PGP-DSBA PROJECT REPORT

FRA – Guided Project

PART A

BY

ISHAAN SHAKTI JAYARAMAN

PGPDSBA.O.JULY24.A

Contents

LIST OF FIGURES 2

INTRODUCTION 3

1.1 Objective 3

EXPLORATORY DATA ANALYSIS 4

2.1 Problem Definition 4

2.2 Data Contents 4

2.3 Data Dictionary 4

2.4 Statistical Summary 7

2.5 Univariate Analysis 9

2.5.1 Default 9

2.5.2 Numerical Variables 10

2.6 Bivariate Analysis 13

2.6.1 Numerical Variables 13

2.6.2 Correlation Matrix 15

DATA PRE-PROCESSING 16

3.1 Data Preparation 16

MODEL BUILDING 17

4.1 Logistic Regression Model 17

4.2 Random Forest Model 19

MODEL PERFORMANCE IMPROVEMENT 22

5.1 Optimised Logistic Regression Model 22

5.2 Optimised Random Forest Model 26

MODEL PERFORMANCE COMPARISION AND FINAL MODEL SELECTION 28

ACTIONABLE INSIGHTS AND RECOMMENDATIONS 30

LIST OF FIGURES

Figure 1 - Statistical Summary of the dataset 7
Figure 2 - Countplot of "Default" 9
Figure 3 - Boxplots of all Numerical Variables 10
Figure 4 - Distribution plots of all Numerical Variables 11
Figure 5 - Boxplot of all Numerical Variables vs Default 13
Figure 6 - Correlation Matrix 15
Figure 7 - Snippet of the Scaled Training Data 16
Figure 8 - Logistic Regression Model 17
Figure 9 - Confustion Matrix - Logistic Regression Model on Training Data 18
Figure 10 - Confustion Matrix - Logistic Regression on Test Data 19
Figure 11 - Confusion Matrix - Random Forest Model on Training Data 20
Figure 12 - Confusion Matrix - Random Forest Model on Test Data 21
Figure 13 - Optimised Logistic Regression Model 22
Figure 14 - ROC Curve 23
Figure 15 - Confusion Matrix - Optimised Logistic Regression Model on Training Data 24
Figure 16 - Confusion Matrix - Optimised Logistic Regression Model on Test Data 25
Figure 17 - Confusion Matrix - Optimised Random Forest Model on Training Data 26
Figure 18 - Confusion Matrix - Optimised Random Forest Model on Test Data 27
Figure 19 - Training Performance Comparison 28
Figure 20 - Test Performance Comparison 28
Figure 21 - Feature Importances 29

INTRODUCTION

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

1.1 Objective

A renowned credit rating organization wants 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, the organization aims to analyze historical financial statements and extract pertinent insights to facilitate informed decision-making via the tool. Specifically, the organization foresees 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.

As a part of the data science team in the organization, you have been provided 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 default on its debt repayments in the next two quarters. The predictive model will help the organization anticipate potential challenges with the financial performance of the companies and enable proactive risk mitigation strategies.

EXPLORATORY DATA ANALYSIS

2.1 Problem Definition

The objective is to develop a Financial Health Assessment Tool using machine learning to evaluate the creditworthiness of companies based on historical financial data. The tool will analyze key financial indicators to assess debt management practices and predict the likelihood of default in upcoming quarters. By leveraging liquidity ratios, debt-to-equity ratios, and other financial metrics, the model will provide actionable insights for investors and businesses. This predictive system aims to facilitate proactive risk mitigation strategies and improve financial decision-making. Ultimately, it will empower stakeholders with a data-driven mechanism to ensure stability and profitability in uncertain market conditions.

2.2 Data Contents

The dataset (CompData.xlsx) consists of financial metrics from the balance sheets of different companies.

- There are 2058 observations in the dataset.
- There are 58 columns in the dataset.
- The columns consist of 57 numerical variables and 1 object variable.
- 4 of the numerical variables are integer type data and the rest are float type data.
- There are no duplicate entries in the dataset.

2.3 Data Dictionary

- Co_Code – Company Code
- Co_Name – Company Name
- Operating_Expense_Rate – Operating Expenses / Net Sales (Measures operating cost efficiency)
- Research_and_development_expense_rate – R&D Expenses / Net Sales (Investment in innovation)
- Cash_flow_rate – Cash Flow from Operating / Current Liabilities (Liquidity measure)
- Interest_bearing_debt_interest_rate – Interest-bearing Debt / Equity (Debt burden indicator)
- Tax_rate_A – Effective Tax Rate (Percentage of taxable income paid in taxes)
- Cash_Flow_Per_Share – After-tax earnings plus depreciation per share (Financial strength measure)

- Per_Share_Net_profit_before_tax_Yuan – Pretax Income Per Share (Earnings before tax per share)
- Realized_Sales_Gross_Profit_Growth_Rate – Growth in gross profit from sales
- Operating_Profit_Growth_Rate – Growth rate of operating income over the last year
- Continuous_Net_Profit_Growth_Rate – Net income growth excluding disposal gains/losses
- Total_Asset_Growth_Rate – Growth rate of company's assets
- Net_Value_Growth_Rate – Growth in total equity
- Total_Asset_Return_Growth_Rate_Ratio – Return on total assets growth
- Cash_Reinvestment_perc – Percentage of cash flow reinvested into business operations
- Current_Ratio – Ratio between company's current assets and liabilities
- Quick_Ratio – Quick assets / Current liabilities (Liquidity measure)
- Interest_Expense_Ratio – Interest Expenses / Total Revenue
- Total_debt_to_Total_net_worth – Total Liability / Equity Ratio
- Long_term_fund_suitability_ratio_A – (Long-term Liability + Equity) / Fixed Assets
- Net_profit_before_tax_to_Paid_in_capital – Pretax Income / Capital
- Total_Asset_Turnover – Net Sales / Average Total Assets
- Accounts_Receivable_Turnover – How efficiently companies collect outstanding debts
- Average_Collection_Days – Number of days receivables remain outstanding
- Inventory_Turnover_Rate_times – Number of times inventory is sold and replenished
- Fixed_Assets_Turnover_Frequency – Efficiency of fixed asset usage to generate sales
- Net_Worth_Turnover_Rate_times – Sales to stockholder equity ratio (Efficiency measure)
- Operating_profit_per_person – Operating income per employee
- Allocation_rate_per_person – Fixed assets per employee
- Quick_Assets_to_Total_Assets – Ratio of quick assets to total assets
- Cash_to_Total_Assets – Ratio of cash to total assets
- Quick_Assets_to_Current_Liability – Quick assets / Current liabilities
- Cash_to_Current_Liability – Cash / Current liabilities
- Operating_Funds_to_Liability – Operating funds compared to liability

- Inventory_to_Working_Capital – Inventory / Working Capital
- Inventory_to_Current_Liability – Inventory / Current liabilities
- Long_term_Liability_to_Current_Assets – Long-term liabilities compared to current assets
- Retained_Earnings_to_Total_Assets – Retained earnings / Total assets
- Total_income_to_Total_expense – Total income / Total expense
- Total_expense_to_Assets – Total expense / Assets
- Current_Asset_Turnover_Rate – Current Assets / Sales (Efficiency of current asset usage)
- Quick_Asset_Turnover_Rate – Quick Assets / Sales
- Cash_Turnover_Rate – Cash / Sales (Liquidity management)
- Fixed_Assets_to_Assets – Fixed Assets compared to total assets
- Cash_Flow_to_Total_Assets – Cash flow in relation to total assets
- Cash_Flow_to_Liability – Cash flow available for liabilities
- CFO_to_Assets – Cash flow efficiency relative to assets
- Cash_Flow_to_Equity – Cash flow available to equity shareholders
- Current_Liability_to_Current_Assets – Ratio of liabilities to assets within a year
- Liability_Assets_Flag – 1 if total liability exceeds total assets, 0 otherwise
- Total_assets_to_GNP_price – Ratio of total assets to Gross National Product (GNP) price
- No_credit_Interval – Period when a company has no credit activity
- Degree_of_Financial_Leverage_DFL – Sensitivity of profitability to changes in capital structure
- Interest_Coverage_Ratio_Interest_expense_to_EBIT – Ability to cover interest expenses using EBIT
- Net_Income_Flag – 1 if net income is negative for the last two years, 0 otherwise
- Equity_to_Liability – Ratio of equity to liability
- Default – 1 if the company has defaulted (bankrupted), 0 otherwise

