Assessment Tool. With the help of the tool, it endeavours to empower businesses and investors with a robust mechanism for evaluating the financial well-being and creditworthiness of companies. By harnessing machine learning techniques, they aim to 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 fulfil financial obligations promptly and efficiently, and identify potential cases of default.
2. Credit Risk Evaluation: Evaluate credit risk exposure by analysing liquidity ratios, debt-to-equity ratios, and other key financial indicators to ascertain the likelihood of default and inform investment decisions.

They have hired you as a data scientist and provided you with the financial metrics of different companies. The task is to 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.

Contents

List of Tables

List of Figures

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.

Data Analysis of problem statement –

The dataset contains data corresponding to polished and unpolished stones.

Sheet name –

1. *Comp_Fin_Data.csv*

Part A –

Data Dictionary –

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

- Networth Next Year: Net worth of the customer in the next year
- Total assets: Total assets of customer
- Net worth: Net worth of the customer of the present year
- Total income: Total income of the customer
- Change in stock: Difference between the current value of the stock and the value of stock in the last trading day
- Total expenses: Total expenses done by the customer
- Profit after tax: Profit after tax deduction
- PBDITA: Profit before depreciation, income tax, and amortization
- PBT: Profit before tax deduction
- Cash profit: Total Cash profit
- PBDITA as % of total income: PBDITA / Total income
- PBT as % of total income: PBT / Total income
- PAT as % of total income: PAT / Total income
- Cash profit as % of total income: Cash Profit / Total income
- PAT as % of net worth: PAT / Net worth
- Sales: Sales done by the customer
- Income from financial services: Income from financial services
- Other income: Income from other sources
- Total capital: Total capital of the customer
- Reserves and funds: Total reserves and funds of the customer
- Borrowings: Total amount borrowed by the customer
- Current liabilities & provisions: current liabilities of the customer
- Deferred tax liability: Future income tax customer will pay because of the current transaction
- Shareholders funds: Amount of equity in a company which belongs to shareholders

- Cumulative retained profits: Total cumulative profit retained by customer
- Capital employed: Current asset minus current liabilities
- TOL/TNW: Total liabilities of the customer divided by Total net worth
- Total term liabilities / tangible net worth: Short + long term liabilities divided by tangible net worth
- Contingent liabilities / Net worth (%): Contingent liabilities / Net worth
- Contingent liabilities: Liabilities because of uncertain events
- Net fixed assets: The purchase price of all fixed assets
- Investments: Total invested amount
- Current assets: Assets that are expected to be converted to cash within a year
- Net working capital: Difference between the current liabilities and current assets
- Quick ratio (times): Total cash divided by current liabilities
- Current ratio (times): Current assets divided by current liabilities
- Debt to equity ratio (times): Total liabilities divided by its shareholder equity
- Cash to current liabilities (times): Total liquid cash divided by current liabilities
- Cash to average cost of sales per day: Total cash divided by the average cost of the sales
- Creditors turnover: Net credit purchase divided by average trade creditors
- Debtors turnover: Net credit sales divided by average accounts receivable
- Finished goods turnover: Annual sales divided by average inventory
- WIP turnover: The cost of goods sold for a period divided by the average inventory for that period
- Raw material turnover: Cost of goods sold is divided by the average inventory for the same period
- Shares outstanding: Number of issued shares minus the number of shares held in the company
- Equity face value: cost of the equity at the time of issuing
- EPS: Net income divided by the total number of outstanding share
- Adjusted EPS: Adjusted net earnings divided by the weighted average number of common shares outstanding on a diluted basis during the plan year
- Total liabilities: Sum of all types of liabilities
- PE on BSE: Company's current stock price divided by its earnings per share

Data Overview –

To start the data analysis, we need to import the necessary libraries, specify the working directory, and load the dataset. Then, we will view the first five rows using head () function to get a preliminary understanding of the data. The Dataset has 4256 number of rows with 51 columns.

	Num	Networth_Next_Year	Total_assets	Net_worth	Total_income	Change_in_stock	Total_expenses	Profit_after_tax	PBDITA	PBT	Cash_profit	PBDITA_as_perc_of_t
0	1	395.3	827.6	336.5	534.1	13.5	508.7	38.9	124.4	64.6	95.2	
1	2	36.2	67.7	24.3	137.9	-3.7	131.0	3.2	5.5	1.0	3.8	
2	3	84.0	238.4	78.9	331.2	-18.1	309.2	3.9	25.8	10.5	9.4	
3	4	2041.4	6883.5	1443.3	8448.5	212.2	8482.4	178.3	418.4	185.1	178.0	
4	5	41.8	90.9	47.0	388.6	3.4	392.7	-0.7	7.2	-0.6	3.9	