2.4 Statistical Summary

Figure 1 - Statistical Summary of the dataset

	count	mean	std	min	25%	50%	75%	max
Operating_Expense_Rate	2,058.00	2,052,388,835.76	3,252,623,690.29	0.00	0.00	0.00	4,110,000,000.00	9,980,000,000.00
Research_and_development_expense_rate	2,058.00	1,208,634,256.56	2,144,568,158.08	0.00	0.00	0.00	1,550,000,000.00	9,980,000,000.00
Cash_flow_rate	2,058.00	0.47	0.02	0.00	0.46	0.46	0.47	1.00
Interest_bearing_debt_interest_rate	2,058.00	11,130,223.52	90,425,949.04	0.00	0.00	0.00	0.00	990,000,000.00
Tax_rate_A	2,058.00	0.11	0.15	0.00	0.00	0.04	0.22	1.00
Cash_Flow_Per_Share	1,891.00	0.32	0.02	0.17	0.31	0.32	0.33	0.46
Per_Share_Net_profit_before_tax_Yuan_	2,058.00	0.18	0.03	0.00	0.17	0.18	0.19	0.79
Realized_Sales_Gross_Profit_Growth_Rate	2,058.00	0.02	0.02	0.00	0.02	0.02	0.02	1.00
Operating_Profit_Growth_Rate	2,058.00	0.85	0.00	0.74	0.85	0.85	0.85	1.00
Continuous_Net_Profit_Growth_Rate	2,058.00	0.22	0.01	0.00	0.22	0.22	0.22	0.23
Total_Asset_Growth_Rate	2,058.00	5,287,663,257.05	2,912,614,769.58	0.00	4,315,000,000.00	6,225,000,000.00	7,220,000,000.00	9,980,000,000.00
Net_Value_Growth_Rate	2,058.00	5,189,504.37	207,791,797.86	0.00	0.00	0.00	0.00	9,330,000,000.00
Total_Asset_Return_Growth_Rate_Ratio	2,058.00	0.26	0.00	0.25	0.26	0.26	0.26	0.36
Cash_Reinvestment_perc	2,058.00	0.38	0.03	0.03	0.37	0.38	0.39	1.00
Current_Ratio	2,058.00	1,336,248.80	60,619,173.20	0.00	0.01	0.01	0.01	2,750,000,000.00
Quick_Ratio	2,058.00	27,755,102.05	444,865,390.47	0.00	0.00	0.01	0.01	9,230,000,000.00
Interest_Expense_Ratio	2,058.00	0.63	0.01	0.53	0.63	0.63	0.63	0.81
Total_debt_to_Total_net_worth	2,037.00	10,714,285.73	269,696,017.59	0.00	0.00	0.01	0.01	9,940,000,000.00
Long_term_fund_suitability_ratio_A	2,058.00	0.01	0.03	0.00	0.01	0.01	0.01	1.00
Net_profit_before_tax_to_Paid_in_capital	2,058.00	0.18	0.03	0.00	0.17	0.17	0.18	0.79
Total_Asset_Turnover	2,058.00	0.13	0.10	0.00	0.06	0.10	0.17	0.92
Accounts_Receivable_Turnover	2,058.00	41,598,639.46	504,767,266.59	0.00	0.00	0.00	0.00	9,740,000,000.00
Average_Collection_Days	2,058.00	26,297,862.01	410,996,733.83	0.00	0.00	0.01	0.01	8,800,000,000.00
Inventory_Turnover_Rate_times	2,058.00	2,030,227,259.48	3,077,250,265.27	0.00	0.00	19,100,000.00	3,815,000,000.00	9,990,000,000.00
Fixed_Assets_Turnover_Frequency	2,058.00	1,230,897,959.18	2,649,288,936.44	0.00	0.00	0.00	0.01	9,990,000,000.00
Net_Worth_Turnover_Rate_times	2,058.00	0.04	0.04	0.01	0.02	0.03	0.04	1.00
Operating_profit_per_person	2,058.00	0.40	0.05	0.00	0.39	0.40	0.40	1.00
Allocation_rate_per_person	2,058.00	5,725,558.82	197,949,961.06	0.00	0.00	0.01	0.02	8,280,000,000.00
Quick_Assets_to_Total_Assets	2,058.00	0.34	0.21	0.00	0.17	0.31	0.48	0.99
Cash_to_Total_Assets	1,962.00	0.08	0.10	0.00	0.02	0.05	0.10	0.93
Quick_Assets_to_Current_Liability	2,058.00	11,904,761.91	312,292,270.93	0.00	0.00	0.01	0.01	8,820,000,000.00
Cash_to_Current_Liability	2,058.00	92,825,072.90	785,189,881.95	0.00	0.00	0.00	0.01	9,170,000,000.00
Operating_Funds_to_Liability	2,058.00	0.35	0.04	0.03	0.34	0.35	0.35	1.00

Inventory_to_Working_Capital	2,058.00	0.28	0.02	0.00	0.28	0.28	0.28	1.00
Inventory_to_Current_Liability	2,058.00	57,863,459.68	627,879,536.23	0.00	0.00	0.01	0.01	9,600,000,000.00
Long_term_Liability_to_Current_Assets	2,058.00	73,401,069.01	669,352,618.01	0.00	0.00	0.00	0.01	9,310,000,000.00
Retained_Earnings_to_Total_Assets	2,058.00	0.93	0.03	0.00	0.93	0.94	0.94	0.97
Total_income_to_Total_expense	2,058.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Total_expense_to_Assets	2,058.00	0.03	0.04	0.00	0.01	0.02	0.04	1.00
Current_Asset_Turnover_Rate	2,058.00	1,273,303,377.07	2,839,740,987.63	0.00	0.00	0.00	0.00	9,990,000,000.00
Quick_Asset_Turnover_Rate	2,058.00	2,571,767,687.08	3,453,544,121.67	0.00	0.00	0.00	5,790,000,000.00	10,000,000,000.00
Cash_Turnover_Rate	2,058.00	2,653,695,544.22	2,821,244,732.19	0.00	0.00	1,730,000,000.00	4,550,000,000.00	9,990,000,000.00
Fixed_Assets_to_Assets	2,058.00	4,042,760.23	183,400,553.09	0.00	0.10	0.21	0.42	8,320,000,000.00
Cash_Flow_to_Total_Assets	2,058.00	0.64	0.05	0.00	0.63	0.64	0.65	1.00
Cash_Flow_to_Liability	2,058.00	0.46	0.03	0.03	0.46	0.46	0.46	0.91
CFO_to_Assets	2,058.00	0.58	0.06	0.00	0.55	0.58	0.61	0.98
Cash_Flow_to_Equity	2,058.00	0.31	0.01	0.00	0.31	0.31	0.32	0.57
Current_Liability_to_Current_Assets	2,044.00	0.04	0.05	0.00	0.02	0.03	0.04	1.00
Total_assets_to_GNP_price	2,058.00	27,793,974.74	471,771,444.55	0.00	0.00	0.00	0.01	9,820,000,000.00
No_credit_Interval	2,058.00	0.62	0.01	0.41	0.62	0.62	0.62	0.96
Degree_of_Financial_Leverage_DFL	2,058.00	0.03	0.01	0.01	0.03	0.03	0.03	0.46
Interest_Coverage_Ratio_Interest_expense_to_EBIT	2,058.00	0.57	0.01	0.17	0.57	0.57	0.57	0.67
Equity_to_Liability	2,058.00	0.04	0.06	0.00	0.02	0.03	0.04	1.00
Default	2,058.00	0.11	0.31	0.00	0.00	0.00	0.00	1.00

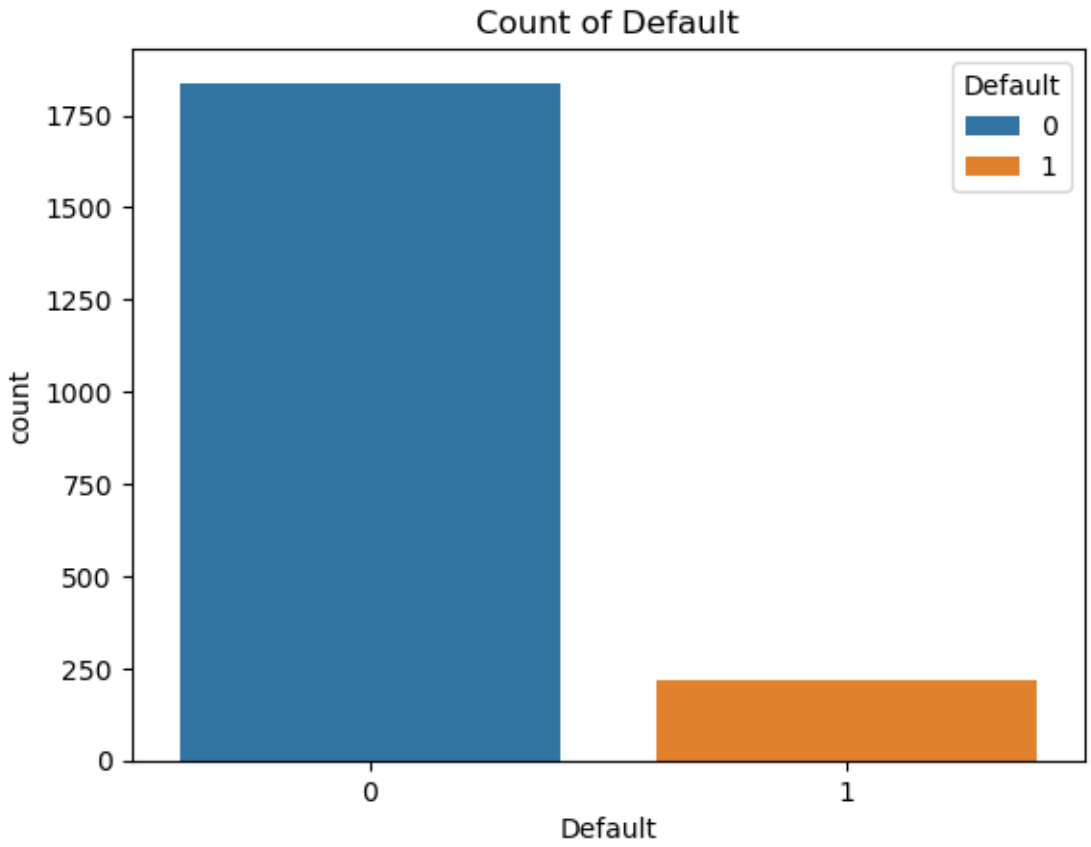
Observations:

- Operating Expense Rate & R&D Expense Rate have extremely high variability, with maximum values reaching 9.98 billion, indicating that some companies invest heavily in operations and innovation while others spend little to none.
- Interest-bearing Debt Interest Rate has a staggering maximum of 990 million, showing that certain firms rely heavily on debt financing, potentially increasing bankruptcy risks.
- Cash Flow Rate has a mean of ~0.47, implying that most firms maintain stable liquidity, but some have zero cash flow, which could signal financial instability.
- Net Value Growth Rate maxes out at 9.33 billion, yet the median is 0, meaning that while some firms experience massive equity growth, many have stagnated or lost value entirely.
- Total Asset Turnover is low on average ~0.13, suggesting that most firms struggle to generate revenue efficiently from their assets.
- Cash Flow to Total Assets has a relatively high mean ~0.64, implying that many companies have strong cash generation compared to their total asset base.

2.5 Univariate Analysis

2.5.1 Default

Figure 2 - Countplot of "Default"



Observations:

- The total number of non-defaulters is approximately 1850 companies.
- Approximately 10.69% of the companies in the dataset have defaulted.

2.5.2 Numerical Variables

Figure 3 - Boxplots of all Numerical Variables

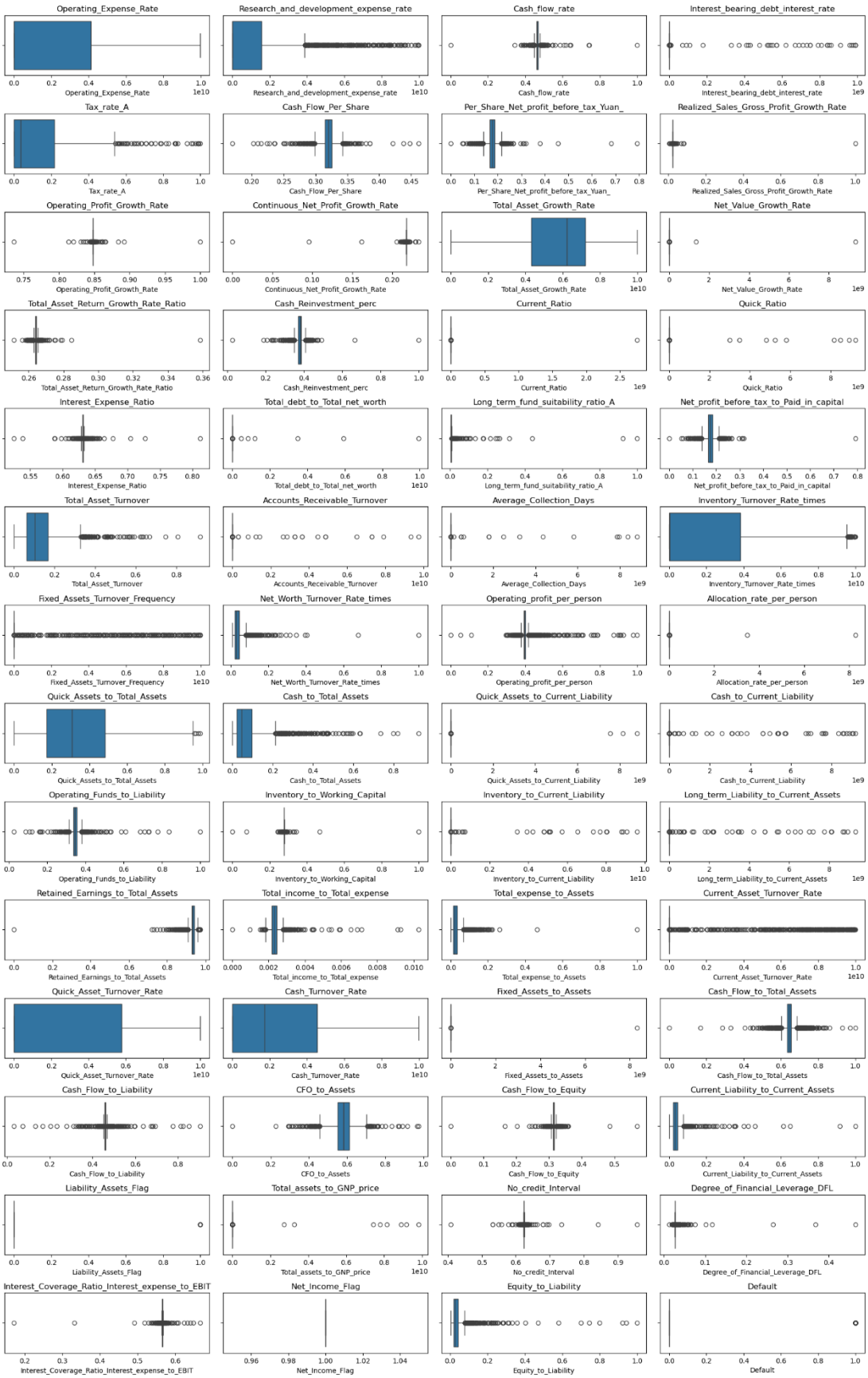
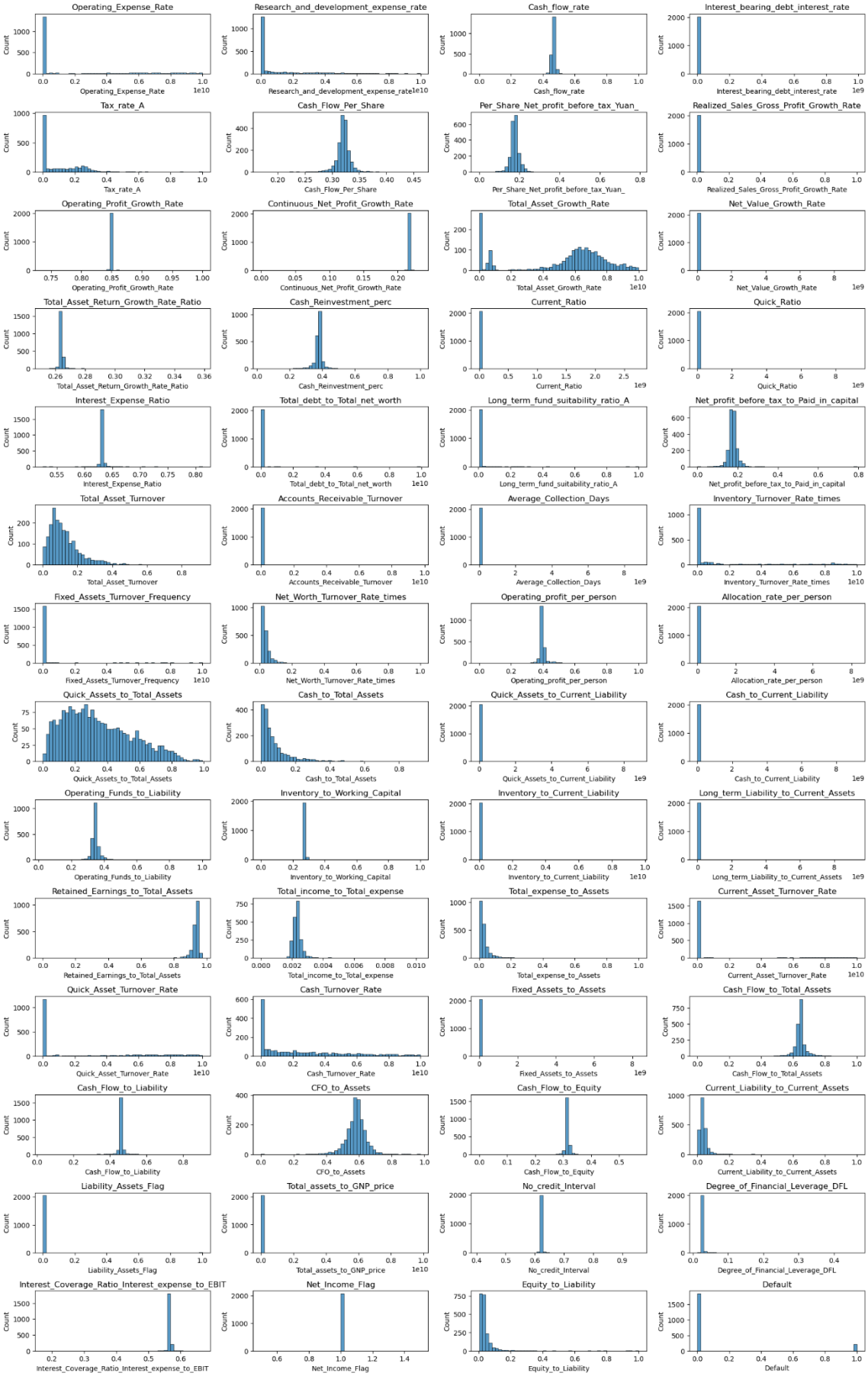