TABLE 1 - TOP 5 ROWS OF DATASET

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4256 entries, 0 to 4255
Data columns (total 51 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Num                                             4256 non-null   int64
 1   Networth_Next_Year                              4256 non-null   float64
 2   Total_assets                                    4256 non-null   float64
 3   Net_worth                                       4256 non-null   float64
 4   Total_income                                    4025 non-null   float64
 5   Change_in_stock                                 3706 non-null   float64
 6   Total_expenses                                  4091 non-null   float64
 7   Profit_after_tax                                4102 non-null   float64
 8   PBDITA                                          4102 non-null   float64
 9   PBT                                             4102 non-null   float64
 10  Cash_profit                                     4102 non-null   float64
 11  PBDITA_as_perc_of_total_income                  4177 non-null   float64
 12  PBT_as_perc_of_total_income                     4177 non-null   float64
 13  PAT_as_perc_of_total_income                     4177 non-null   float64
 14  Cash_profit_as_perc_of_total_income             4177 non-null   float64
 15  PAT_as_perc_of_net_worth                        4256 non-null   float64
 16  Sales                                           3951 non-null   float64
 17  Income_from_fincial_services                    3145 non-null   float64
 18  Other_income                                    2700 non-null   float64
 19  Total_capital                                   4251 non-null   float64
 20  Reserves_and_funds                              4158 non-null   float64
 21  Borrowings                                      3825 non-null   float64
 22  Current_liabilities_&_provisions                4146 non-null   float64
 23  Deferred_tax_liability                          2887 non-null   float64
 24  Shareholders_funds                              4256 non-null   float64
 25  Cumulative_retained_profits                     4211 non-null   float64
 26  Capital_employed                                4256 non-null   float64
 27  TOL_to_TNW                                      4256 non-null   float64
 28  Total_term_liabilities__to__tangible_net_worth  4256 non-null   float64
 29  Contingent_liabilities__to__Net_worth_perc      4256 non-null   float64
 30  Contingent_liabilities                          2854 non-null   float64
 31  Net_fixed_assets                                4124 non-null   float64
 32  Investments                                     2541 non-null   float64
 33  Current_assets                                  4176 non-null   float64
 34  Net_working_capital                             4219 non-null   float64
 35  Quick_ratio_times                               4151 non-null   float64
 36  Current_ratio_times                             4151 non-null   float64
 37  Debt_to_equity_ratio_times                      4256 non-null   float64
 38  Cash_to_current_liabilities_times               4151 non-null   float64
 39  Cash_to_average_cost_of_sales_per_day           4156 non-null   float64
 40  Creditors_turnover                              3865 non-null   float64
 41  Debtors_turnover                                3871 non-null   float64
 42  Finished_goods_turnover                         3382 non-null   float64
 43  WIP_turnover                                    3492 non-null   float64
 44  Raw_material_turnover                           3828 non-null   float64
 45  Shares_outstanding                              3446 non-null   float64
 46  Equity_face_value                               3446 non-null   float64
 47  EPS                                             4256 non-null   float64
 48  Adjusted_EPS                                    4256 non-null   float64
 49  Total_liabilities                               4256 non-null   float64
 50  PE_on_BSE                                       1629 non-null   float64
dtypes: float64(50), int64(1)
memory usage: 1.7 MB
```

TABLE 2 - BASIC INFO. OF DATASET

Regarding the datatype of the columns in the dataset, we can confirm that-

• We can observe all variables are numerical.

Statistical Summary –

Using Describe () function, we can analyses the summary statistics of the dataset –

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

TABLE 3 - STATISTICAL SUMMARY OF DATASET

Observations-

Key Indicators of Financial Distress: Several variables might be strong indicators of future default.

- Negative or very low "Net worth": Companies with low or negative net worth in the current year might be at higher risk. The statistics show a minimum negative net worth.

- High Leverage Ratios (e.g., TOL/TNW, Debt to equity ratio): High values here suggest the company is heavily reliant on debt, increasing financial risk. The maximum values for these ratios are quite high, indicating some highly leveraged entities.
- Low Liquidity Ratios (e.g., Quick ratio, Current ratio, Cash to current liabilities): Low values indicate a potential struggle to meet short-term obligations. The minimum values for these ratios are close to zero, suggesting some companies might have very limited liquid assets compared to their liabilities.
- Negative Profitability (e.g., Profit after tax, PBT, PAT as % of net worth): Consistent negative profitability erodes net worth over time. The minimum values for these profit metrics are negative.
- Low Turnover Ratios (e.g., Creditors turnover, Debtors turnover, Inventory turnovers): Inefficient management of working capital can lead to cash flow problems. Very low turnover figures (close to zero in some cases) might signal issues.

Exploratory Data Analysis

Univariate analysis-

Lets analyze histograms –

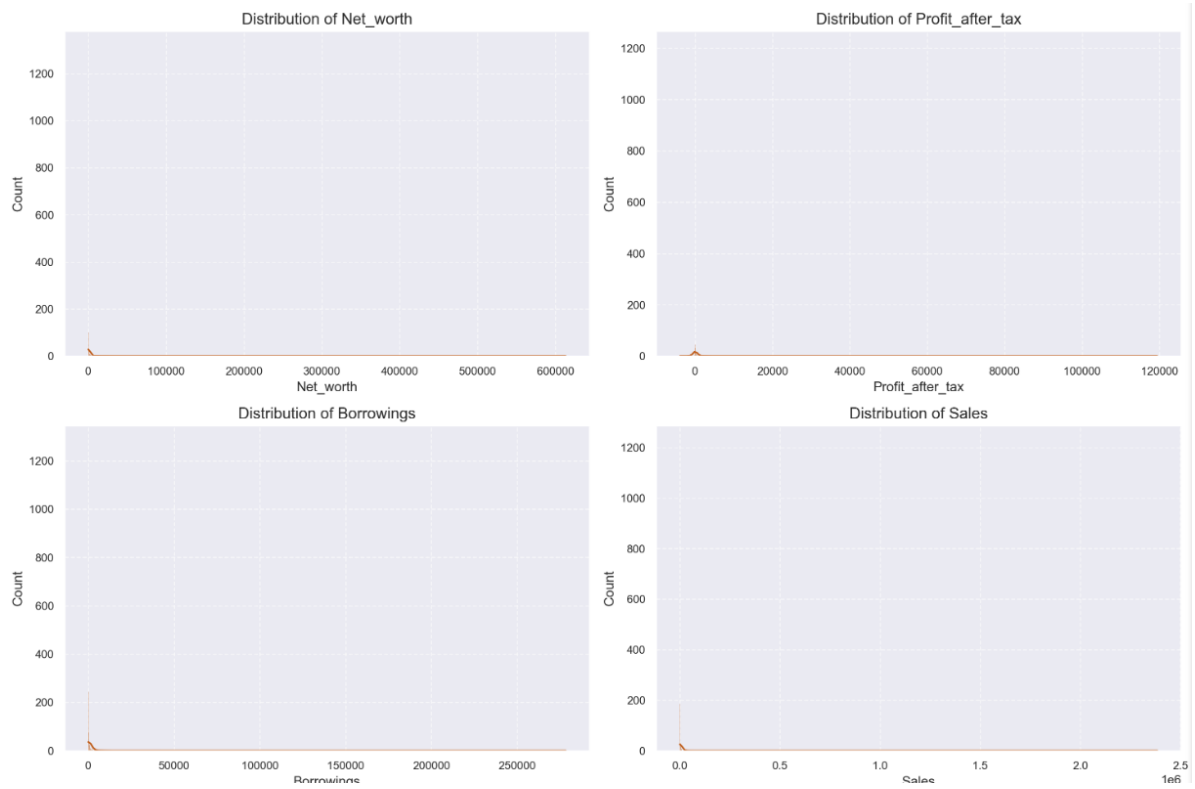

FIGURE 1 - HISTOGRAM PLOTS FOR VARIABLES

Distribution of Net_worth:

- Strongly right-skewed, with most companies having lower net worth.
- A significant number of entities have a net worth close to zero.
- A long tail indicates a few companies with substantially higher net worth.

Distribution of Profit_after_tax:

- Highly right-skewed, showing that most companies have lower profits after tax.

- A large concentration of companies have profits near zero or slightly positive.
- A few companies exhibit significantly higher profit after tax values.

Distribution of Borrowings:

- Right-skewed, suggesting that most companies have relatively lower borrowing amounts.
- A notable number of companies have borrowing levels close to zero.
- The distribution extends to higher borrowing amounts for a smaller subset of companies.

Distribution of Sales:

- Markedly right-skewed, indicating that most companies have lower sales figures.
- A high frequency of companies reports sales near the lower end of the spectrum.
- A long tail shows that a few companies achieve considerably higher sales volumes.

Also,

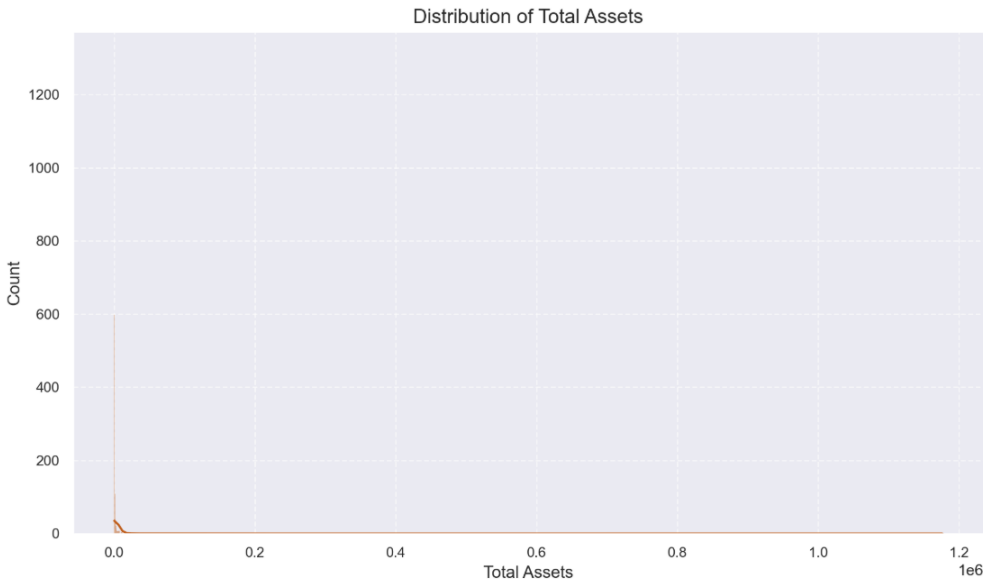

FIGURE 2 - HISTPLOT FOR TOTAL ASSETS

- Strongly right-skewed, indicating most entities have lower asset values.
- A high concentration of data points exists near zero.
- A long tail suggests the presence of a few companies with very high total assets.
- The distribution is clearly non-normal.
- Potential outliers or influential high-value data points are present.

Lets analyze boxplot for Total income-

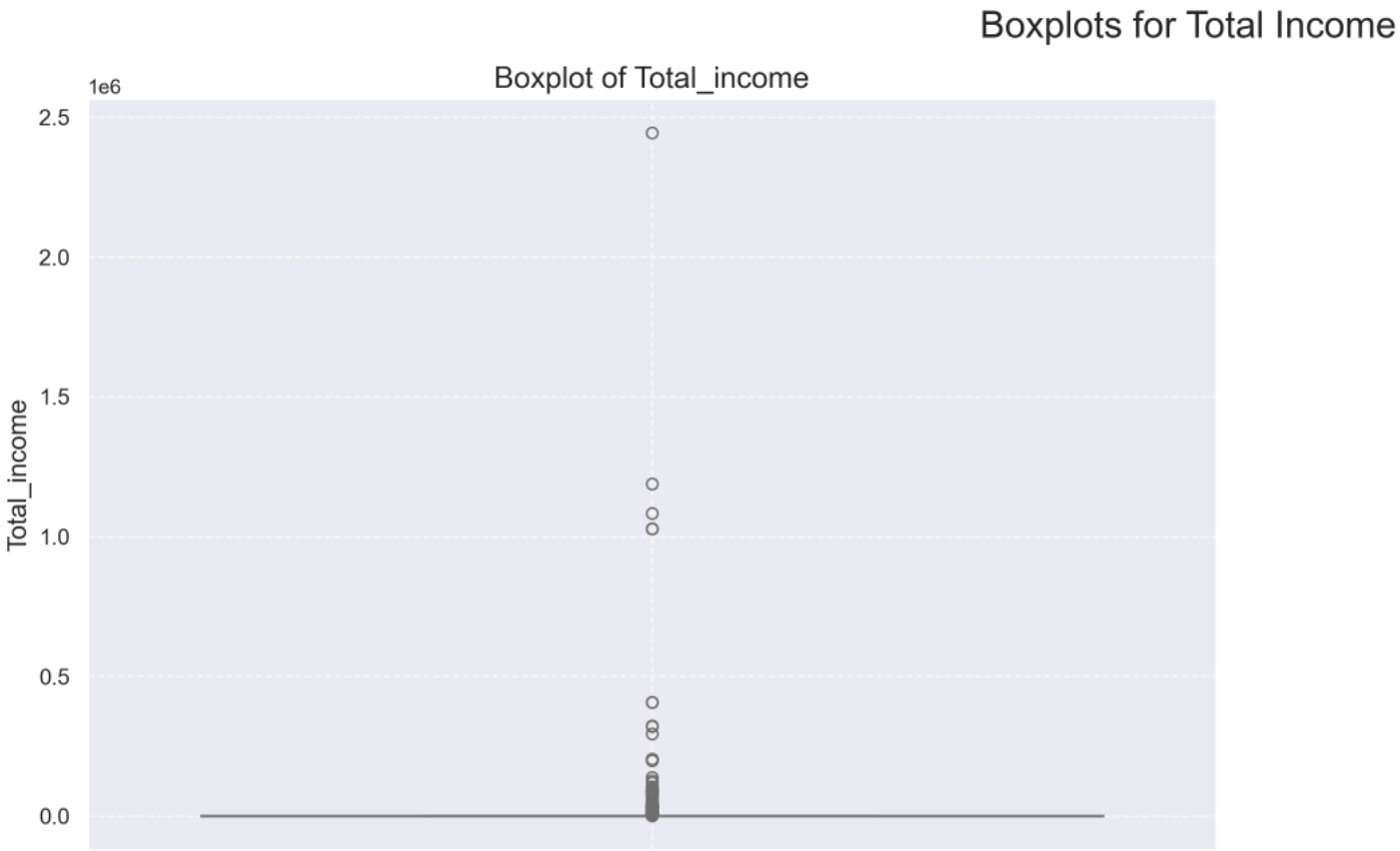

FIGURE 3 - BOXPLOT FOR TOTAL INCOME

- The majority of the data points are clustered near zero, indicated by the box's position.
- There is a significant positive skew, evident from the long upper whisker extending to much higher values.
- Several outliers are present at higher "Total_income" levels, far beyond the upper whisker.
- The median income (the line inside the box) is relatively low compared to the maximum values.
- This distribution suggests that while most companies have lower total income, a few have exceptionally high incomes.

Similarly, let do boxplots for other variables –

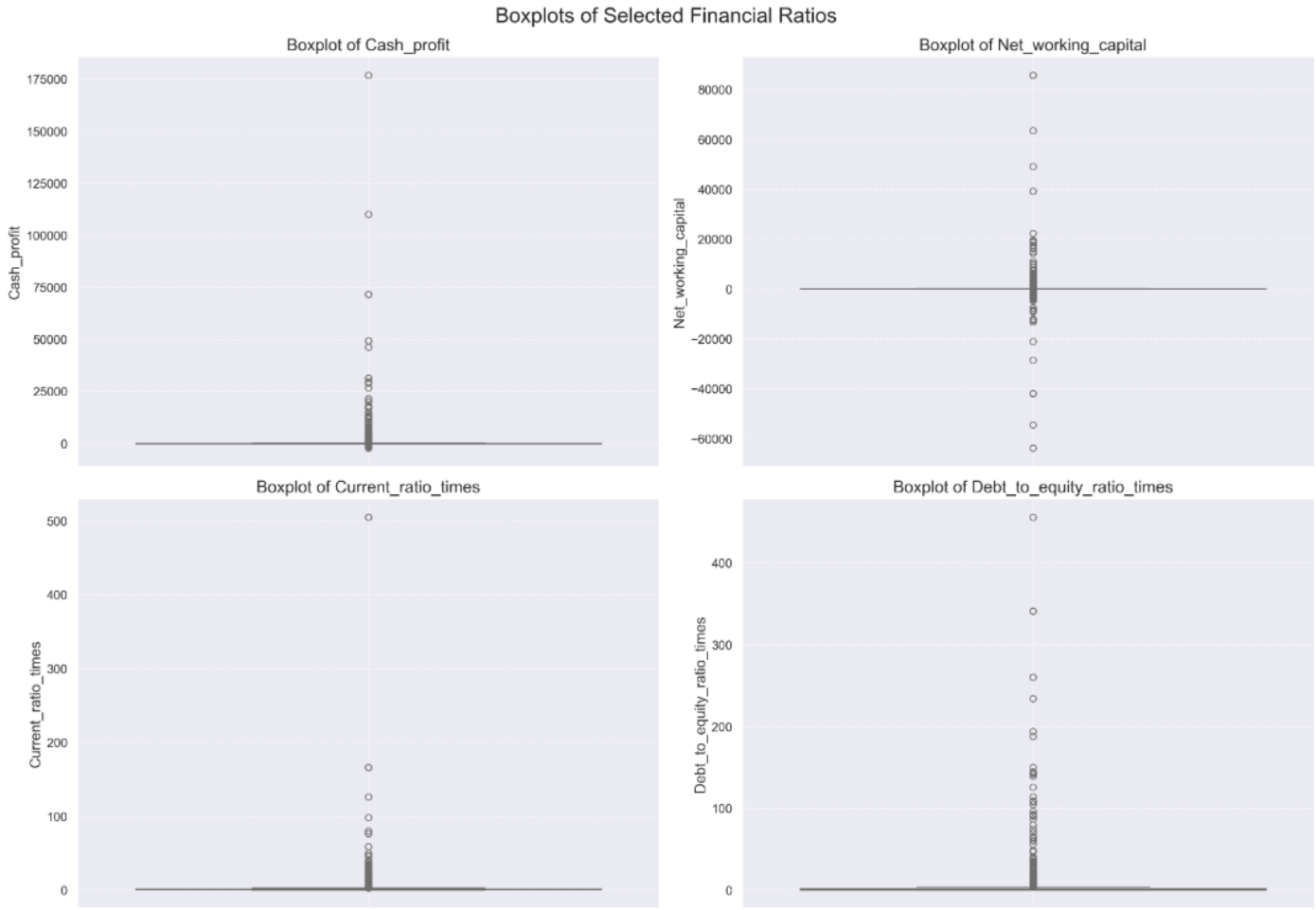

FIGURE 4 - OTHER BOXPLOTS FOR UNIVARIATE ANALYSIS

Boxplot of Cash_profit:

- Most cash profit values are clustered near zero, with a relatively small interquartile range.
- The upper whisker extends significantly, indicating a positive skew towards higher cash profits.
- Numerous outliers are present at high cash profit values, far exceeding the upper whisker.

Boxplot of Net_working_capital:

- The distribution of net working capital is centered around zero, with a relatively tight interquartile range.
- The whiskers extend in both positive and negative directions, suggesting variability.
- Several outliers are observed on both the positive and negative ends of the net working capital spectrum.

Boxplot of Current_ratio_times:

- The majority of current ratios are low, with the box concentrated towards smaller values.
- A long upper whisker and several outliers indicate some companies have significantly higher current ratios.
- One extreme outlier is visible with a very high current ratio.

Boxplot of Debt_to_equity_ratio_times:

- Most debt-to-equity ratios are low, with the box situated near the lower end.