Figure 4 - Distribution plots of all Numerical Variables



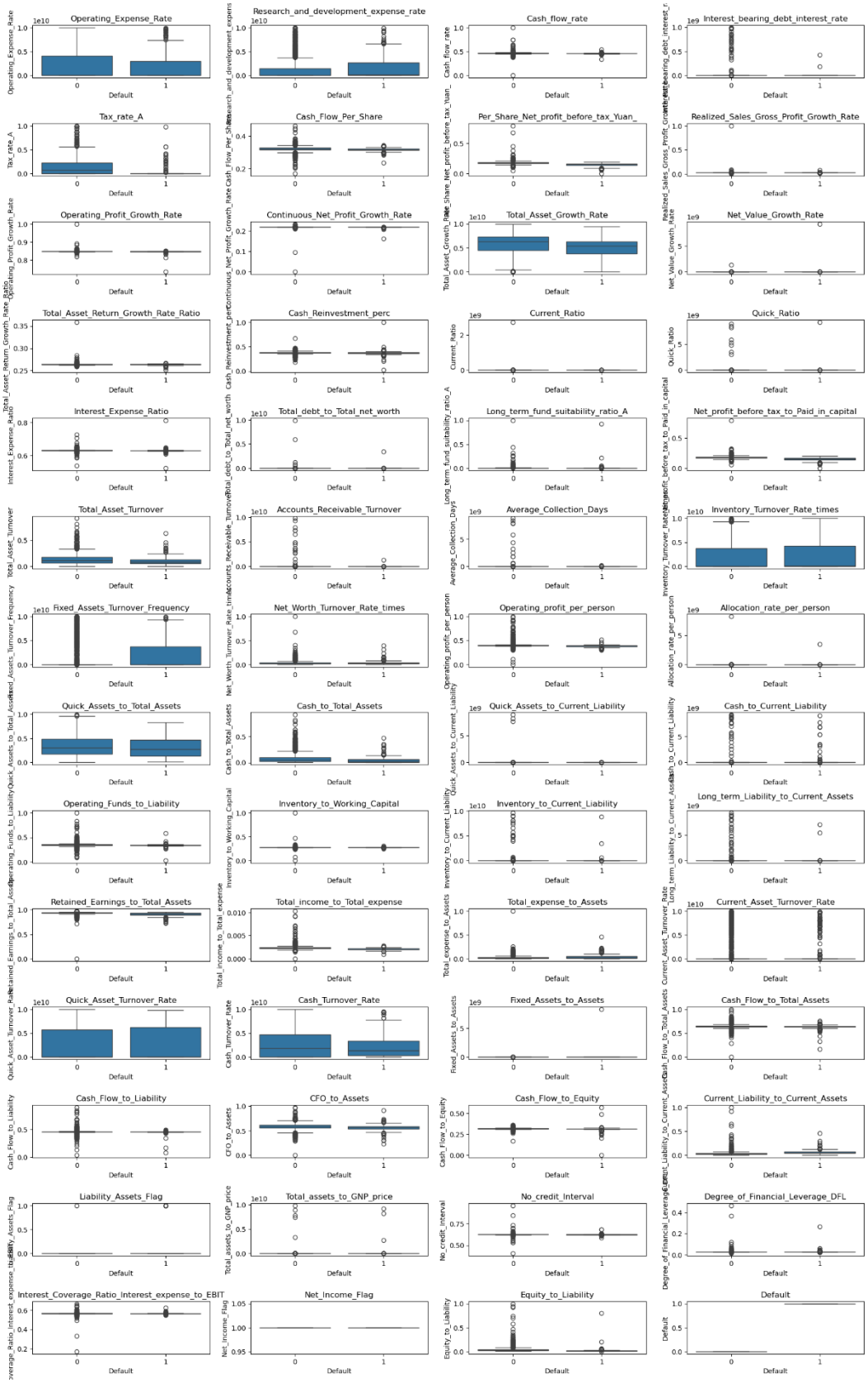
Observations:

- Operating expense ratios appear concentrated at the lower end, this suggests a general tendency towards cost efficiency across a significant portion of these companies.
- Operating Profit Growth rate is largely focused around point 0.85. Some observations can be seen at values above and below that values exhibiting some variability.
- Research and development expense ratios show a wide spread. This highlights the varying importance of innovation and future growth investment across different businesses.
- Cash flow rates exhibit a positive skew, this shows strong positive cash generation which is a fundamental indicator of financial health for many of these companies.
- Current ratios are mostly above 1, while quick ratios are slightly lower, this indicates that while most companies can cover their short-term liabilities with liquid assets, inventory plays a noticeable role in their current asset composition.
- Debt to equity ratios seem skewed towards lower values. a preference for equity financing over debt could point to a more conservative financial structure for many of these companies.

2.6 Bivariate Analysis

2.6.1 Numerical Variables

Figure 5 - Boxplot of all Numerical Variables vs Default

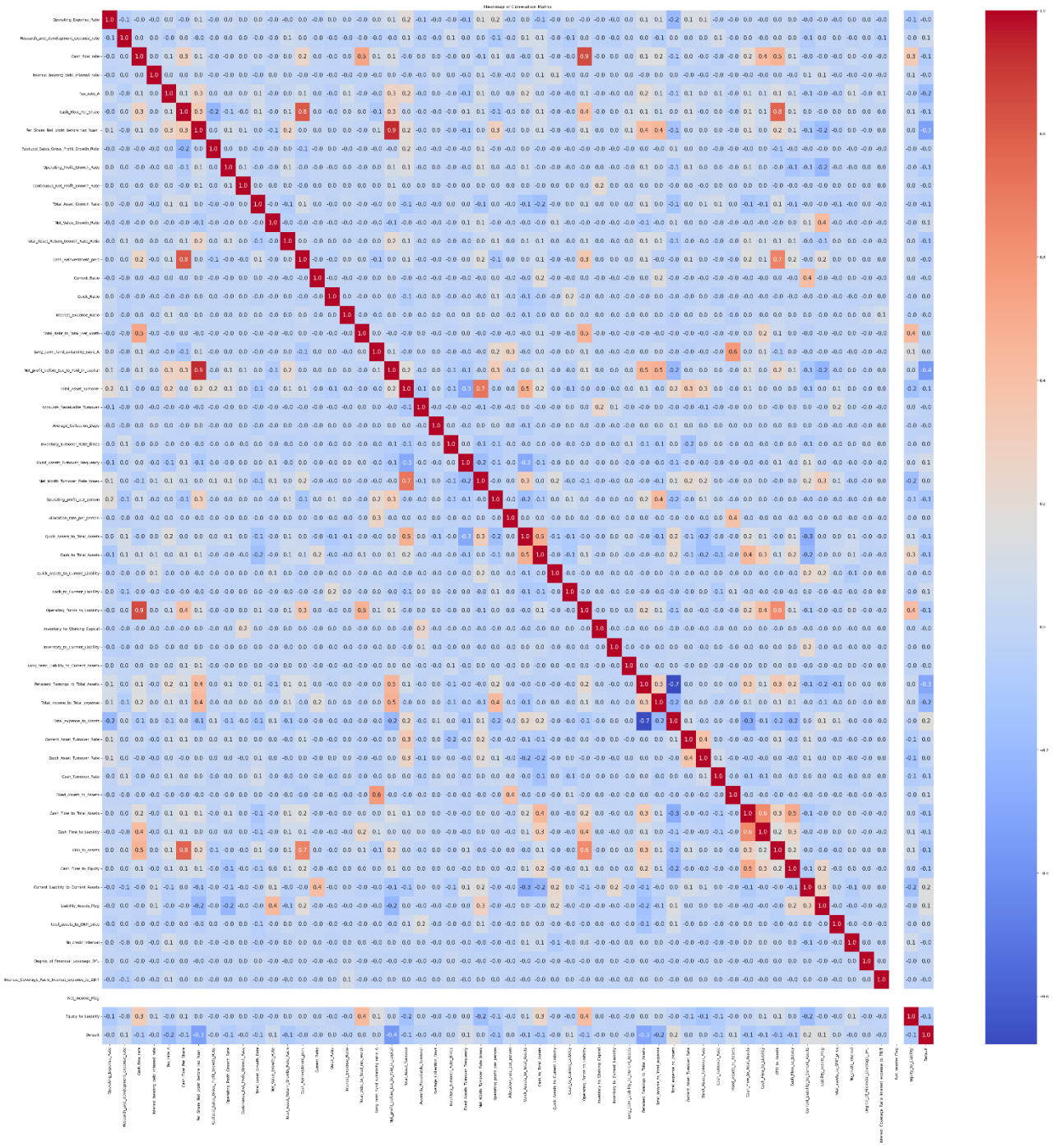


Observations:

- Companies that ended up defaulting seem to have generally carried a noticeably higher proportion of interest-bearing debt relative to their invested capital compared to those who didn't default.
- The non-defaulting companies on average show a considerably higher ratio of net profit before tax relative to their paid-in capital than the companies that defaulted.
- It shows that companies that did not default tended to have a higher long-term fund suitability ratio suggesting a better alignment of long-term funding with long-term assets.
- The non-defaulting group generally exhibits a higher quick ratio, indicating a greater ability to meet short-term obligations with their most liquid assets.
- Similar to the quick ratio, the non-defaulting companies tend to have a higher current ratio suggesting a stronger short-term liquidity position overall.
- Companies that defaulted appear to have a higher debt-to-equity ratio indicating greater reliance on debt financing compared to equity.
- The defaulting companies seem to have a higher degree of financial leverage, implying a greater sensitivity of their earnings per share to changes in operating income.