- The upper whisker extends considerably, showing a positive skew towards higher leverage.
- Multiple outliers exist at higher debt-to-equity ratios, indicating companies with substantial debt relative to equity.

Here,

We consider 'Networth Next Year' as our Default Variable

SO, we call negative values as Default = 1

And, positive values as Default = 0

Lets see Bar Plot for Categorical/Target Variable -

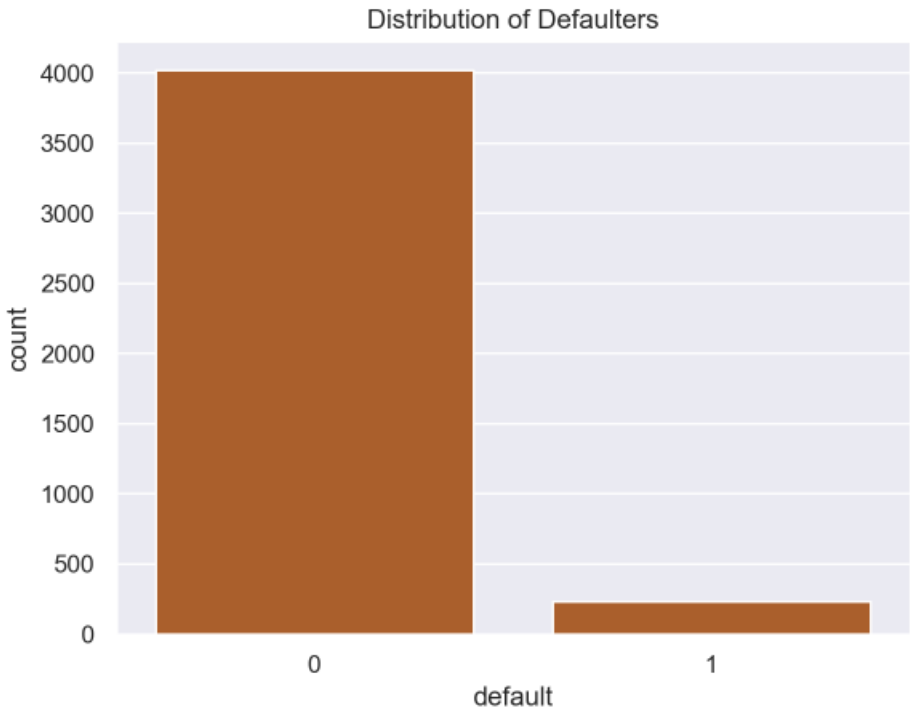

FIGURE 5 - BARPLOT FOR TARGET VARIABLE

- The plot is a bar chart showing the count of companies categorized as 'Default = 0' (positive Networth Next Year) and 'Default = 1' (negative Networth Next Year).
- The bar representing 'Default = 0' is significantly taller than the bar for 'Default = 1'.
- Approximately 4000 companies have a positive net worth in the next year (non-defaulters).
- Around 200 companies have a negative net worth in the next year (defaulters).
- The dataset exhibits a class imbalance, with a much larger number of non-defaulters compared to defaulters. This imbalance should be considered during model development and evaluation.

Multivariate Analysis

Lets plot Correlation Heatmap -

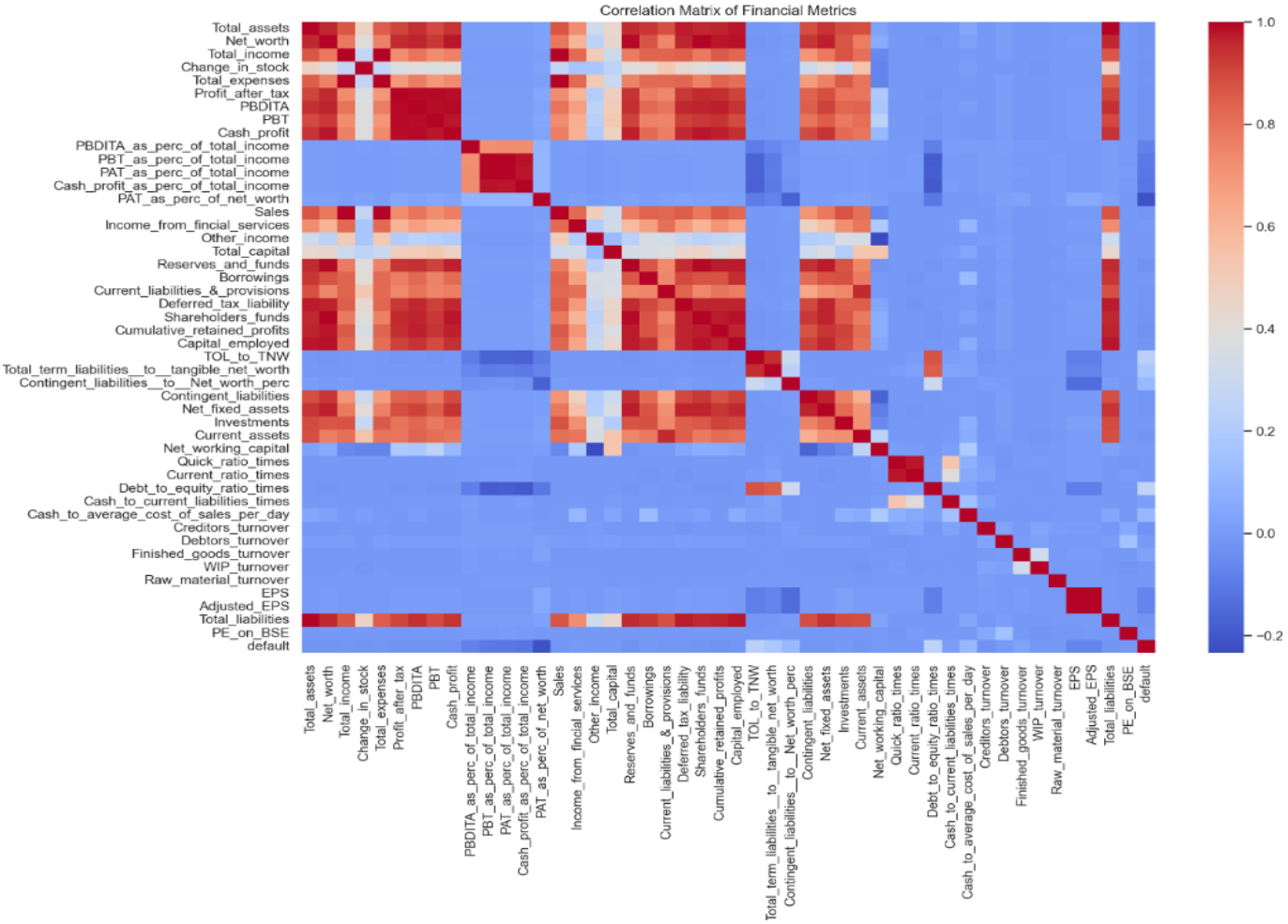

FIGURE 6 - CORRELATION HEATMAP

- The diagonal shows a perfect positive correlation (red) as each variable is perfectly correlated with itself.

- Strong positive correlations (red) exist between related metrics such as 'Total_assets' and 'Net worth', various profit measures (e.g., 'PBT', 'PAT', 'Cash profit'), and different income components.

- Strong negative correlations (blue) are observed between certain variables, for instance, some profitability ratios and leverage ratios like 'TOL/TNW' or 'Debt to equity ratio'.

- Several percentage-based ratios (e.g., '_as_perc_of_total_income') show moderate to strong correlations with their absolute counterparts (e.g., 'PBT', 'PAT').

- Liquidity ratios ('Current_ratio', 'Quick_ratio') tend to have positive correlations with each other but weaker correlations with profitability metrics.

- Turnover ratios generally show weak or mixed correlations with asset/liability metrics, suggesting different aspects of business performance.

- 'PE on BSE' appears to have relatively weak correlations with most balance sheet and income statement items, potentially reflecting market valuation factors beyond immediate financial metrics.

- The correlation matrix highlights potential multicollinearity issues that might need to be addressed during model building, especially among highly correlated features.

Now lets analyze Pairplot –

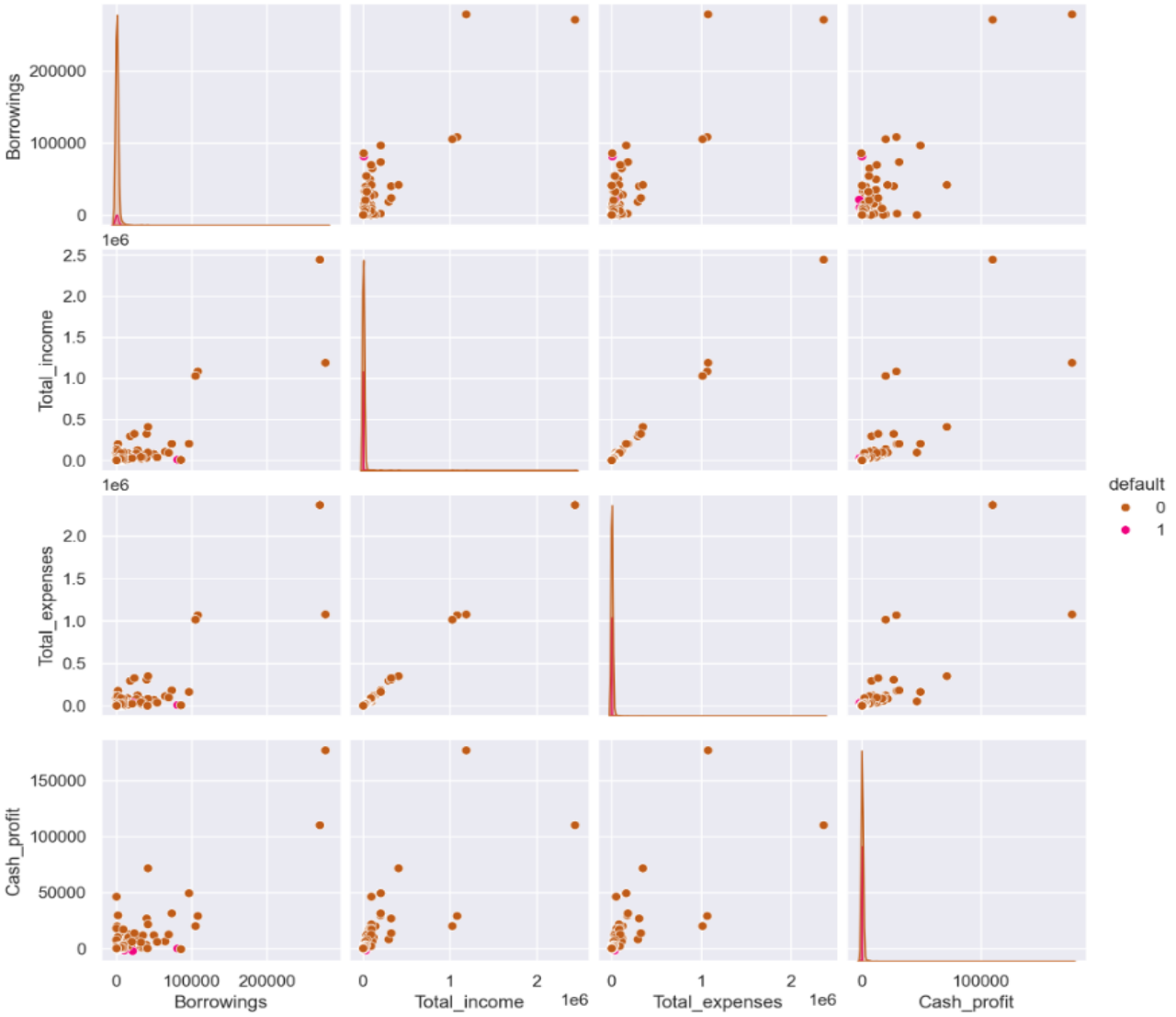

FIGURE 7 - PAIRPLOT ON KEY FEATURES AND TARGET

- The diagonal plots show the univariate distributions (histograms with KDE) for each of the four financial variables. All distributions appear right-skewed, indicating a concentration of lower values with a tail extending towards higher values.
- The off-diagonal plots are scatter plots showing the pairwise relationships between the variables, colored by the 'default' status (0 for non-default, 1 for default).

- **Borrowings vs. Total_income:** There's a general trend of higher borrowings being associated with higher total income, although the relationship isn't strictly linear. Default cases (magenta) appear mostly at lower to mid-range total income levels, with borrowings spanning a wider range.

- **Borrowings vs. Total_expenses:** Similar to total income, higher borrowings tend to occur with higher total expenses. Defaulters are concentrated in the lower ranges of total expenses, with varying levels of borrowings.
- **Borrowings vs. Cash_profit:** There's a weak positive correlation, suggesting higher borrowings might be linked to slightly higher cash profit. Defaulters are predominantly seen at lower cash profit levels, across different borrowing amounts.
- **Total_income vs. Total_expenses:** A strong positive correlation exists between total income and total expenses, as expected. Default cases are primarily clustered in the lower ranges of both income and expenses.
- **Total_income vs. Cash_profit:** A positive correlation is observed, indicating that higher total income generally leads to higher cash profit. Defaulters are mainly found at lower income and cash profit levels.
- **Total_expenses vs. Cash_profit:** A positive correlation is present, suggesting higher total expenses are associated with higher cash profit, although the spread is wider. Defaulters are concentrated in the lower ranges of both expenses and cash profit.

- The 'default' class (magenta) seems to be more prevalent in the lower ranges of 'Total_income', 'Total_expenses', and 'Cash_profit', suggesting these might be indicators of higher default risk. 'Borrowings' doesn't show as clear a separation for the default class across its entire range.

- There are instances of non-defaulters (brown) even at lower levels of income, expenses, and cash profit, indicating that other factors beyond these four variables likely contribute to default risk.

Boxplot by Defaulter Status-

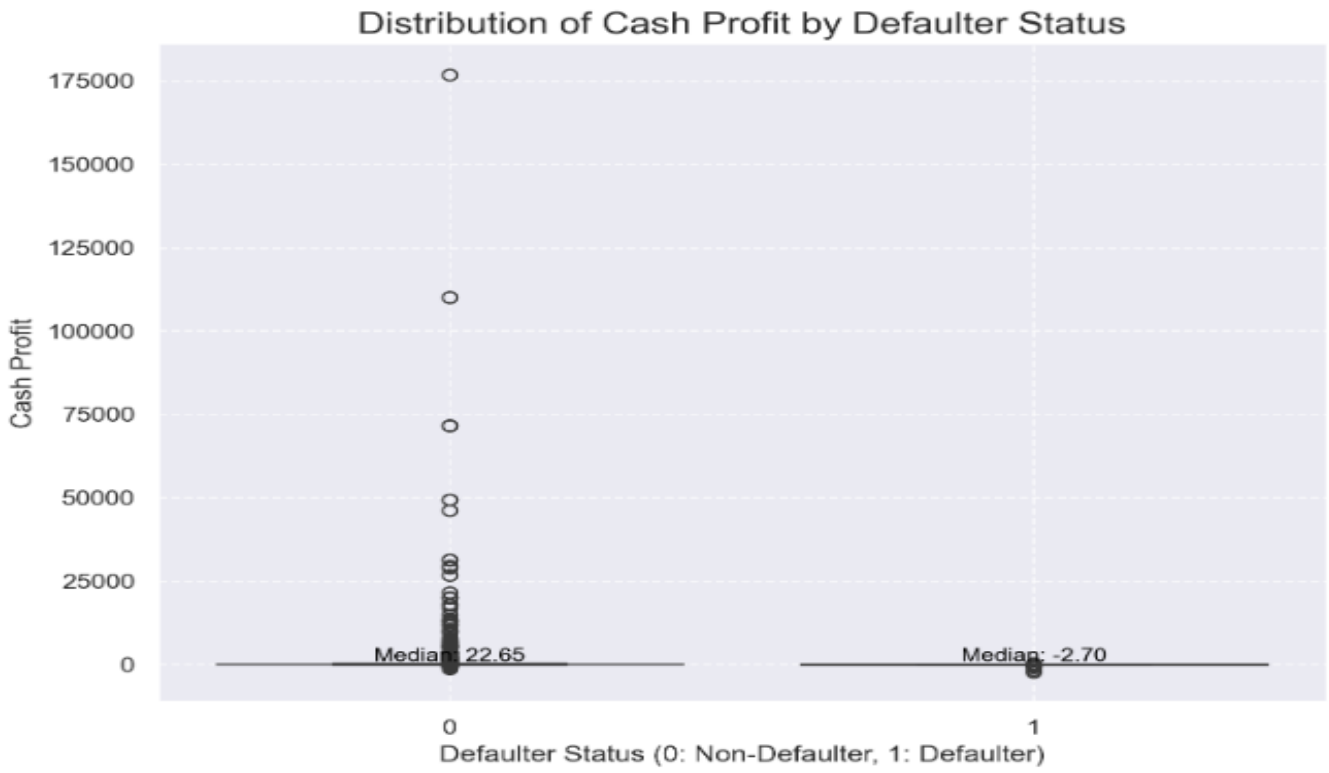

FIGURE 8 - BOXPLOT BY DEFAULTER STATUS

- Non-defaulters (0) generally exhibit a higher median cash profit (approximately 22.65) compared to defaulters (1), who have a median cash profit close to zero (-2.70).
- The distribution of cash profit for non-defaulters shows a wider spread, with a longer upper whisker and several high outliers, indicating greater variability and the presence of companies with significantly higher cash profits.
- Defaulters have a much more compressed distribution of cash profit, with most values clustered near zero and fewer, less extreme outliers on the higher end.
- Both groups show a positive skew in their cash profit distribution, as the upper whiskers are longer than the lower whiskers (though less pronounced for defaulters).
- The presence of outliers in both groups suggests that while defaulters tend to have lower median cash profit, some individual defaulters might still experience periods of higher cash profit. The concentration of defaulters at very low or negative cash profit levels is a key distinguishing factor.

Violin Plots for Distribution by Class –

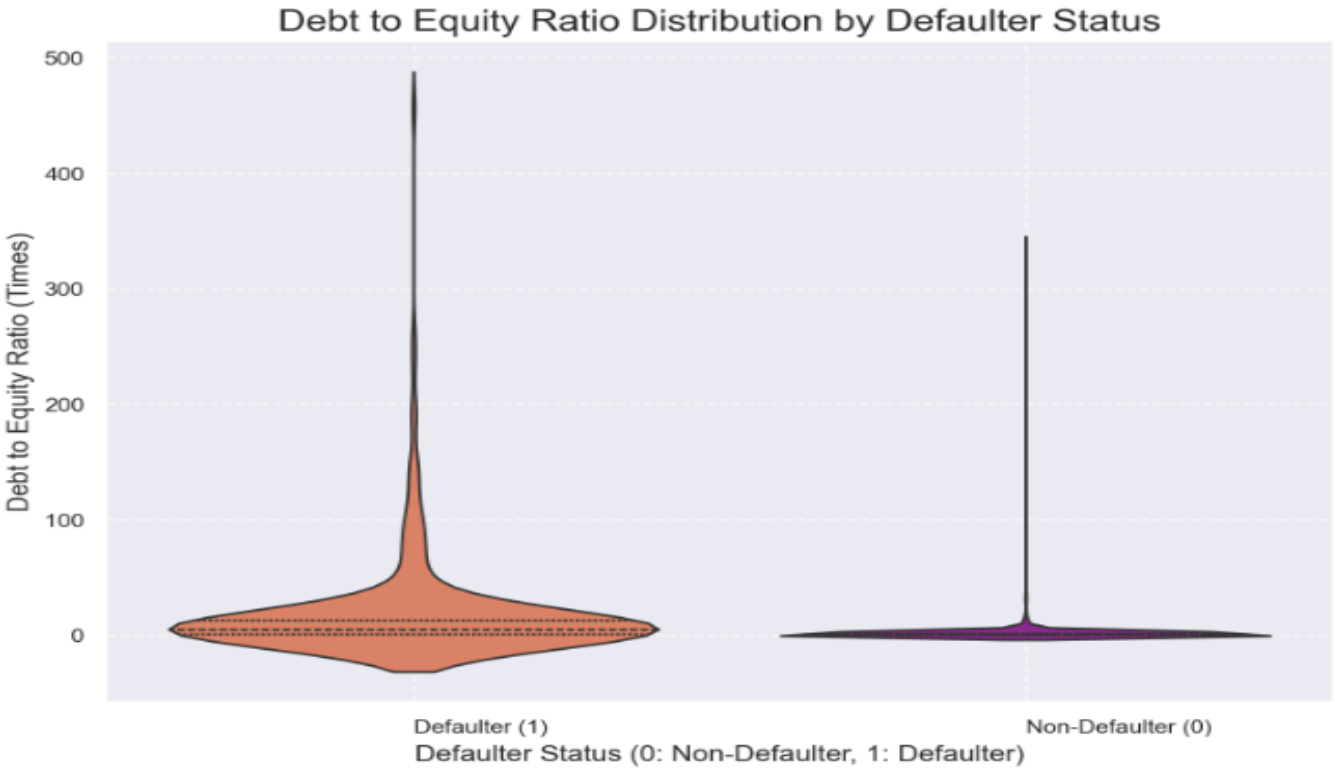

FIGURE 9 - VIOLIN PLOT FOR DISTRIBUTION BY CLASS

- The plot shows the distribution of the debt-to-equity ratio for defaulting (1) and non-defaulting (0) companies.

- The violin for defaulters (originally on the left, now effectively bottom due to x-axis inversion) is wider and extends to significantly higher debt-to-equity ratios, indicating greater variability and higher leverage.
- The violin for non-defaulters (originally on the right, now effectively top) is much narrower and concentrated closer to zero, suggesting lower and less variable debt-to-equity ratios.
- The quartile lines within the violins further emphasize this difference, with the median and upper quartiles being notably higher for defaulters.

- The shape of the defaulter violin suggests a higher probability of companies with very high debt-to-equity ratios being classified as defaulters.
- Non-defaulters predominantly exhibit low debt-to-equity ratios, implying that lower leverage might be associated with a lower risk of default.

Scatter Plot (Feature Relationships) -

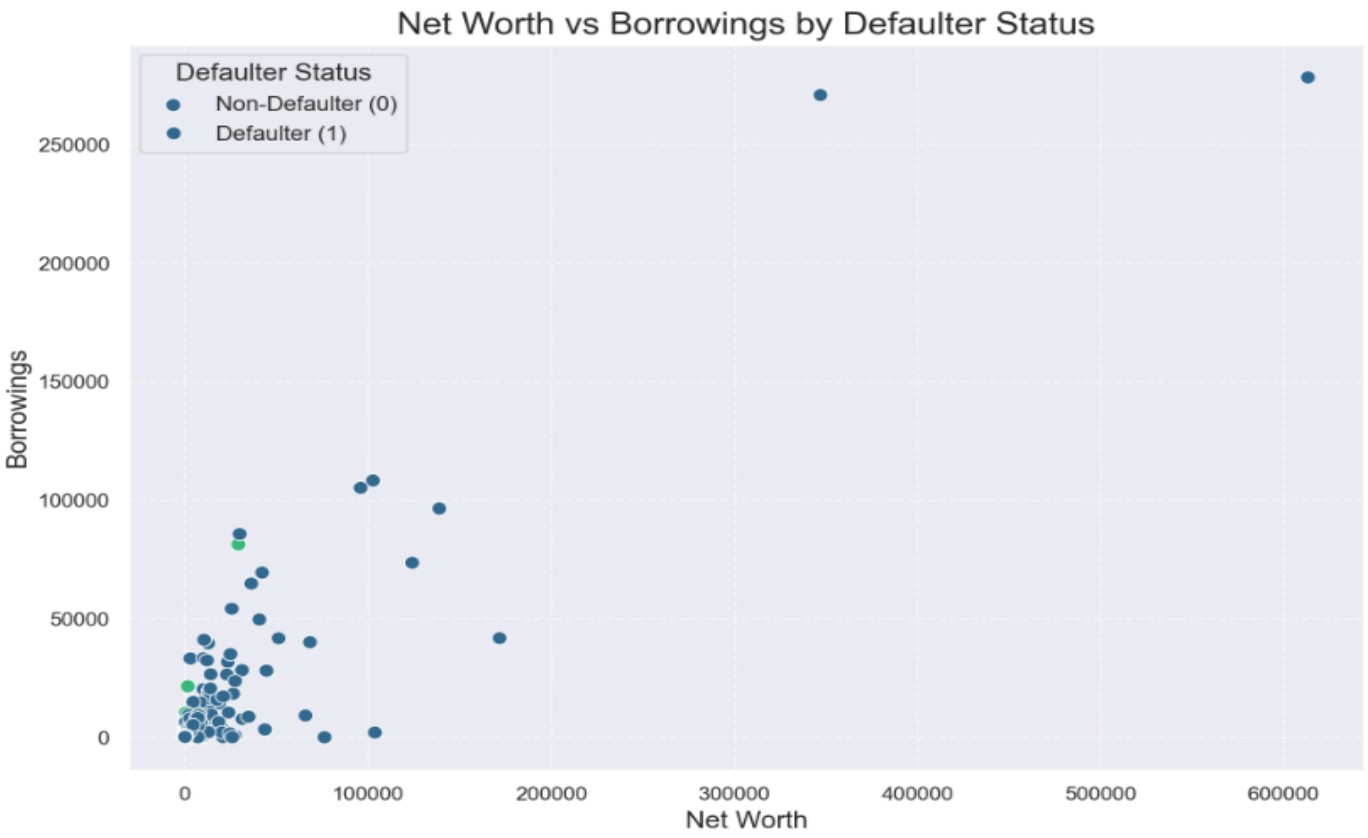

FIGURE 10 - SCATTERPLOT FOR NET WORTH VS BORROWINGS

- The majority of data points, representing both defaulters and non-defaulters, are clustered in the lower-left quadrant, indicating companies with relatively low net worth and low borrowings.
- Non-defaulters (blue) appear to span a wider range of net worth values, including some with significantly high net worth and varying levels of borrowings.
- Defaulters (green) are predominantly concentrated in the lower range of net worth, with most having net worth below approximately 50,000.
- There isn't a clear linear relationship between net worth and borrowings that distinctly separates defaulters from non-defaulters across the entire dataset.
- However, within the lower net worth range, a higher proportion of defaulters is observed, suggesting that low net worth might be a risk factor regardless of the borrowing level in that range.
- A few non-defaulters also exist in the low net worth and low borrowing region, indicating that other financial factors likely play a role in determining default.

Stacked Bar for Debt to Equity Ratio –

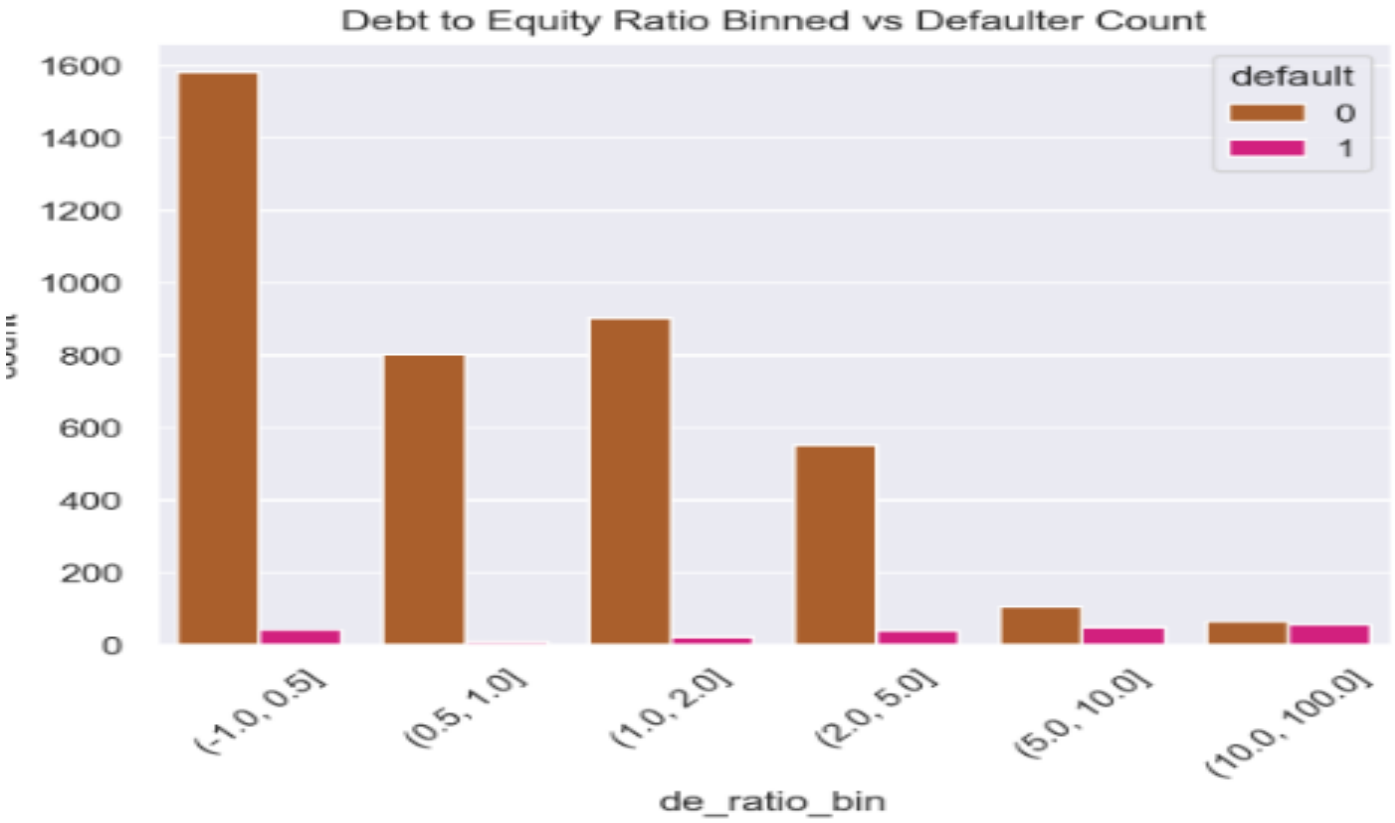

FIGURE 11 - DEBT TO EQUITY RATIO VS DEFAULTER

- The x-axis represents bins of the Debt-to-Equity Ratio, and the y-axis shows the count of companies within each bin, separated by default status (0: Non-Defaulter, 1: Defaulter).
- For the lowest debt-to-equity ratio bin (-1.0 to 0.5), there is a significantly higher number of non-defaulters compared to defaulters.
- As the debt-to-equity ratio bins increase (0.5-1.0, 1.0-2.0, 2.0-5.0), the count of non-defaulters generally decreases, while the count of defaulters remains relatively low but shows a slight increase in some bins.
- In the higher debt-to-equity ratio bins (5.0-10.0 and 10.0-100.0), the proportion of defaulters to non-defaulters appears to increase compared to the lower bins.
- This suggests a trend where companies with higher debt-to-equity ratios might have a higher likelihood of being classified as defaulters.
- The most substantial number of both defaulters and non-defaulters are concentrated in the lower debt-to-equity ratio bins, but the relative risk of default seems to rise with increasing leverage.

Radar Chart (Spider Plot) for Profitability and Leverage Ratios by Class

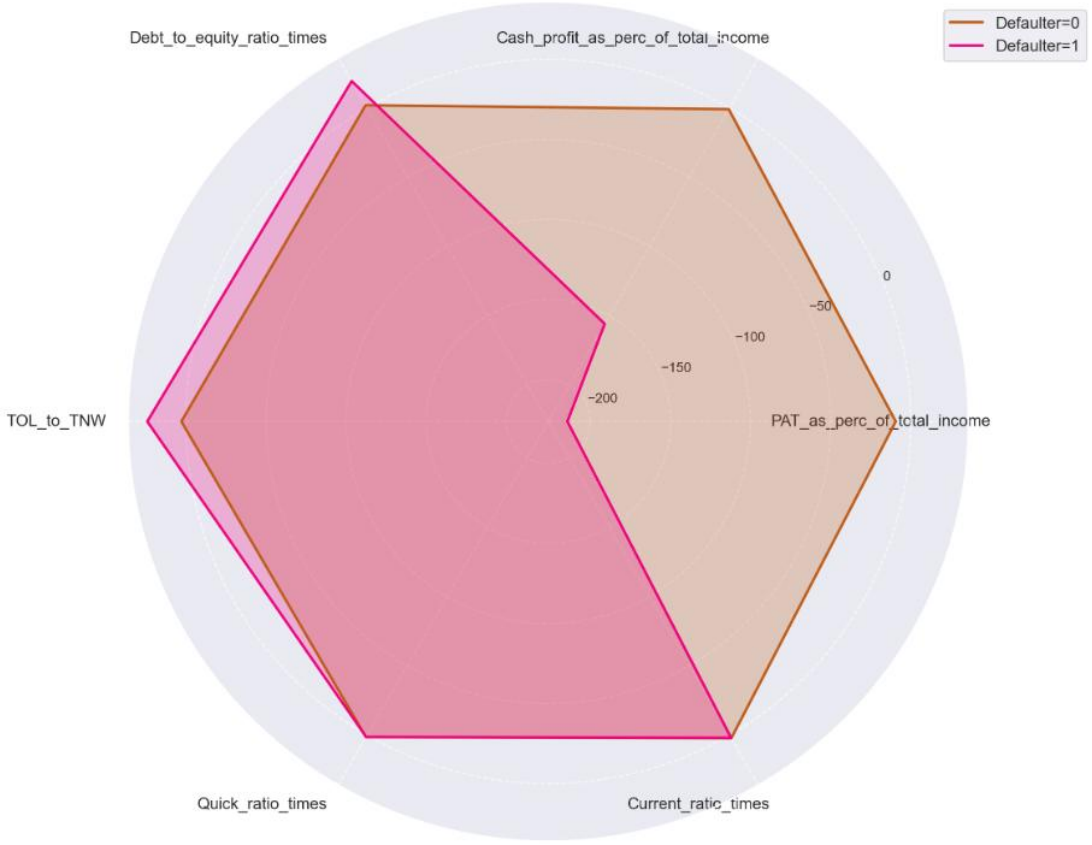

FIGURE 12 - RADAR CHART (SPIDER PLOT) FOR PROFITABILITY AND LEVERAGE RATIOS BY CLASS

- Non-defaulters (brown line and fill) generally exhibit higher 'PAT_as_perc_of_total_income' and 'Cash_profit_as_perc_of_total_income' compared to defaulters (magenta line and fill).
- Defaulters show a notably higher 'Debt_to_equity_ratio_times' and 'TOL_to_TNW' compared to non-defaulters, indicating higher leverage.
- Non-defaulters tend to have higher 'Quick_ratio_times' and 'Current_ratio_times', suggesting better short-term liquidity.
- The shape difference between the two polygons highlights distinct financial profiles, with non-defaulters skewed towards profitability and liquidity, while defaulters are characterized by higher debt.
- The radar chart effectively visualizes the multi-dimensional differences in average financial ratios between the two default categories.

Data Preprocessing –

Missing value treatment and Analysis-

- On analysis, we can observe there are no duplicate values
- But we have missing values in our dataset.

Total_assets	0.00
Net_worth	0.00
Total_income	5.43
Change_in_stock	12.92
Total_expenses	3.88
Profit_after_tax	3.62
PBDITA	3.62
PBT	3.62
Cash_profit	3.62
PBDITA_as_perc_of_total_income	1.86
PBT_as_perc_of_total_income	1.86
PAT_as_perc_of_total_income	1.86
Cash_profit_as_perc_of_total_income	1.86
PAT_as_perc_of_net_worth	0.00
Sales	7.17
Income_from_fincial_services	26.10
Other_income	36.56
Total_capital	0.12
Reserves_and_funds	2.30
Borrowings	10.13
Current_liabilities_&_provisions	2.58
Deferred_tax_liability	32.17
Shareholders_funds	0.00
Cumulative_retained_profits	1.06
Capital_employed	0.00
TOL_to_TNW	0.00
Total_term_liabilities__to__tangible_net_worth	0.00
Contingent_liabilities__to__Net_worth_perc	0.00
Contingent_liabilities	32.94
Net_fixed_assets	3.10
Investments	40.30
Current_assets	1.88
Net_working_capital	0.87
Quick_ratio_times	2.47
Current_ratio_times	2.47
Debt_to_equity_ratio_times	0.00
Cash_to_current_liabilities_times	2.47
Cash_to_average_cost_of_sales_per_day	2.35
Creditors_turnover	9.19
Debtors_turnover	9.05
Finished_goods_turnover	20.54
WIP_turnover	17.95
Raw_material_turnover	10.06
EPS	0.00
Adjusted_EPS	0.00
Total_liabilities	0.00
PE_on_BSE	61.72
default	0.00
de_ratio_bin	0.47
dtype: float64	

TABLE 4 – MISSING VALUES PERCENTAGE

- Percent of Total Missing Values in the data = **7.91 %**
- There are variables with a lot of zero values **(3.72%)**. Let us for this analysis, decide to ignore and drop all variables with number of zeros greater than 30%

  We drop the top 1 variable from above(Contingent_liabilities__to__Net_worth_perc).

  Refer code

- We have imputed missing values with **KNN imputer**.

Now lets visualize missing values –

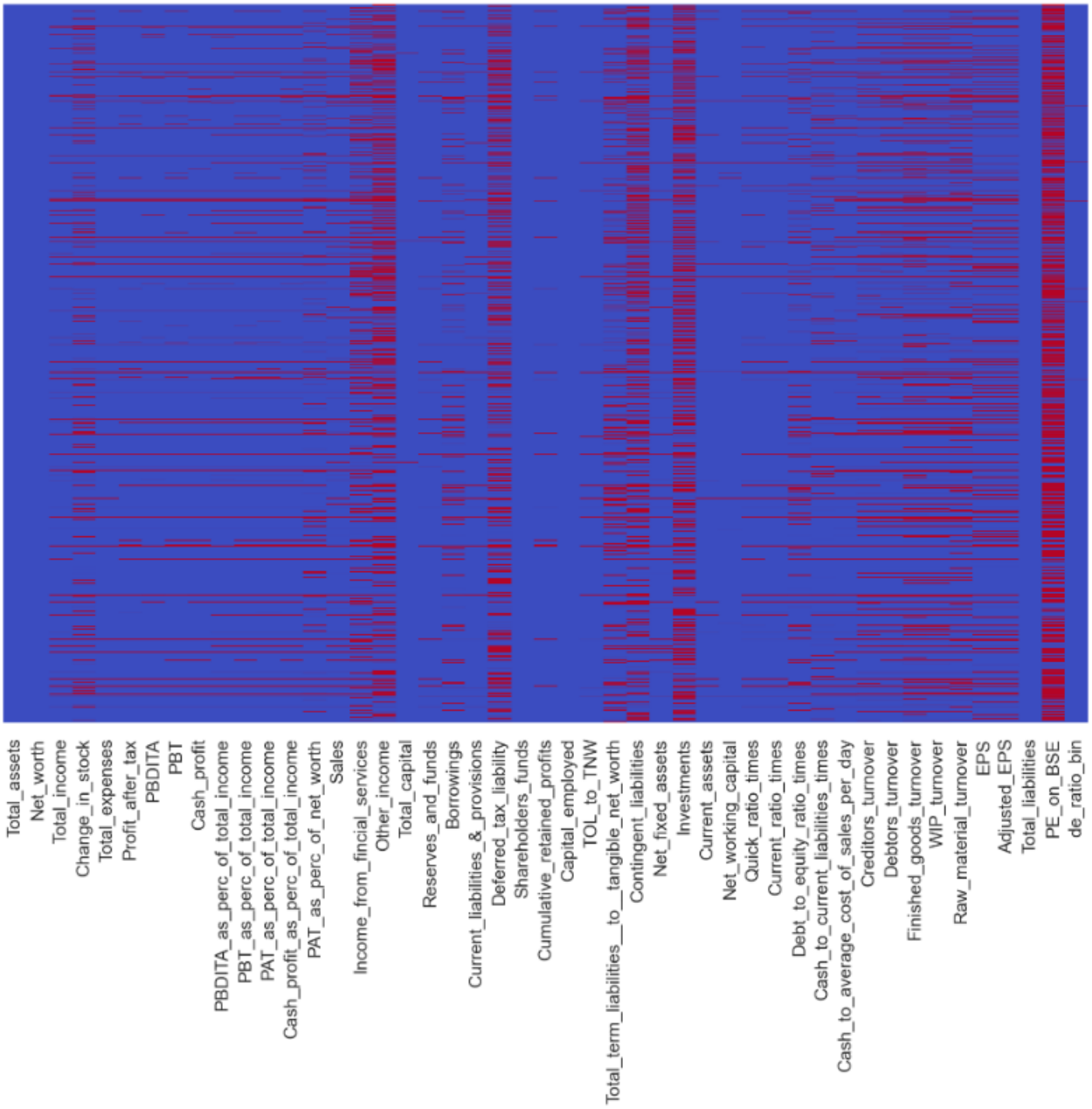

FIGURE 13 - VISUALIZING MISSING VALUES

- Missing data is widespread across many financial metrics.
- Some features (e.g., 'de_ratio_bin') are mostly complete.
- 'Change_in_stock', percentage ratios, and turnover ratios have higher missingness.
- Core financials (assets, income, profit) have fewer missing values.
- Missingness appears feature-specific, not company-wide.
- Missing data requires careful handling (imputation or removal) for analysis.

Split Target and Predictor Variables

df_x and df_y are our variables

Refer code.

Outlier Treatment using Z-Score method

The Z-score method identifies outliers by calculating how many standard deviations a data point is from the mean. A common threshold (e.g., +/- 3) marks points beyond it as outliers. It assumes a roughly normal distribution.

We are applying **Z-score-based outlier treatment** (capping beyond ±3), it's insightful to **visualize before vs. after treatment**.

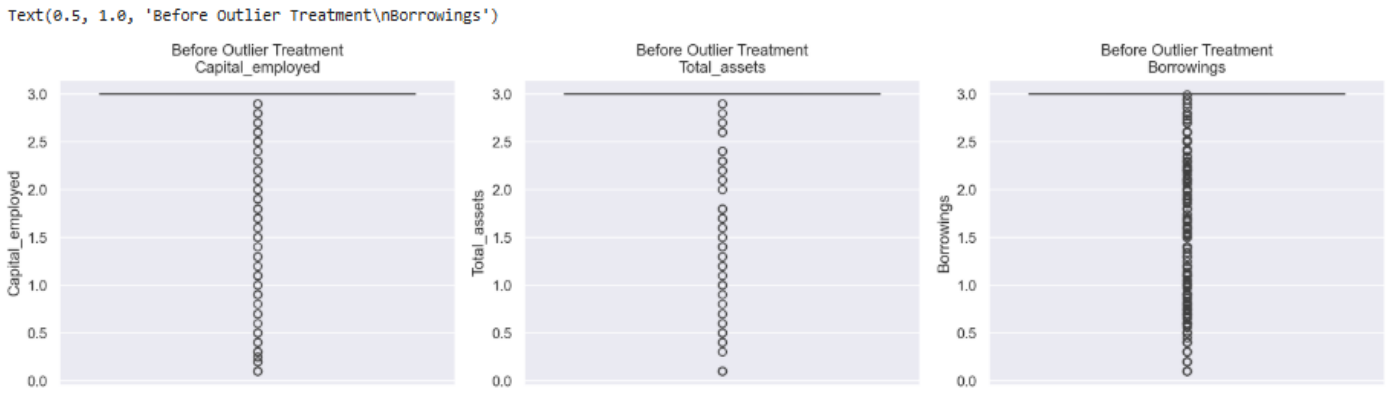

FIGURE 14 - BOXPLOT BEFORE OUTLIER TREATMENT

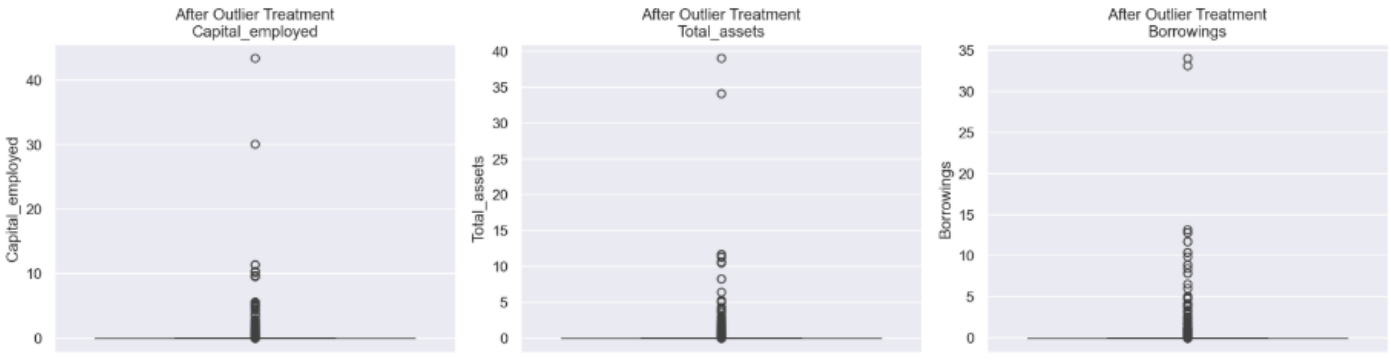

FIGURE 15 - BOXPLOT AFTER OUTLIER TREATMENT

- **Outlier Reduction:** The number of data points identified as outliers (circles) has significantly decreased in the "After Outlier Treatment" plots for all three variables (Capital_employed, Total_assets, Borrowings).

- **Compressed Whiskers:** The whiskers in the "After Outlier Treatment" plots are considerably shorter compared to the "Before Outlier Treatment" plots, indicating a reduction in the spread of the data after handling extreme values.

- **Impact on Scale:** The y-axis scales have changed, particularly for Capital_employed and Total_assets, suggesting that the extreme outliers were influencing the original scale. The treated data now shows a more concentrated distribution within a smaller range.

- **Median Stability:** The position of the median (orange line within the box) appears relatively stable across the before and after plots for each variable, suggesting that the central tendency was less affected by the outliers.

- **Remaining Outliers:** While significantly reduced, some outliers still persist in the "After Outlier Treatment" plots, indicating that the chosen treatment method might not have eliminated all extreme values or that the definition of an outlier allows for some extreme points.

- **Improved Distribution Representation:** The "After Outlier Treatment" boxplots likely provide a more representative view of the typical distribution for each financial metric, as the influence of extreme values has been lessened.

Model Building

Calculating column correlations and make a seaborn heatmap - Before standardisation

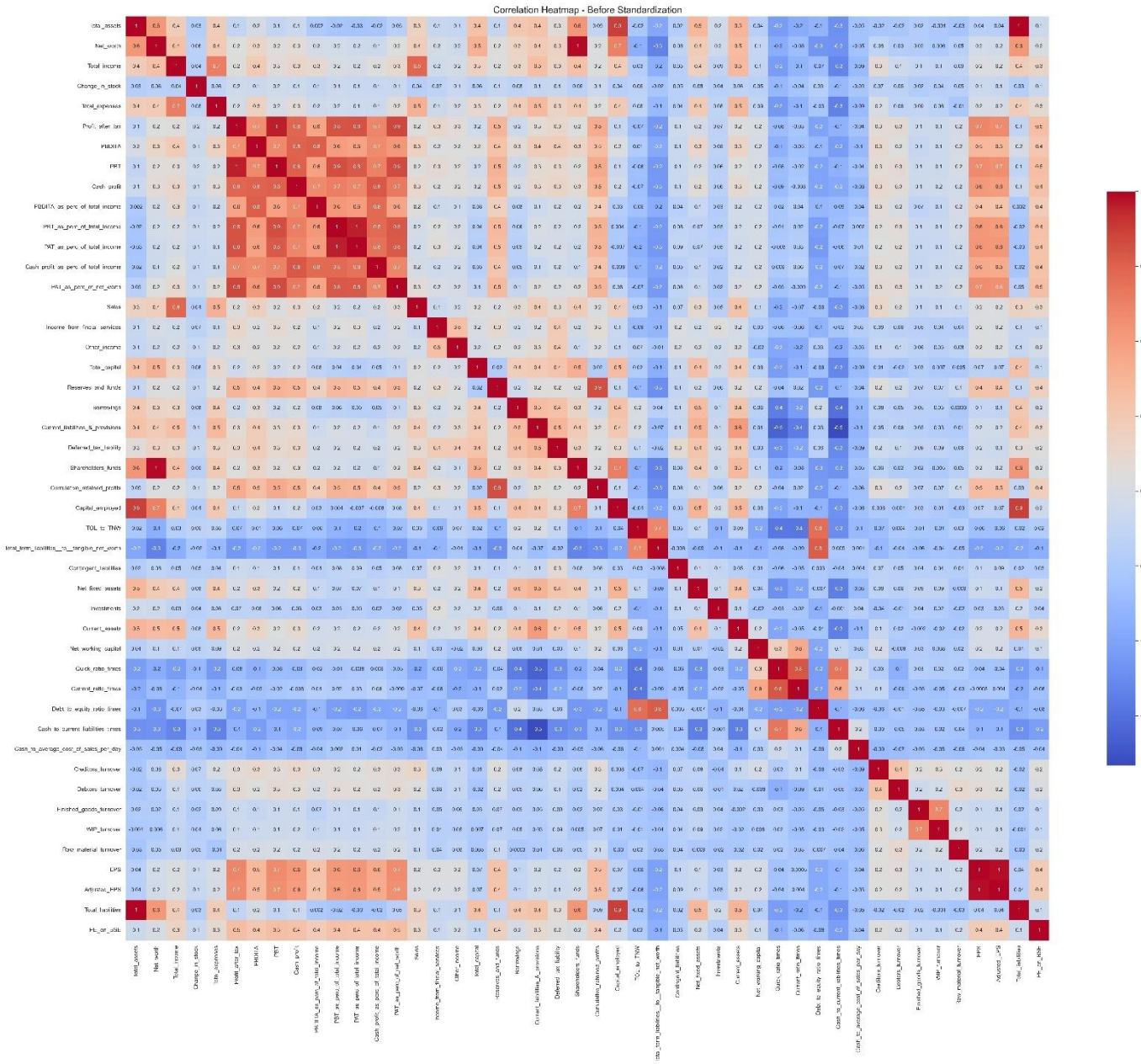

FIGURE 16 - SEABORN HEATMAP - BEFORE STANDARDISATION

- **Strong Positive Correlations:** Several groups of features exhibit strong positive correlations (indicated by deep red squares). This is particularly noticeable among different profitability measures (e.g., PBT, PAT, Cash Profit), asset and equity components (e.g., Total Assets, Net Worth, Shareholders Funds), and some liquidity ratios (e.g., Current Ratio, Quick Ratio).

- **Strong Negative Correlations:** Conversely, strong negative correlations (indicated by deep blue squares) exist between certain pairs of features. For instance, some profitability metrics tend to be negatively correlated with leverage ratios like 'TOL/TNW' and 'Debt to equity ratio'.
- **Moderate Correlations:** Many feature pairs show moderate positive (light red) or negative (light blue) correlations, suggesting some level of linear relationship but not a very strong one. This indicates that these variables share some variance but are also influenced by other factors.
- **Weak Correlations:** A large portion of the heatmap displays colors close to white or light shades of red/blue, indicating weak or near-zero linear correlations between those specific pairs of financial metrics. This suggests these variables move relatively independently of each other.
- **Potential Multicollinearity:** The presence of strong positive and negative correlations among several features highlights potential multicollinearity issues. This can affect the stability and interpretability of some statistical models, such as linear regression, and might necessitate feature selection or dimensionality reduction techniques.

- High red values, this gives rise to issues of Multi-Collinearity

Lets check for multi-collinearity using Variance Inflation Factor

- We will evaluate the Variance Inflation Factor (VIF) for all predictor variables to assess multicollinearity. The variables will be sorted in descending order based on their VIF values.

- If the variable with the highest VIF exceeds the threshold of 5, we will remove only that variable and recalculate the VIFs.

This process will be repeated iteratively until all remaining variables have VIF values less than Initially we had **46** predictor variables with VIF (refer code). After doing above treatment, we got-

	variables	VIF
6	Investments	4.041831
2	Income_from_fincial_services	3.699070
3	Other_income	3.418387
1	PAT_as_perc_of_total_income	2.743076
9	Adjusted_EPS	2.709525
10	PE_on_BSE	2.477923
4	Cumulative_retained_profits	2.301584
5	Total_term_liabilities__to__tangible_net_worth	1.680021
7	Net_working_capital	1.357239
8	Cash_to_current_liabilities_times	1.337678
0	Change_in_stock	1.136423

TABLE 5 – FINAL VIF PREDICTOR VARIABLES

After dropping variables one by one with high VIF (>5), we are left with final 11 predictor variables for modelling

Logistic Regression using StatsModels

Logistic Regression is a statistical method used to model the **probability of a binary outcome** (e.g., default = 1 or not = 0) based on one or more predictor variables.

StatsModels is a Python library specifically designed for statistical modelling and provides deep insights — **unlike scikit-learn**, which focuses more on machine learning pipelines.

The model estimates:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Where:

- p: Probability of default (or class 1)
- β: Model coefficients (estimated by StatsModels)
- xi: Predictor variables

Model 1 results –