2.6.2 Correlation Matrix

Figure 6 - Correlation Matrix



Observations:

- Net Profit before tax to paid in capital and Per Share Net Profit before tax Yuan have the highest correlation at 0.9 which indicates a good relationship.
- Operating funds to liability and Cash flow rate have a high correlation score of 0.9 indicating there is a relationship.

- CFO to assets and Cash flow per share have a high correlation score is 0.8.
- Few other variables are positively correlated with each other; however, majority of the numerical variables don't show much correlation to each other.

DATA PRE-PROCESSING

3.1 Data Preparation

The following steps are taken so that the data is pre-processed so the modified dataset can used to build the various models to predict if the company will default on its debt repayment in the next 2 quarters.

- The columns "Net_Income_Flag" and "Liability_Assets_Flag" are dropped as they have very few unique values.
- There are many outliers present in the dataset, since these are all actual values and provided insight into the data the values will remain and not be treated.
- The data is split into Training and Test data in the ration of 75:25, meaning 25% of the dataset will become the testing data and 75% of the dataset will be used for training the different models.
- The training and testing dataset consists of missing values, these values are imputed based on the KNN Imputer which estimates missing values by averaging the k-nearest neighbors in a dataset based on feature similarity.
- The Dataset is then scaled since there is considerable variation in the values of each variable. Scaling is done so that all variables contribute equally to a building a model and prevents any bias towards larger numerical variables.

Figure 7 - Snippet of the Scaled Training Data

	Operating_Expense_Rate	Research_and_development_expense_rate	Cash_flow_rate	Interest_bearing_debt_interest_rate	Tax_rate_A	Cash_Flow_Per_Share
0	-0.633296	-0.396806	-0.132455	-0.128462	-0.754347	0.088170
1	-0.633296	-0.561672	-0.934352	-0.128462	-0.754347	-1.224514
2	-0.633296	0.361946	-0.290335	-0.128462	0.061964	-0.409659
3	-0.633296	-0.561672	-0.179548	-0.128462	-0.754347	-0.077773
4	-0.633296	-0.561672	-0.123892	-0.128462	-0.754347	-0.168422

MODEL BUILDING

4.1 Logistic Regression Model

Logistic regression is a classification algorithm used to predict binary outcomes (0 or 1). It models the probability of a default class using a sigmoid function applied to a linear combination of input features. For the current dataset, A constant is added to ensure that the intercept is included in the regression model to allow for a more accurate fit.

Figure 8 - Logistic Regression Model

Find

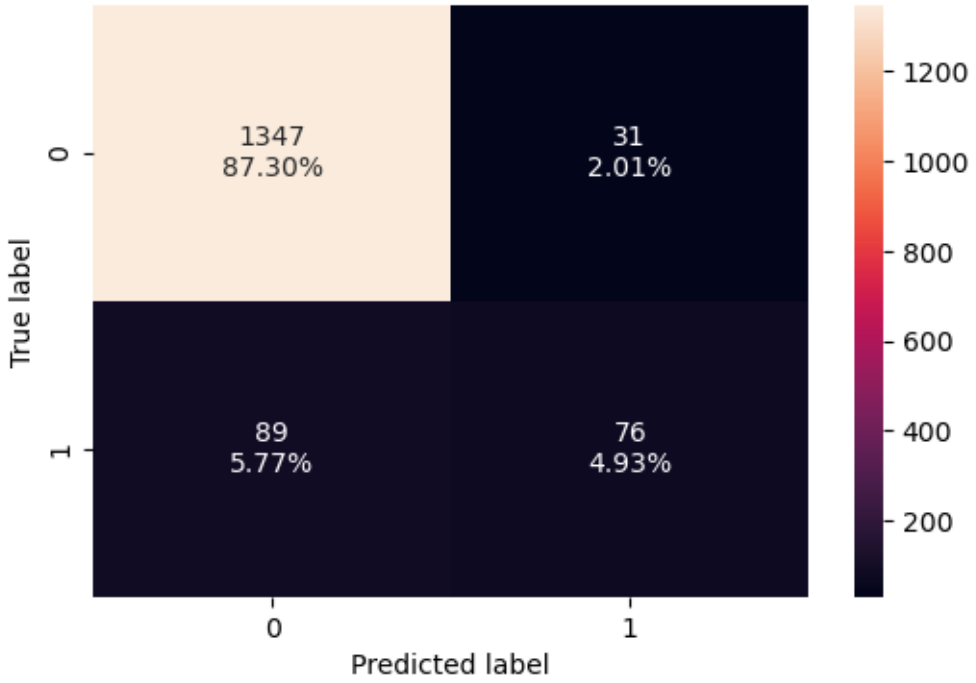
Logit Regression Results

Dep. Variable:	Default	No. Observations:	1543
Model:	Logit	Df Residuals:	1489
Method:	MLE	Df Model:	53
Date:	Sat, 10 May 2025	Pseudo R-squ.:	0.4297
Time:	16:59:16	Log-Likelihood:	-299.26
converged:	False	LL-Null:	-524.71
Covariance Type:	nonrobust	LLR p-value:	1.764e-64

	coef	std err	z	P>|z|	[0.025	0.975]
const	-7.4685	2410.787	-0.003	0.998	-4732.523	4717.586
Operating_Expense_Rate	0.2077	0.121	1.713	0.087	-0.030	0.445
Research_and_development_expense_rate	0.3556	0.104	3.433	0.001	0.153	0.559
Cash_flow_rate	-0.1837	1.016	-0.181	0.857	-2.175	1.808
Interest_bearing_debt_interest_rate	0.1755	0.151	1.163	0.245	-0.120	0.471
Tax_rate_A	-0.2580	0.174	-1.481	0.139	-0.599	0.083
Cash_Flow_Per_Share	-0.3533	0.281	-1.260	0.208	-0.903	0.196
Per_Share_Net_profit_before_tax_Yuan_	0.2518	1.276	0.197	0.844	-2.249	2.752
Realized_Sales_Gross_Profit_Growth_Rate	0.1012	0.118	0.859	0.390	-0.130	0.332
Operating_Profit_Growth_Rate	-0.1546	0.267	-0.579	0.563	-0.678	0.369
Continuous_Net_Profit_Growth_Rate	0.1736	0.132	1.317	0.188	-0.085	0.432
Total_Asset_Growth_Rate	-0.0640	0.131	-0.487	0.626	-0.321	0.193
Net_Value_Growth_Rate	0.5177	3097.843	0.000	1.000	-6071.142	6072.178
Total_Asset_Return_Growth_Rate_Ratio	-0.3299	0.361	-0.915	0.360	-1.037	0.377
Cash_Reinvestment_perc	0.1700	0.346	0.491	0.624	-0.509	0.849
Current_Ratio	-1.6114	0.925	-1.742	0.081	-3.424	0.201
Quick_Ratio	-2.7355	2.57e+04	-0.000	1.000	-5.05e+04	5.05e+04
Interest_Expense_Ratio	0.0197	0.065	0.303	0.762	-0.107	0.147
Total_debt_to_Total_net_worth	1.9035	0.623	3.058	0.002	0.683	3.124
Long_term_fund_suitability_ratio_A	0.1675	0.223	0.751	0.452	-0.269	0.604
Net_profit_before_tax_to_Paid_in_capital	-1.0834	1.179	-0.919	0.358	-3.394	1.227
Total_Asset_Turnover	-0.2122	0.319	-0.666	0.506	-0.837	0.413
Accounts_Receivable_Turnover	-1.0019	0.642	-1.560	0.119	-2.261	0.257
Average_Collection_Days	-15.1938	2.49e+04	-0.001	1.000	-4.89e+04	4.88e+04
Inventory_Turnover_Rate_times	-0.0490	0.117	-0.420	0.675	-0.278	0.180
Fixed_Assets_Turnover_Frequency	0.1775	0.106	1.678	0.093	-0.030	0.385
Net_Worth_Turnover_Rate_times	-0.2559	0.211	-1.212	0.225	-0.670	0.158
Operating_profit_per_person	0.0505	0.195	0.259	0.796	-0.331	0.432
Allocation_rate_per_person	-80.4893	153.634	-0.524	0.600	-381.606	220.628

The model is trained on the training dataset using the logistic regression algorithm from the statsmodels library of Python.