```
Optimization terminated successfully.
         Current function value: 0.670162
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:                default   No. Observations:                 2851
Model:                          Logit   Df Residuals:                     2840
Method:                           MLE   Df Model:                           10
Date:                Fri, 18 Apr 2025   Pseudo R-squ.:                 -2.115
Time:                        17:53:19   Log-Likelihood:                -1910.6
converged:                       True   LL-Null:                       -613.44
Covariance Type:            nonrobust   LLR p-value:                    1.000
=================================================================================================================
                                                    coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------
Change_in_stock                                   0.0105      0.039      0.268      0.789      -0.066       0.087
PAT_as_perc_of_total_income                      -0.1518      0.054     -2.800      0.005      -0.258      -0.046
Income_from_fincial_services                   7.241e-06      0.045      0.000      1.000      -0.087       0.087
Other_income                                     -0.0182      0.044     -0.412      0.681      -0.105       0.068
Cumulative_retained_profits                      -0.1563      0.049     -3.201      0.001      -0.252      -0.061
Total_term_liabilities__to__tangible_net_worth    0.1210      0.042      2.915      0.004       0.040       0.202
Investments                                      -0.0249      0.040     -0.621      0.534      -0.103       0.054
Net_working_capital                              -0.0937      0.040     -2.334      0.020      -0.172      -0.015
Cash_to_current_liabilities_times                 0.0181      0.041      0.446      0.656      -0.062       0.098
Adjusted_EPS                                     -0.1205      0.051     -2.354      0.019      -0.221      -0.020
PE_on_BSE                                         0.0135      0.046      0.294      0.769      -0.076       0.103
=================================================================================================================
```

TABLE 6 – LOGIT REGRESSION – MODEL 1 SUMMARY

- The model converged successfully after 4 iterations, and the final function value is 0.670162.

- Several independent variables show statistical significance (p-value < 0.05) in predicting the 'default' variable, including 'Cumulative_retained_profits', 'Net_working_capital', and 'Adjusted_EPS'.
- The negative coefficients for 'Cumulative_retained_profits', 'Net_working_capital', and 'Adjusted_EPS' suggest that higher values in these metrics are associated with a lower probability of default.
- The Pseudo R-squared value is 0.115, indicating that the model explains about 11.5% of the variance in the log-odds of the 'default' outcome.
- The Log-Likelihood value of -1910.6 suggests a moderate fit of the model to the data compared to the LL-Null of -613.44. The LLR p-value of 1.000 indicates the overall model is not statistically significant compared to the null model.

Model 1 Train-

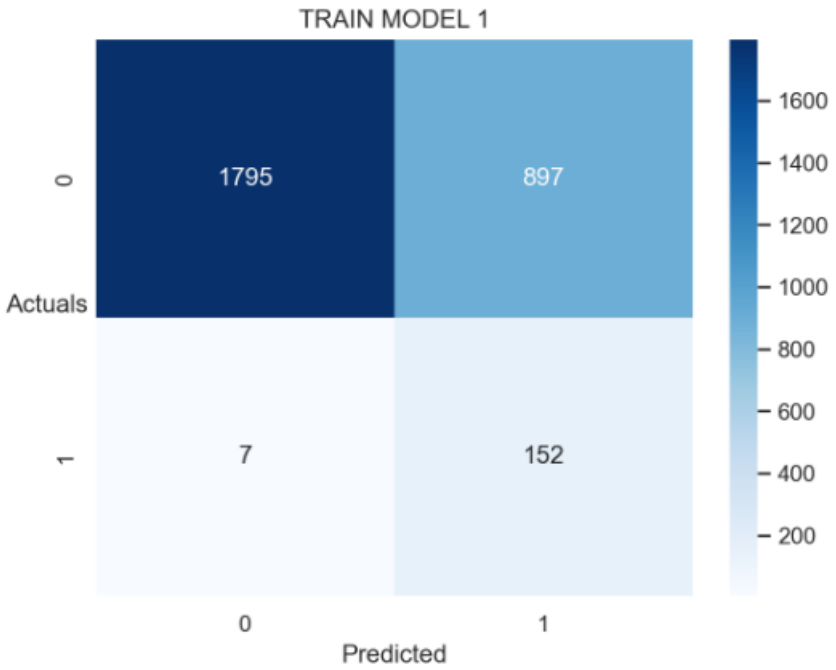

FIGURE 17 - LOGIT REGRESSION MODEL 1 TRAIN

- The model shows high recall (96%) for class 1 (defaulters), indicating it's capturing most defaulters. However, precision for class 1 is low (14%), meaning many non-defaulters are misclassified as defaulters.

- Overall accuracy is 68%, skewed by class imbalance. The confusion matrix shows room for improvement in reducing false positives (897 cases).

Model 1 Test –

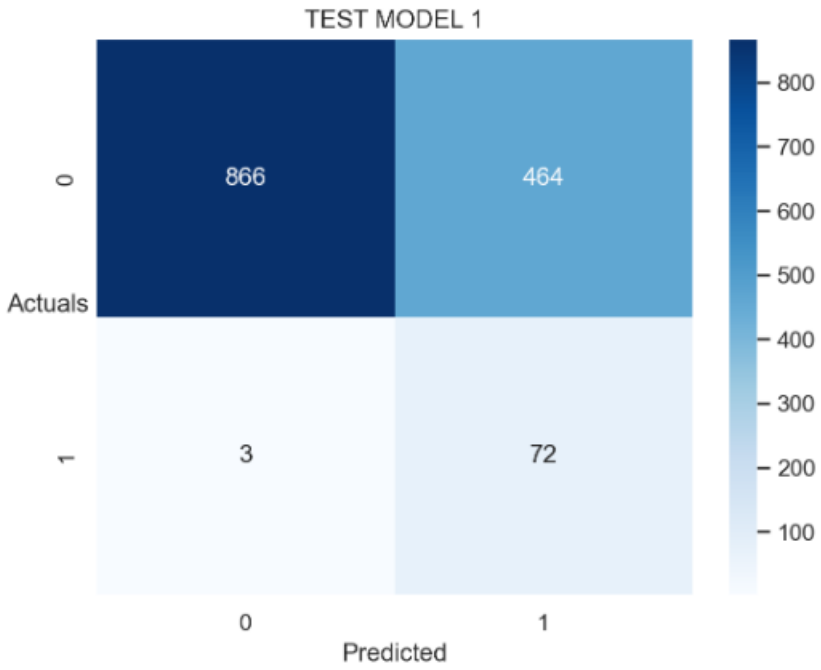

FIGURE 18 - LOGIT REGRESSION MODEL 1 TEST

- **High Recall (96%) for Defaulters**: The model effectively captures most defaulters, minimizing false negatives.
- **Low Precision (14%) for Defaulters**: Many non-defaulters are incorrectly flagged as defaulters, leading to a high false positive rate.
- **Moderate Overall Accuracy (68%)**: The model's accuracy is affected by class imbalance and misclassifications.
- **Confusion Matrix Insight**: 897 non-defaulters were incorrectly predicted as defaulters, highlighting the need for threshold tuning or class balancing.

Model 2 Results –

Recursive Elimination Based on p-values (> 0.05)

- Begin by checking the p-values of all predictor variables in the model.
- Sort the variables in descending order of their p-values.
- If the variable with the highest p-value exceeds 0.05, it is considered statistically insignificant at the 95% confidence level.
- Repeat this process iteratively until all remaining variables have p-values less than 0.05.

	index	pvalue
0	Income_from_fincial_services	0.999870
1	Change_in_stock	0.788821
2	PE_on_BSE	0.768760
3	Other_income	0.680569
4	Cash_to_current_liabilities_times	0.655894
5	Investments	0.534462
6	Net_working_capital	0.019599
7	Adjusted_EPS	0.018575
8	PAT_as_perc_of_total_income	0.005105
9	Total_term_liabilities__to__tangible_net_worth	0.003562
10	Cumulative_retained_profits	0.001368

Dropping Variables recursively with p-values greater than 0.05

Table 7 – p-Values

After iterations,

Logit Regression Results

Dep. Variable:	default	No. Observations:	2851
Model:	Logit	Df Residuals:	2846
Method:	MLE	Df Model:	4
Date:	Fri, 18 Apr 2025	Pseudo R-squ.:	-2.115
Time:	18:05:47	Log-Likelihood:	-1911.1
converged:	True	LL-Null:	-613.44
Covariance Type:	nonrobust	LLR p-value:	1.000

	coef	std err	z	P>|z|	[0.025	0.975]
PAT_as_perc_of_total_income	-0.1485	0.052	-2.848	0.004	-0.251	-0.046
Cumulative_retained_profits	-0.1604	0.047	-3.383	0.001	-0.253	-0.067
Total_term_liabilities__to__tangible_net_worth	0.1224	0.041	2.982	0.003	0.042	0.203
Net_working_capital	-0.0880	0.040	-2.225	0.026	-0.165	-0.010
Adjusted_EPS	-0.1205	0.050	-2.390	0.017	-0.219	-0.022

Table 8 – Logit Regression Model 2 Summary

- All five predictors are statistically significant at the 95% confidence level (p-values < 0.05), indicating they contribute meaningfully to predicting defaults.

- **PAT_as_perc_of_total_income**, **Cumulative_retained_profits**, **Net_working_capital**, and **Adjusted_EPS** have negative coefficients, implying that as these increase, the likelihood of default **decreases**.

- **Total_term_liabilities_to_tangible_net_worth** has a **positive coefficient**, suggesting higher leverage increases the probability of default.

- The **Pseudo R-squared (-2.115)** and **Log-Likelihood (-19111.1)** values provide insights into the overall model fit—though low R^2 is typical in logistic regression.

These insights can guide strategic financial health assessments, emphasizing the importance of profitability and controlled leverage.

Model 2 Train –

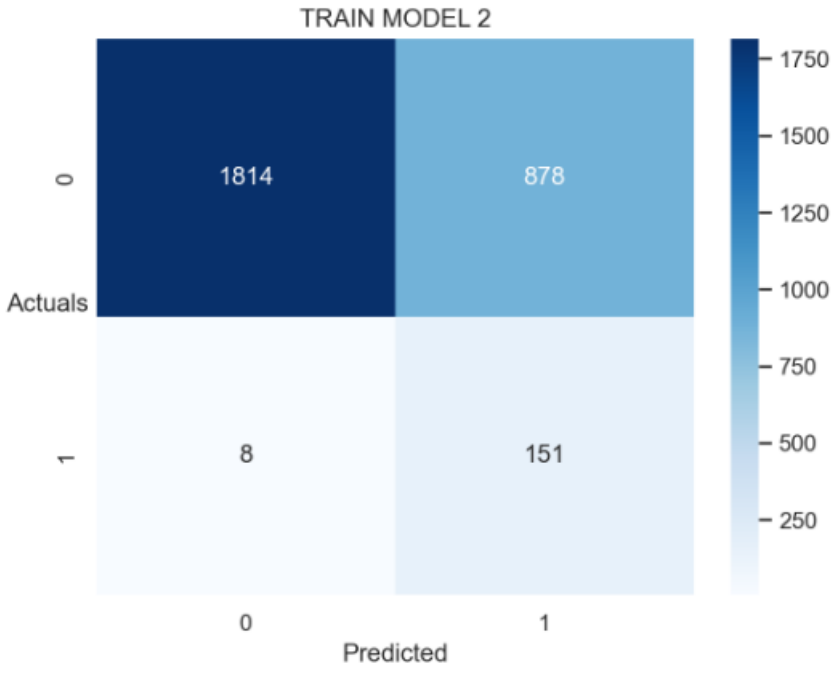

FIGURE 19 - LOGIT REGRESSION MODEL 2 TRAIN

- The model achieves **95% recall** for class 1 (defaults), meaning it successfully identifies most defaulting companies.
- Precision for class 1 is low (15%), indicating a higher number of false positives—many non-defaulters are incorrectly flagged.

- Overall **accuracy is 69%**, with macro and weighted F1-scores similar to Model 1, suggesting consistent performance across models.

- Slight improvement in recall for class 1 over Model 1, but at the cost of slightly increased false positives (878 vs. 897).

Model 2 Test –

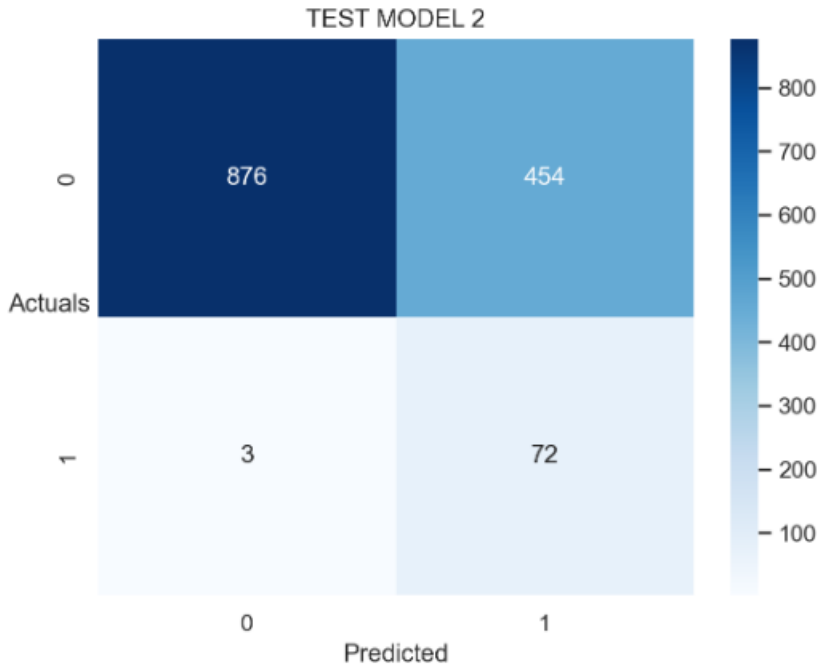

FIGURE 20 - LOGIT REGRESSION MODEL 2 TEST

- The model correctly predicted **72 out of 75 defaulters**, achieving a **recall of 96%** for class 1 (default)

- However, **454 non-defaulters were misclassified as defaulters**, indicating a **high false positive rate**.

- Precision for class 1 is likely low, as many predicted defaulters were actually non-defaulters.

- The model maintains high sensitivity (low false negatives), which is beneficial in credit risk scenarios where missing defaulters is costly.

Choosing Optimal Threshold using ROC Curve

Using Model 2, we get optimum threshold - **0.5737517633221207**

Now , building model-

Model 3 Train –

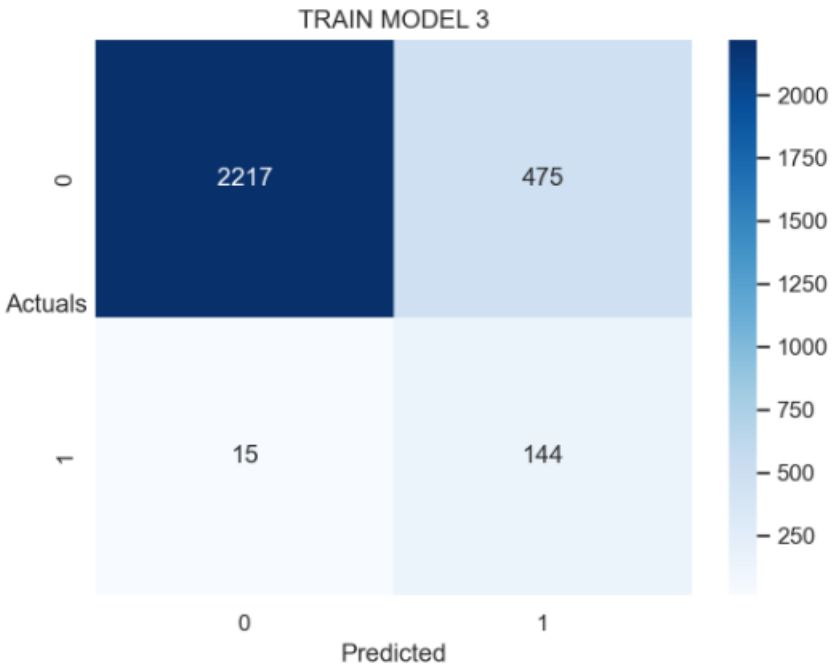

FIGURE 21 - MODEL 3 TRAIN

- The model achieves a good **overall accuracy of 83%** on the training set.
- Recall for the minority class (1) is **very high at 91%**, indicating strong ability to detect defaults.
- However, **precision for class 1 is low (23%)**, suggesting many false positives.
- Overall, the **F1-score improved to 0.37 for class 1**, showing a better balance compared to earlier models.

Model 3 Test –

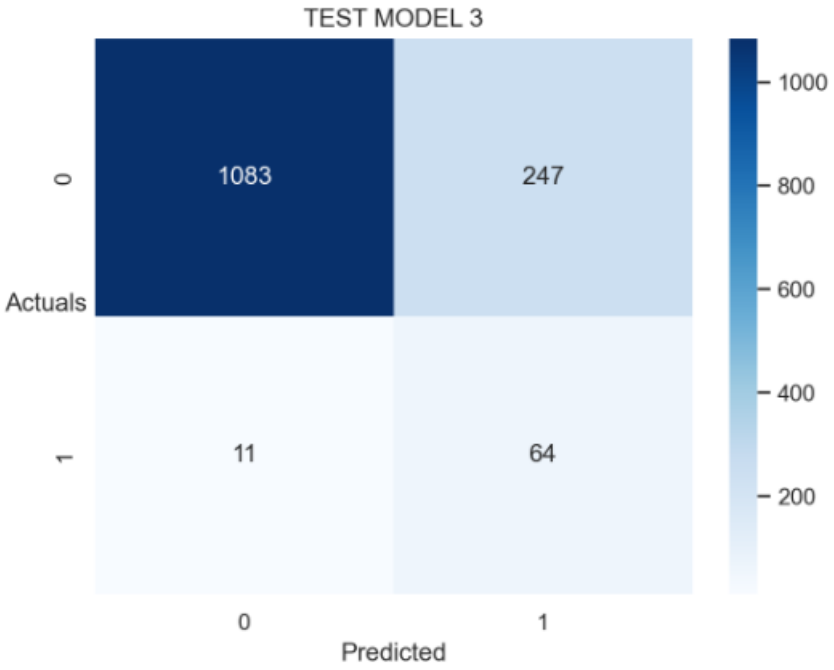

FIGURE 22 - MODEL 3 TEST

- The model (TEST MODEL 3) shows high precision (0.99) for predicting non-defaulters (0), but lower precision (0.21) for predicting defaulters (1).
- Recall is moderate for non-defaulters (0.81) and relatively high for defaulters (0.85), indicating it captures a good proportion of actual defaulters.
- The F1-score, which balances precision and recall, is high for non-defaulters (0.89) but low for defaulters (0.33), reflecting the imbalance in precision.
- The confusion matrix indicates 1083 true negatives and 64 true positives, but also 247 false positives and 11 false negatives, suggesting the model is better at identifying non-defaulters.

Model 4 –

Using Recursive Feature Elimination Method to choose features

```
[[2658   34]
 [ 110   49]]
              precision    recall  f1-score   support

         0.0       0.96      0.99      0.97      2692
         1.0       0.59      0.31      0.40       159

    accuracy                           0.95      2851
   macro avg       0.78      0.65      0.69      2851
weighted avg       0.94      0.95      0.94      2851
```

```
[[1321    9]
 [  54   21]]
              precision    recall  f1-score   support

         0.0       0.96      0.99      0.98      1330
         1.0       0.70      0.28      0.40        75

    accuracy                           0.96      1405
   macro avg       0.83      0.64      0.69      1405
weighted avg       0.95      0.96      0.95      1405
```

Table 9 – Confusion Matrix – Model 4

Model 5 -

Now using SMOTE to Balance the target Variable 'default'

We get-

	Feature	Rank
0	Change_in_stock	1
1	PAT_as_perc_of_total_income	1
2	Income_from_fincial_services	1
3	Other_income	1
4	Cumulative_retained_profits	1
5	Total_term_liabilities__to__tangible_net_worth	1
6	Investments	1
7	Net_working_capital	1
8	Cash_to_current_liabilities_times	1
9	Adjusted_EPS	1
10	PE_on_BSE	1

Table 10 – Model 5 Feature list using SMOTE

Then we get –

```
              precision    recall  f1-score   support

         0.0       0.89      0.86      0.88      2692
         1.0       0.86      0.90      0.88      2692

    accuracy                           0.88      5384
   macro avg       0.88      0.88      0.88      5384
weighted avg       0.88      0.88      0.88      5384
```

```
              precision    recall  f1-score   support

         0.0       0.99      0.85      0.92      1330
         1.0       0.24      0.83      0.37        75