Figure 9 - Confustion Matrix - Logistic Regression Model on Training Data

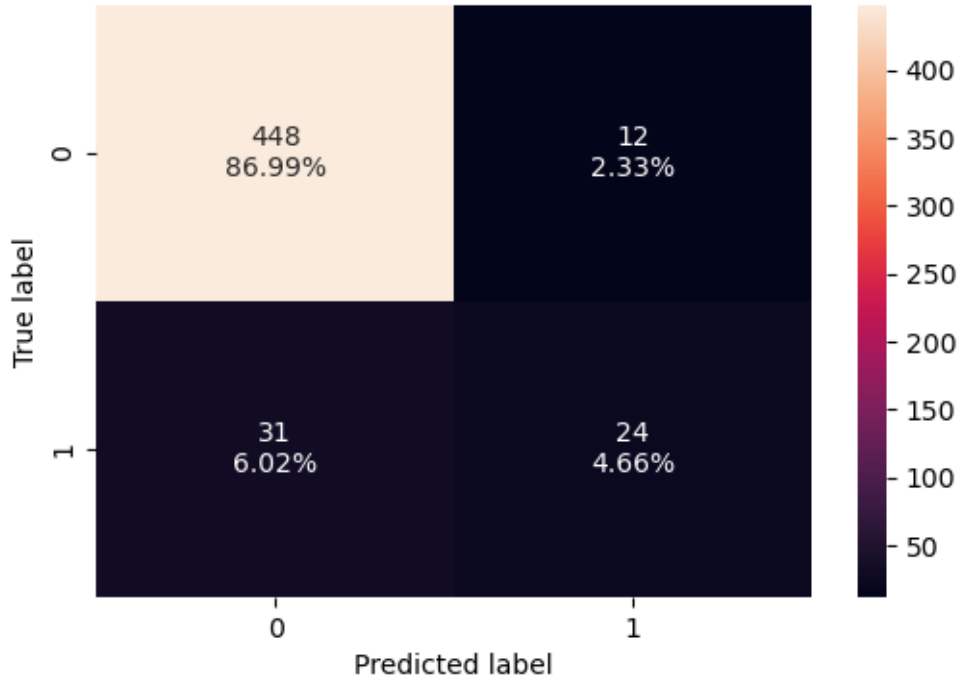


	Accuracy	Recall	Precision	F1
0	0.922229	0.460606	0.71028	0.558824

The model has correctly predicted 92.23% defaulters and non-defaulter companies of the data on the training set. 7.77% of the data is incorrectly predicted. There is a higher chance of the model predicting a company will not default when it will default.

An accuracy score of 92.22% is observed which indicates a good overall classification rate. However, the recall score of 46.06% is significantly lower than the accuracy, suggesting the model misses a substantial portion of actual defaults. The precision for the default class is 71.03%, meaning that when the model predicts a default, it's correct a good portion of the time, but the lower recall indicates many defaults are not being caught. This imbalance between precision and recall, along with the lower F1-score of 0.558824 shows that the logistic regression model is better at identifying non-defaults but needs improvement in accurately predicting defaults.

Figure 10 - Confustion Matrix - Logistic Regression on Test Data



	Accuracy	Recall	Precision	F1
0	0.916505	0.436364	0.666667	0.527473

The model correctly predicts a high percentage of 91.65% of non-defaulters. There is a notable number of false negatives 6.02%, where defaults are missed, while the false positive rate is relatively low around 2.33%. The model performs slightly worse than the training set.

An accuracy score of 91.65% is observed but the recall score is low at 43.64%, indicating it misses a significant portion of actual defaults. While the precision for defaults is reasonable at 66.67%, meaning when it predicts a default it's often correct, the low recall and resulting F1-score of 0.5275 suggest the model's ability to identify defaulting companies is limited. This reinforces the observation from the confusion matrix that the model is better at predicting non-defaults than catching defaults in the unseen data.

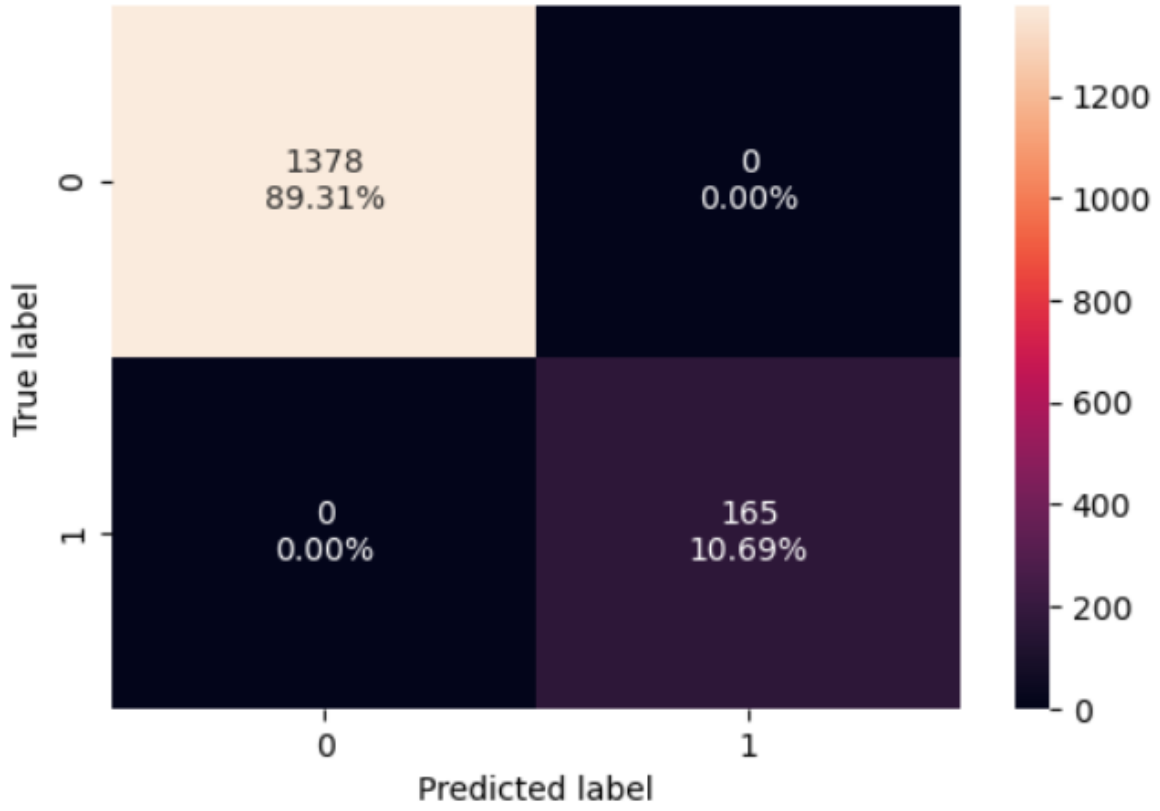
4.2 Random Forest Model

Random Forest is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. It works by randomly selecting subsets of data and features to train individual trees, then aggregates their predictions for a final result. The model is highly effective for classification and regression tasks, handling large datasets with

complex relationships. It offers strong resistance to noise and missing data while maintaining high predictive performance. Random Forest is widely used in finance, healthcare, and marketing due to its robustness and interpretability

The Random Forest Model is built using the scikit-learn library on Python

Figure 11 - Confusion Matrix - Random Forest Model on Training Data

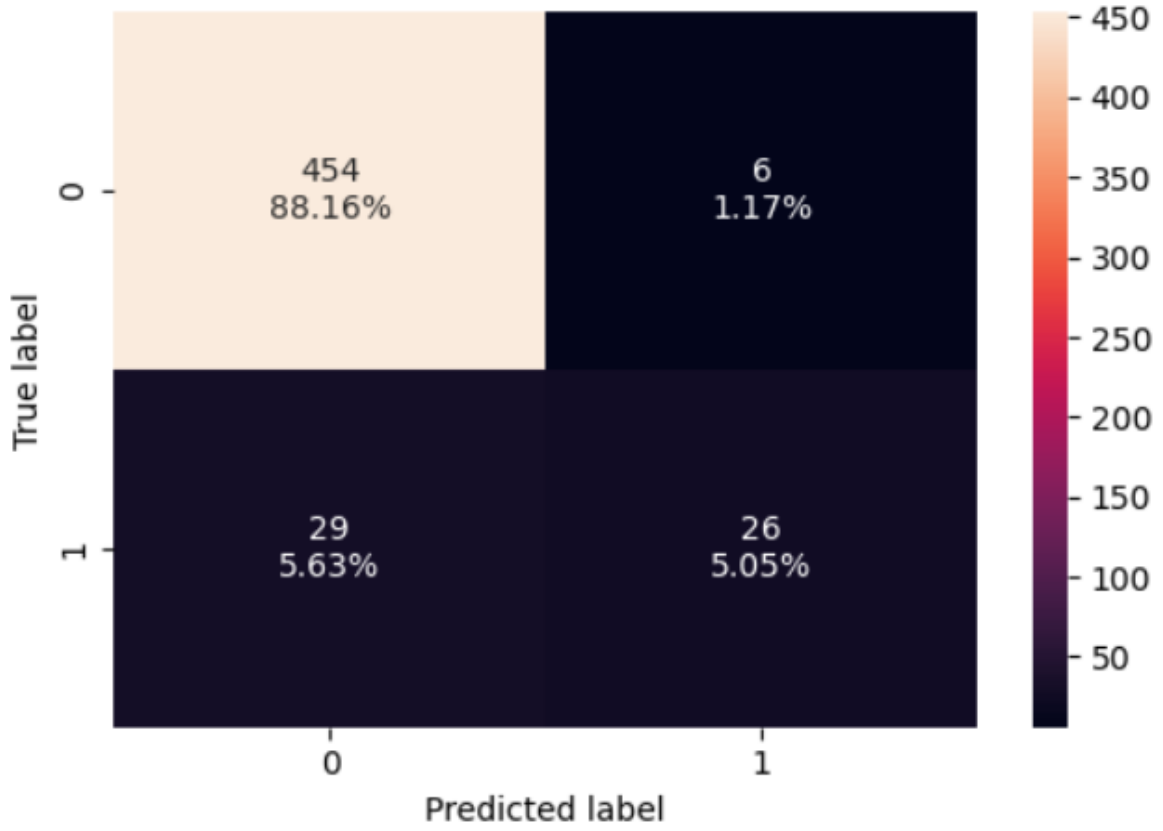


	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

The model performs extremely well on the training data with perfect prediction rate of 89.31% true positives and 10.69% of true negatives.

The accuracy, recall, precision and F1 scores are all perfect indicating 100% prediction rate on the training set.

Figure 12 - Confusion Matrix - Random Forest Model on Test Data



	Accuracy	Recall	Precision	F1
0	0.932039	0.472727	0.8125	0.597701

The model correctly predicted a large portion of the non-defaulting companies 88.16% of the total. It also correctly identified some of the companies that actually defaulted 5.05% of the total. However, there were some errors: 1.17% non-defaulting companies were incorrectly predicted and 5.63% of the total that actually defaulted were incorrectly predicted as non-defaults.

With an accuracy score of 93.20% the model indicates a good overall performance in classifying companies. The recall for the default class is relatively low at 0.4727, meaning the model only identifies about 47% of the actual defaulting companies in the test set. The precision for the default class is quite high at 0.8125. The F1-score of 0.5977, which balances recall and precision, indicates a moderate performance in correctly identifying defaults,

highlighting a trade-off where the model is very reliable when it predicts a default, but it misses a significant number of actual defaults.

MODEL PERFORMANCE IMPROVEMENT

5.1 Optimised Logistic Regression Model

Figure 13 - Optimised Logistic Regression Model

Logit Regression Results

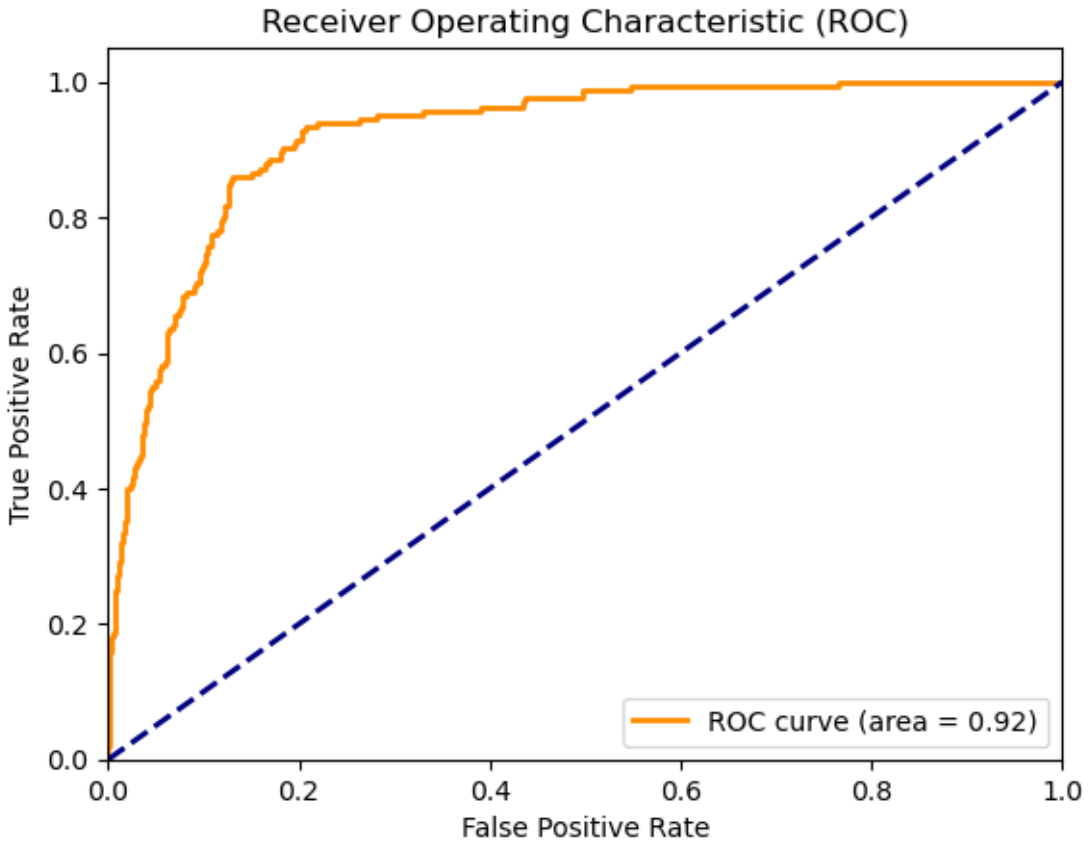
Dep. Variable:	Default	No. Observations:	1543
Model:	Logit	Df Residuals:	1496
Method:	MLE	Df Model:	46
Date:	Sat, 10 May 2025	Pseudo R-squ.:	0.4130
Time:	17:11:18	Log-Likelihood:	-308.02
converged:	True	LL-Null:	-524.71
Covariance Type:	nonrobust	LLR p-value:	1.893e-64

	coef	std err	z	P>|z|	[0.025	0.975]
const	-14.6665	nan	nan	nan	nan	nan
Operating_Expense_Rate	0.1757	0.118	1.495	0.135	-0.055	0.406
Research_and_development_expense_rate	0.3704	0.100	3.712	0.000	0.175	0.566
Interest_bearing_debt_interest_rate	0.1968	0.153	1.289	0.197	-0.102	0.496
Tax_rate_A	-0.3632	0.180	-2.021	0.043	-0.715	-0.011
Cash_Flow_Per_Share	-0.1428	0.137	-1.040	0.299	-0.412	0.126
Realized_Sales_Gross_Profit_Growth_Rate	0.1213	0.117	1.033	0.302	-0.109	0.351
Operating_Profit_Growth_Rate	-0.2182	0.285	-0.767	0.443	-0.776	0.340
Continuous_Net_Profit_Growth_Rate	0.1526	0.123	1.245	0.213	-0.088	0.393
Total_Asset_Growth_Rate	-0.0620	0.126	-0.491	0.623	-0.309	0.185
Net_Value_Growth_Rate	2.2725	nan	nan	nan	nan	nan
Total_Asset_Return_Growth_Rate_Ratio	-0.6885	0.366	-1.879	0.060	-1.407	0.030
Current_Ratio	-1.9831	0.659	-3.009	0.003	-3.275	-0.691
Quick_Ratio	-7.8291	5.18e+07	-1.51e-07	1.000	-1.02e+08	1.02e+08
Interest_Expense_Ratio	0.0278	0.065	0.428	0.669	-0.100	0.155
Total_debt_to_Total_net_worth	2.8741	0.578	4.971	0.000	1.741	4.007
Long_term_fund_suitability_ratio_A	0.1444	0.195	0.740	0.459	-0.238	0.527
Accounts_Receivable_Turnover	-1.0561	0.618	-1.709	0.088	-2.268	0.155
Average_Collection_Days	-72.6735	nan	nan	nan	nan	nan
Inventory_Turnover_Rate_times	-0.0490	0.114	-0.430	0.667	-0.272	0.174
Fixed_Assets_Turnover_Frequency	0.1547	0.104	1.485	0.138	-0.049	0.359
Net_Worth_Turnover_Rate_times	-0.1966	0.128	-1.536	0.124	-0.448	0.054
Operating_profit_per_person	0.0566	0.187	0.302	0.763	-0.311	0.424
Allocation_rate_per_person	-188.2483	2.26e+04	-0.008	0.993	-4.45e+04	4.42e+04
Quick_Assets_to_Total_Assets	0.0965	0.163	0.593	0.553	-0.222	0.415
Cash_to_Total_Assets	-0.3878	0.212	-1.833	0.067	-0.802	0.027
Quick_Assets_to_Current_Liability	-11.4719	nan	nan	nan	nan	nan
Cash_to_Current_Liability	0.0738	0.075	0.986	0.324	-0.073	0.221
Inventory_to_Working_Capital	-0.1546	0.144	-1.071	0.284	-0.438	0.128

The Logistic Regression model is optimized by first dealing with multicollinearity. Multicollinearity occurs when independent variables are highly correlated, causing instability in coefficient estimates. Variance Inflation Factor is used to detect such correlations.