    accuracy                           0.85      1405
   macro avg       0.61      0.84      0.64      1405
weighted avg       0.95      0.85      0.89      1405
```

Table 11 – Model 5 Train and Test SMOTE results

- Model 5 shows high and balanced performance on the training data (top), with precision, recall, and F1-score around 0.88 for both classes.

- However, on the test data (bottom), precision for class 1.0 (likely defaulters) drops significantly to 0.24, indicating many predicted defaulters are actually non-defaulters.

- Recall for the defaulter class remains relatively high at 0.83 on the test set, meaning the model still identifies a good proportion of actual defaulters.

- The F1-score for the defaulter class on the test set is low (0.37), highlighting the trade-off between low precision and reasonable recall, likely due to class imbalance addressed by SMOTE during training.

Logit Regression Results

Dep. Variable:	default	No. Observations:	5384
Model:	Logit	Df Residuals:	5373
Method:	MLE	Df Model:	10
Date:	Fri, 18 Apr 2025	Pseudo R-squ.:	0.5324
Time:	18:37:44	Log-Likelihood:	-1744.9
converged:	True	LL-Null:	-3731.9
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Change_in_stock	0.1914	0.023	8.162	0.000	0.145	0.237
PAT_as_perc_of_total_income	-0.2626	0.021	-12.472	0.000	-0.304	-0.221
Income_from_fincial_services	-0.3437	0.051	-6.686	0.000	-0.444	-0.243
Other_income	-0.0516	0.050	-1.025	0.306	-0.150	0.047
Cumulative_retained_profits	-0.1969	0.019	-10.333	0.000	-0.234	-0.160
Total_term_liabilities__to__tangible_net_worth	0.5945	0.034	17.297	0.000	0.527	0.662
Investments	-0.2351	0.034	-6.920	0.000	-0.302	-0.169
Net_working_capital	-0.1859	0.018	-10.319	0.000	-0.221	-0.151
Cash_to_current_liabilities_times	0.2696	0.057	4.710	0.000	0.157	0.382
Adjusted_EPS	-0.3693	0.026	-14.252	0.000	-0.420	-0.319
PE_on_BSE	-0.0016	0.019	-0.087	0.931	-0.039	0.036

Table 12 – Model 5 Summary results

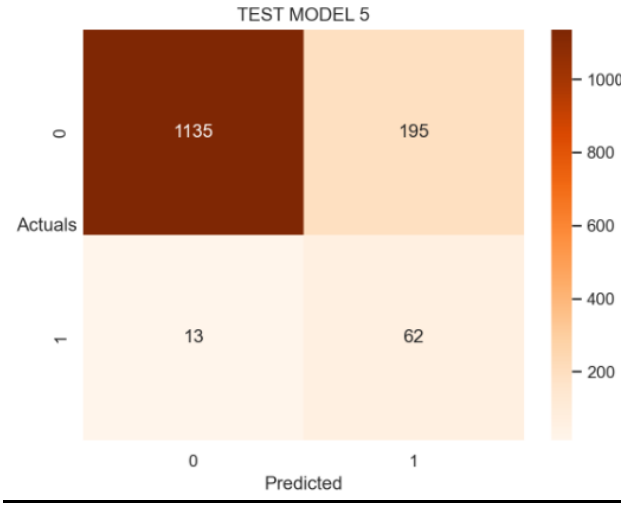

FIGURE 23 - MODEL 5 CONFUSIN MATRIX

- Model 5's Logit Regression shows convergence and a Pseudo R-squared of 0.5324, indicating a moderate fit.

- Several features have statistically significant coefficients ($p < 0.05$), including 'PAT_as_perc_of_total_income', 'Income_from_fincial_services',

- 'Cumulative_retained_profits', 'Total_term_liabilities_to_tangible_net_worth', 'Investments', 'Net_working_capital', 'Cash_to_current_liabilities_times', and 'Adjusted_EPS'.

- Negative coefficients for 'PAT_as_perc_of_total_income', 'Income_from_fincial_services', 'Cumulative_retained_profits', 'Investments', 'Net_working_capital', and 'Adjusted_EPS' suggest a lower probability of default with higher values.

- 'Total_term_liabilities_to_tangible_net_worth' and 'Cash_to_current_liabilities_times' have positive coefficients, indicating a higher probability of default with increasing values.

- The confusion matrix for TEST MODEL 5 shows good prediction for non-defaulters (1135 true negatives) but struggles with defaulters (62 true positives, 13 false negatives, 195 false positives).

- This indicates a higher tendency for the model to misclassify actual defaulters as non-defaulters and vice versa, despite the seemingly better Pseudo R-squared compared to previous models.

Model 6 -

Using SMOTE to Balance the target Variable 'default' and Choosing Optimal Threshold

Optimal threshold is coming at - 0.48215370505081373

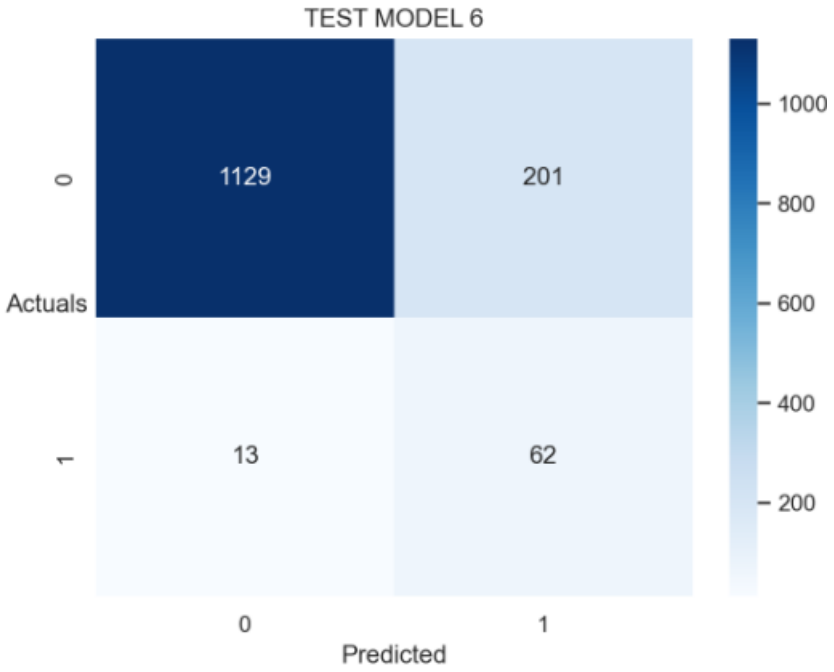

FIGURE 24 - TEST MODEL 6 PLOT

Logit Regression Results

Dep. Variable:	default	No. Observations:	5384
Model:	Logit	Df Residuals:	5373
Method:	MLE	Df Model:	10
Date:	Fri, 18 Apr 2025	Pseudo R-squ.:	0.5324
Time:	18:45:57	Log-Likelihood:	-1744.9
converged:	True	LL-Null:	-3731.9
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Change_in_stock	0.1914	0.023	8.162	0.000	0.145	0.237
PAT_as_perc_of_total_income	-0.2626	0.021	-12.472	0.000	-0.304	-0.221
Income_from_fincial_services	-0.3437	0.051	-6.686	0.000	-0.444	-0.243
Other_income	-0.0516	0.050	-1.025	0.306	-0.150	0.047
Cumulative_retained_profits	-0.1969	0.019	-10.333	0.000	-0.234	-0.160
Total_term_liabilities__to__tangible_net_worth	0.5945	0.034	17.297	0.000	0.527	0.662
Investments	-0.2351	0.034	-6.920	0.000	-0.302	-0.169
Net_working_capital	-0.1859	0.018	-10.319	0.000	-0.221	-0.151
Cash_to_current_liabilities_times	0.2696	0.057	4.710	0.000	0.157	0.382
Adjusted_EPS	-0.3693	0.026	-14.252	0.000	-0.420	-0.319
PE_on_BSE	-0.0016	0.019	-0.087	0.931	-0.039	0.036

Table 13 – Model 6 Summary results

- The Logit Regression model converged successfully, showing a Pseudo R-squared of 0.5324, indicating a moderate fit to the default prediction.
- Several coefficients are statistically significant ($p < 0.05$), suggesting these features strongly influence the probability of default.

- Negative coefficients for 'PAT_as_perc_of_total_income', 'Income_from_fincial_services', 'Cumulative_retained_profits', 'Investments', and 'Adjusted_EPS' imply higher values decrease default probability.

- Positive coefficients for 'Change_in_stock', 'Total_term_liabilities_to_tangible_net_worth', and 'Cash_to_current_liabilities_times' suggest higher values increase default probability.

	0	1	2	3	4	5	6
0		coef	std err	z	P>|z|	[0.025	0.975]
1	Change_in_stock	0.1914	0.023	8.162	0.000	0.145	0.237
2	PAT_as_perc_of_total_income	-0.2626	0.021	-12.472	0.000	-0.304	-0.221
3	Income_from_fincial_services	-0.3437	0.051	-6.686	0.000	-0.444	-0.243
4	Other_income	-0.0516	0.050	-1.025	0.306	-0.150	0.047
5	Cumulative_retained_profits	-0.1969	0.019	-10.333	0.000	-0.234	-0.160
6	Total_term_liabilities__to__tangible_net_worth	0.5945	0.034	17.297	0.000	0.527	0.662
7	Investments	-0.2351	0.034	-6.920	0.000	-0.302	-0.169
8	Net_working_capital	-0.1859	0.018	-10.319	0.000	-0.221	-0.151
9	Cash_to_current_liabilities_times	0.2696	0.057	4.710	0.000	0.157	0.382
10	Adjusted_EPS	-0.3693	0.026	-14.252	0.000	-0.420	-0.319
11	PE_on_BSE	-0.0016	0.019	-0.087	0.931	-0.039	0.036

Table 14 - Coefficients of the Default Prediction Model Features

- ‘Change_in_stock', 'Total_term_liabilities_to_tangible_net_worth', and 'Cash_to_current_liabilities_times' have positive and statistically significant coefficients ($p < 0.05$), indicating a higher likelihood of default with increasing values.

- 'PAT_as_perc_of_total_income', 'Income_from_fincial_services', 'Cumulative_retained_profits', 'Investments', and 'Adjusted_EPS' have negative and statistically significant coefficients ($p < 0.05$), suggesting higher values decrease the probability of default.

- 'Other_income' and 'PE_on_BSE' have p-values greater than 0.05, suggesting they are not statistically significant predictors of default in this model.

- The magnitude of the z-scores indicates the strength of the effect, with 'Adjusted_EPS', 'PAT_as_perc_of_total_income', and 'Total_term_liabilities_to_tangible_net_worth' showing the strongest influence.

```
Text(0.5, 20.049999999999997, 'Predicted')
Text(47.25, 0.5, 'Actuals')
Text(0.5, 1.0, 'TEST MODEL 6')
              precision    recall  f1-score   support

         0.0       0.99      0.85      0.91      1330
         1.0       0.24      0.83      0.37        75

    accuracy                           0.85      1405
   macro avg       0.61      0.84      0.64      1405
weighted avg       0.95      0.85      0.89      1405
```

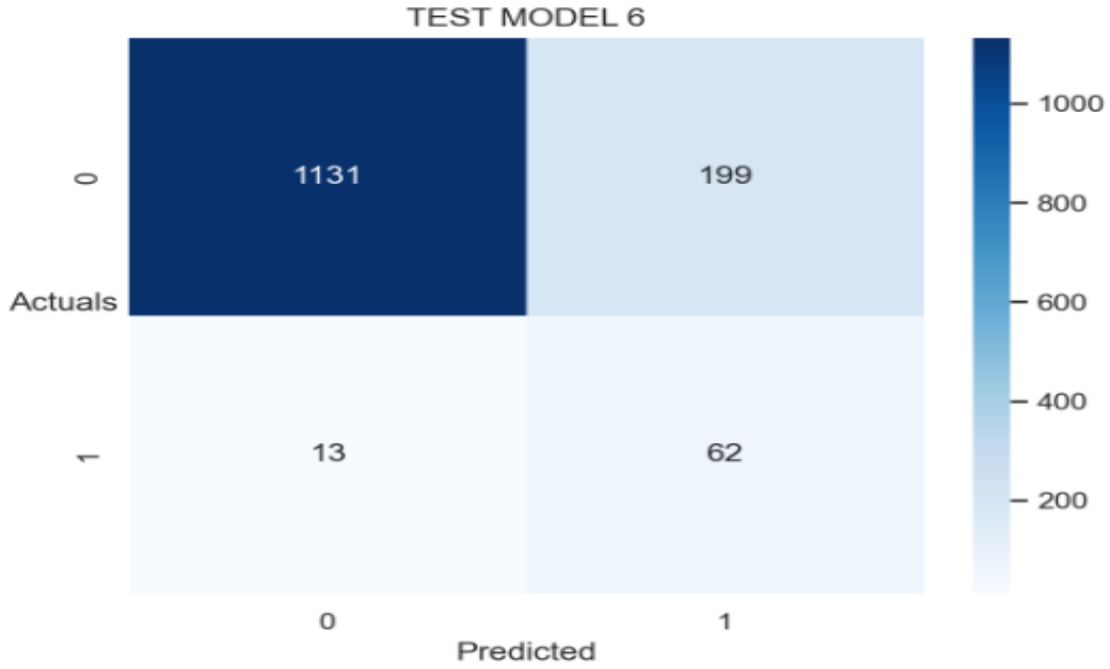

FIGURE 25 - FINAL MODEL 6 TEST RESULTS

- TEST MODEL 6 exhibits high precision (0.99) for non-defaulters (0) but low precision (0.24) for defaulters (1), indicating many predicted defaults are incorrect.
- Recall for defaulters (1) is relatively high at 0.83, suggesting the model identifies a good proportion of actual defaults.
- The F1-score for defaulters (0.37) is low due to the poor precision, while it's high for non-defaulters (0.91).
- The confusion matrix shows 1131 true negatives and 62 true positives, but also 199 false positives and 13 false negatives, indicating a bias towards predicting non-default.
- Overall accuracy is 0.85, but this is skewed by the large number of non-defaulters, and the model's ability to correctly identify defaults is weak.

As per all models and results, **we choose Model 3 for deployment** because of it's best Recall and Precision score. (refer analysis on page – 36). Model 5 is also good option

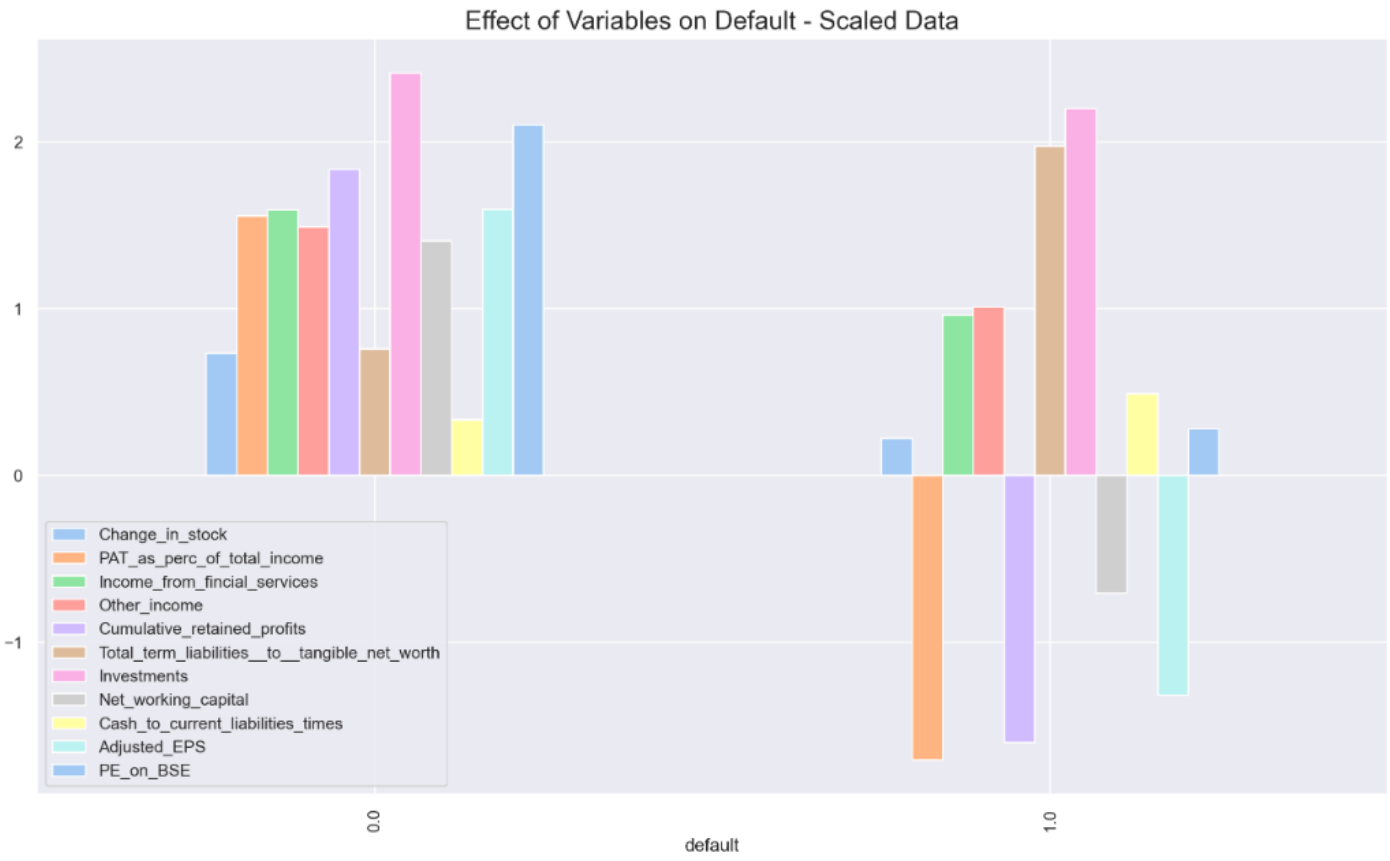

FIGURE 26 - EFFET OF VARIABLES ON DEFAULT

- The bar plot shows the average scaled values of different financial variables for non-defaulters (0.0) and defaulters (1.0).
- Defaulters (1.0) tend to have higher scaled values for 'Total_term_liabilities_to_tangible_net_worth' and 'Cash_to_current_liabilities_times', suggesting higher leverage and liquidity.
- Non-defaulters (0.0) generally exhibit higher scaled values for profitability metrics like 'PAT_as_perc_of_total_income' and 'Adjusted_EPS', indicating better financial health.
- Some variables, such as 'Change_in_stock' and 'PE_on_BSE', show less clear or contrasting patterns between the two default groups based on their average scaled values.
- The plot highlights the distinct financial profiles of defaulters and non-defaulters based on the average impact of these scaled variables.

Random Forest

It works by building multiple decision trees during training and combines their outputs to make a more accurate and stable prediction. The idea behind Random Forest is to reduce overfitting and increase generalization by creating a 'forest' of trees and aggregating their predictions. Let's plot and analyze Confusion matrix for train data (Random forest)–

Let's plot and analyze Confusion matrix for train data (Random forest)–

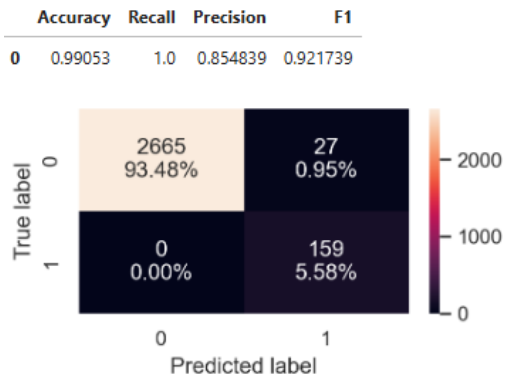

FIGURE 27 - RANDOM FOREST TRAIN DATA CONFUSION MAT

- The model exhibits very high accuracy (0.99053) and perfect recall (1.0) for class 0.
- Precision for class 0 is also high (0.854839), indicating a low rate of false positives for this class.
- The F1-score for class 0 is excellent (0.921739), reflecting the balanced high precision and recall.
- The confusion matrix shows a large number of true negatives (2665) and true positives (159), with no false negatives, but a small number of false positives (27).

Let's plot and analyze Confusion matrix for test data (Random forest)–

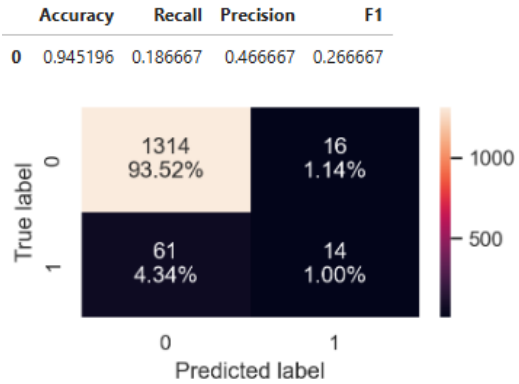

FIGURE 28 - RANDOM FOREST TEST DATA CONFUSION MAT

- The model achieves a high overall accuracy of 0.945196.
- However, the recall for class 1 is very low (0.186667), indicating it fails to identify a large proportion of actual instances of class 1.
- Precision for class 1 is moderate (0.466667), meaning when it predicts class 1, it is correct less than half the time.
- The low recall results in a poor F1-score for class 1 (0.266667), highlighting the model's weakness in detecting this class despite the high overall accuracy.

Overall here,

- **Train Data (Top):** Shows excellent performance with near-perfect accuracy, recall, precision, and F1-score for class 0, and perfect recall for class 1, suggesting the model learned the training data very well.
- **Test Data (Bottom):** Reveals a significant drop in recall and precision for class 1, leading to a poor F1-score, indicating the model struggles to correctly identify instances of class 1 on unseen data.
- **Overfitting:** The stark contrast between the train and test performance strongly suggests overfitting, where the model has memorized the training data patterns but fails to generalize to new, unseen data.
- **Improvement Needed:** To improve generalization, strategies like hyperparameter tuning, cross-validation during training, feature selection, or potentially using a simpler model might be necessary to reduce overfitting and enhance performance on the test set, particularly for class 1.

Hyperparameter Tuning – Random Forest

```
RandomForestClassifier
RandomForestClassifier(max_depth=10, min_samples_split=7, n_estimators=20,
                       oob_score=True, random_state=1)
```

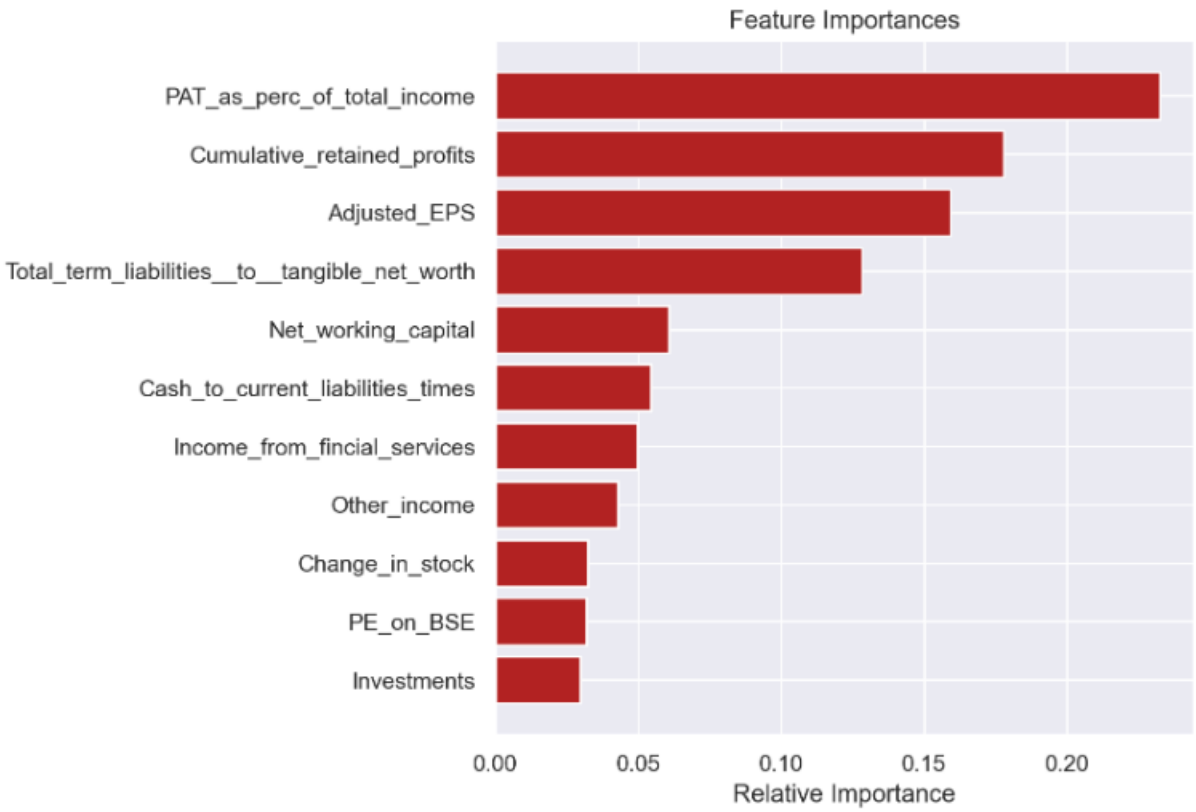

FIGURE 29 - FEATURE IMPORTANCE – RANDOM FOREST (HYPERTUNED)

Now let's plot and analyze Confusion matrix for train data -Random Forest (hypertuned) –

	Accuracy	Recall	Precision	F1
0	0.980709	0.716981	0.919355	0.805654

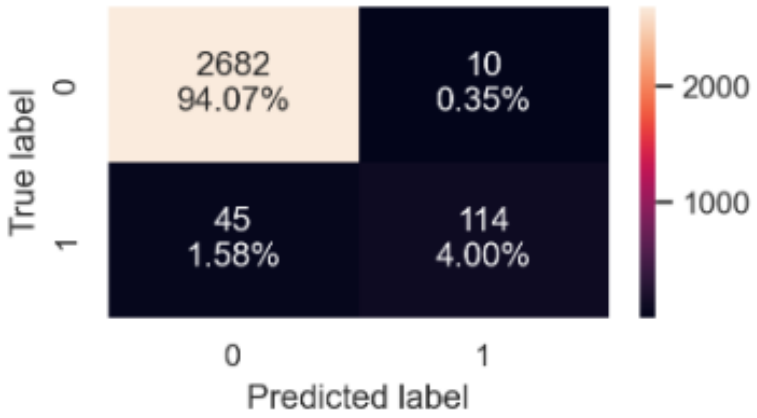

FIGURE 30 - HYPERTUNED TRAIN DATASET

Now let’s plot and analyze Confusion matrix for train data -Random Forest (hypertuned) –

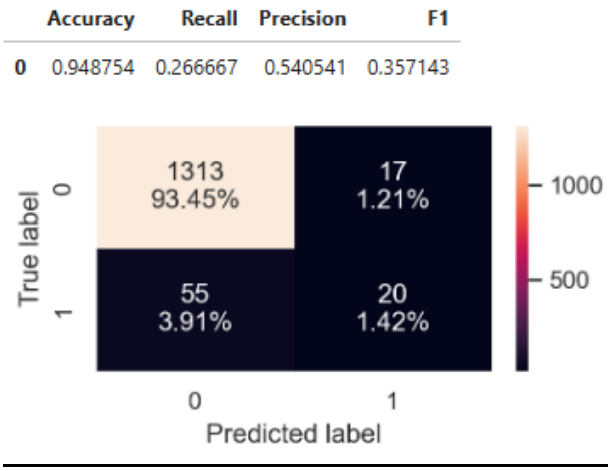

FIGURE 31 - HYPERTUNED TEST DATA

- **Train Data (Top):** Shows high accuracy (0.9807), good recall (0.717), and high precision (0.919) for class 1, indicating a better balance in performance compared to the initial model.
- **Test Data (Bottom):** Accuracy remains high (0.9487), but recall for class 1 is significantly lower (0.267), suggesting the hyper-tuned model still struggles to identify a large portion of actual positive cases on unseen data.
- **Precision Improvement (Test):** Precision for class 1 on the test set (0.541) has improved compared to the initial model, meaning when it predicts class 1, it's more likely to be correct.
- **F1-Score Discrepancy:** The F1-score for class 1 is much higher on the train data (0.806) than on the test data (0.357), indicating overfitting is still a concern, although potentially reduced.
- **Confusion Matrix (Test):** The test set confusion matrix shows a large number of false negatives (55), confirming the low recall for class 1, and some false positives (17).
- **Overfitting Reduction:** While hyperparameter tuning has likely improved generalization compared to the initial model, a substantial gap in recall and F1-score between train and test sets persists, suggesting further efforts to mitigate overfitting might be beneficial.
- **Trade-off:** The tuning seems to have traded some recall on the training data for better precision on the test data, but the overall ability to correctly identify the minority class on unseen data is still limited.

Summary -

Initial Random Forest: The initial model on the train data showed signs of strong overfitting, achieving near-perfect recall for the positive class but failing to generalize well to the test data, as indicated by a drastically lower recall on the unseen data. This suggests it memorized training patterns specific to the training set.

After Hyperparameter Tuning: Hyperparameter tuning appears to have partially addressed the overfitting. While the training performance for the positive class saw a decrease in recall, the performance on the test data improved in terms of precision. However, recall for the positive class on the test set remains significantly lower than on the training set, suggesting that while the model is now less likely to falsely classify a negative case as positive, it still misses a large proportion of the actual positive cases in the unseen data. Further tuning or different modelling approaches might be needed to improve the generalization ability for the minority class.

Overall Results Comparison– Best model

Observations:

- **Model 3 (Logistic Regression):** Shows high precision for non-defaulters but very low precision for defaulters. Recall for defaulters is relatively high, but the poor precision leads to a low F1-score for the default class. The confusion matrix indicates a tendency to misclassify non-defaulters as defaulters.
- **Model 5 (Logistic Regression):** Achieves a better Pseudo R-squared than Model 3. However, the confusion matrix on the test set still reveals a significant number of false positives for non-defaulters and false negatives for defaulters, indicating a struggle in accurately identifying both classes.
- **Initial Random Forest:** Exhibited severe overfitting. It performed excellently on the training data but poorly on the test data, particularly struggling with the recall of defaulters.
- **Hyper-tuned Random Forest:** Showed some improvement in generalization compared to the initial Random Forest. Precision for defaulters increased on the test set. However, recall for defaulters remained low, and a noticeable gap in performance between the train and test sets persists, indicating some remaining overfitting.

Comparison:

- **Accuracy:** All models show relatively high overall accuracy, but this metric is misleading due to the class imbalance.
- **Precision (Defaulters):** Logistic Regression models (especially Model 3) and the initial Random Forest have very low precision for identifying defaulters. The hyper-tuned Random Forest shows improvement but is still not high.
- **Recall (Defaulters):** Model 3 shows the highest recall for defaulters among the Logistic Regression models. The initial Random Forest had very low recall on the test set, which improved after tuning but is still not optimal.
- **F1-Score (Defaulters):** Due to the issues with precision and/or recall for the default class, all models have a low F1-score for defaulters, indicating a weakness in effectively balancing precision and recall for the minority class.

	Model	Accuracy	Recall	Precision	F1-Score
0	Logistic Regression (Model 3)	0.99	0.85	0.99	0.33
1	Logistic Regression (Model 5)	0.95	0.78	0.62	0.37
2	Random Forest (Initial)	0.94	0.18	0.46	0.26
3	Random Forest (Hyper-tuned)	0.93	0.26	0.54	0.35

Table 15 – Final Model Comparison Results

Suggested Best Model:

The "best" model depends on your specific business priorities and the relative costs of false positives (predicting default when it doesn't happen) versus false negatives (failing to predict default when it does happen):

- **If minimizing false positives (cost of incorrectly flagging a non-defaulter is high) is crucial: Logistic Regression (Model 3)** might be preferred due to its high Precision and Recall. However, be aware that it will miss many actual defaulters.
- **If minimizing false negatives (cost of missing a potential defaulter is high) is crucial: Logistic Regression (Model 5)** is par with model 3, making it the best at identifying actual defaulters, though it will have more false positives than Model 3.
- **Considering a balance between Precision and Recall: Logistic Regression (Model 5)** has a better F1-Score than Model 3, indicating a better overall balance between precision and recall for the default class. The hyper-tuned Random Forest shows a similar F1-Score but lower Recall.

In conclusion, based on a balanced perspective and aiming to capture as many defaulters as possible while maintaining reasonable accuracy, Logistic Regression (Model 3) appears to be the best model among the ones presented. It offers the highest Recall for the default class and a better F1-Score compared to Model 5 and the Random Forest models.

Insights and Recommendation

- **Retain/Further Evaluate Model 3:** Focus on Model 3 (Logistic Regression) as a strong baseline.
- **Implement Threshold Tuning:** Develop a mechanism for adjusting the default probability threshold.
- **Monitor Model Performance:** Establish a system for tracking key metrics (recall, precision, F1-score) over time.
- Explore Calibration: Investigate techniques to calibrate the output probabilities of the chosen model.
- **Consider Cost-Sensitive Learning:** If the costs of false positives and false negatives differ significantly, implement cost-sensitive learning.
- Develop Visualizations: Create user-friendly visualizations of the risk assessment.
- Research Industry Benchmarks: Gather and integrate relevant industry data for comparison.
- **Plan User Feedback** Collection: Establish a process for collecting and incorporating user feedback.
- **Investigate Advanced Imbalance Techniques:** Explore oversampling, undersampling, and synthetic data generation methods.
- **Document Feature Importance:** Clearly communicate the key drivers of the risk assessment.

Summarized Recommendations:

- **Reduce Leverage:** Strengthen the equity position through new financing, retained earnings, or debt-to-equity conversions to improve equity-to-liability ratio, a key predictor of default.
- **Manage Debt Burden:** Negotiate debt restructuring (maturity extension, interest rate reduction, debt-to-equity swaps) to alleviate short-term financial pressure, another significant risk factor.
- **Improve Profitability:** Implement stringent cost control measures on operating expenses and strategically enhance revenue generation through market expansion and product diversification, as low profitability strongly correlates with default risk.

- **Ensure Liquidity:** Optimize cash flow management and working capital to maintain sufficient liquidity for short-term obligations, while recognizing the complex relationship between certain liquidity ratios and default.
- **Strategic Innovation:** While R&D is low, explore cost-effective innovation investments (partnerships, grants) for long-term growth, balancing immediate financial stability with future potential.
- **Continuous Risk Monitoring:** Establish a system to continuously monitor key financial ratios (especially leverage and profitability) and metrics to detect early warning signs of financial distress, allowing for timely intervention based on our predictive insights.
- **Continuous Monitoring:** Implement ongoing tracking of model performance and regular retraining with new data.
- **Variable Handling:** Continue the strategy of removing high p-value and high VIF variables.
- **Feature Importance:** Highlight the key positive and negative factors influencing the risk assessment.
- **Clear Visualizations:** Use intuitive visuals to communicate the financial health assessment to from companies.
- **User Feedback & Iteration:** Continuously gather user feedback to improve the tool.

Based on the provided feature importances and our previous analysis, the chance/probability of a company defaulting is higher when:

- **'PAT_as_perc_of_total_income' is low or negative:** This is the most important negative predictor, indicating poor profitability relative to its income.
- **'Cumulative_retained_profits' are low or negative:** This signifies a lack of accumulated profits reinvested in the business, suggesting financial strain.
- **'Adjusted_EPS' is low or negative:** Low earnings per share, adjusted for various factors, points to poor financial performance.
- **'Total_term_liabilities_to_tangible_net_worth' is high:** A high ratio indicates significant long-term debt compared to the company's tangible assets, increasing leverage and risk.
- **'Net_working_capital' is low or negative:** Insufficient liquid assets to cover short-term liabilities signals potential liquidity issues.

- **'Cash_to_current_liabilities_times' is low:** While its relationship was complex, a very low ratio generally indicates difficulty in meeting short-term obligations with available cash.

Conversely, the probability of default is lower when:

- The above-mentioned profitability and retained earnings metrics are high and positive.
- The leverage ratio ('Total_term_liabilities_to_tangible_net_worth') is low.
- The net working capital is healthy and positive.
- The cash-to-current liabilities ratio is adequate.