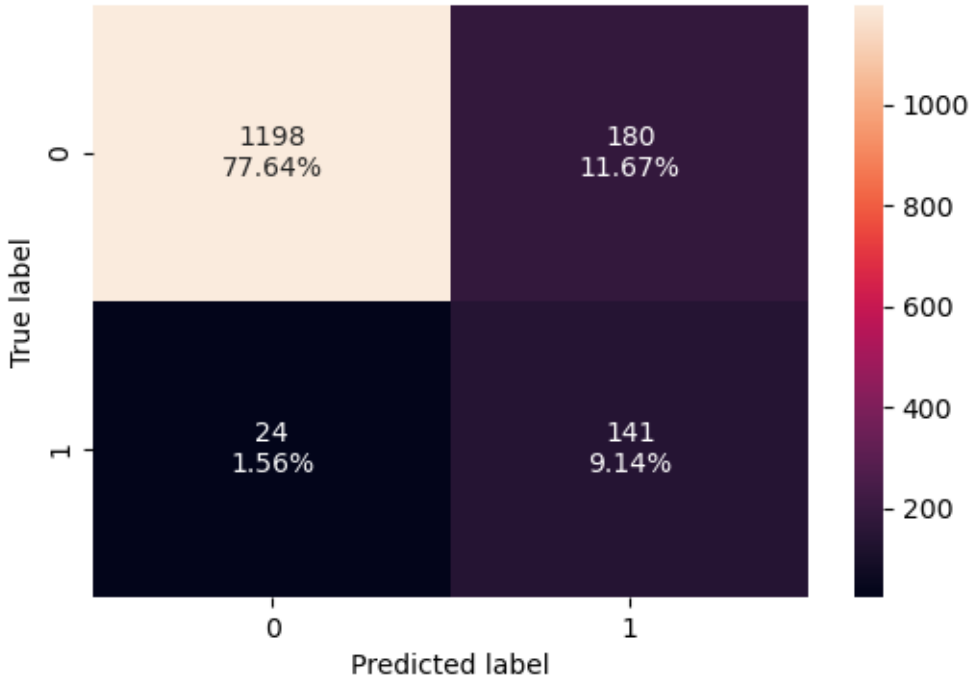
For this model any variable with a VIF score > =5 will be removed and the model built on the rest of the data. 'Cash_flow_rate', 'Per_Share_Net_profit_before_tax_Yuan_', 'Cash_Reinvestment_perc','Net_profit_before_tax_to_Paid_in_capital', 'Total_Asset_Turnover', 'Operating_Funds_to_Liability', 'CFO_to_Assets'. These variables are removed from the data.

Figure 14 - ROC Curve



Logistic Regression predicts probabilities, but choosing the optimal threshold is crucial for classification accuracy. The Receiver Operating Characteristic Curve visualizes how well the model distinguishes between classes. From the above plot an optimal threshold score of 0.92 for the ROC curve is determined.

Figure 15 - Confusion Matrix - Optimised Logistic Regression Model on Training Data

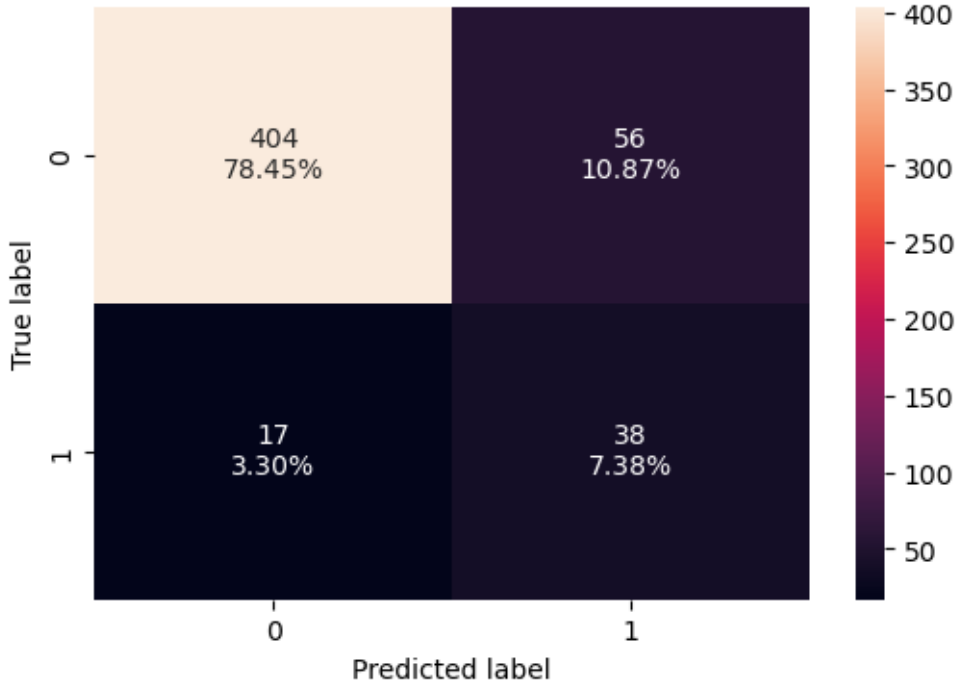


	Accuracy	Recall	Precision	F1
0	0.86779	0.854545	0.439252	0.580247

The optimized logistic regression model correctly predicts 77.64% of the defaulters and 9.14% of the defaulters. 1.56% are false negatives, indicating incorrectly predicted non-defaulters and 11.67% are False positives indicating incorrectly predicted defaulters.

The model shows an accuracy score of 86.77%. It demonstrates a substantial improvement in identifying defaulting companies, achieving a high recall of 0.8545. However, this increased ability to catch defaults comes with a lower precision of 0.4393, meaning more non-defaulting companies are incorrectly flagged. The F1-score of 0.5802, which balances recall and precision, indicates an enhanced ability to predict defaults compared to the initial model. This optimization has shifted the model towards being more sensitive to defaults, potentially at the expense of higher false positive rates.

Figure 16 - Confusion Matrix - Optimised Logistic Regression Model on Test Data



	Accuracy	Recall	Precision	F1
0	0.858252	0.690909	0.404255	0.510067

The optimized logistic regression model correctly predicts 78.45% of the defaulters and 7.38% of the defaulters when performed on the test data. 3.30% are false negatives, indicating incorrectly predicted non-defaulters and 10.87% are False positives indicating incorrectly predicted defaulters.

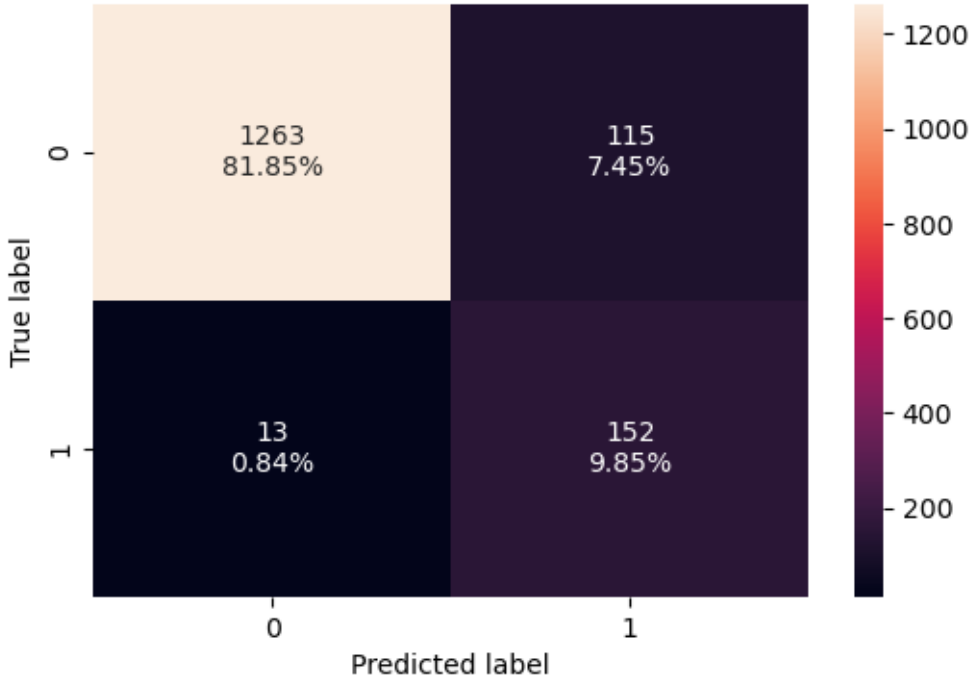
The accuracy score is 85.83%, indicating the overall correctness on test data. The recall for the default class is 0.6909, suggesting the model correctly identifies a decent portion of actual defaults in the test set. However, the precision is lower at 0.4043, meaning a significant number of companies predicted as defaults are actually not. The F1-score of 0.5101, balancing recall and precision, shows a moderate performance in identifying defaults on new data. Compared to the training data, there's a slight drop in recall and precision on the test set, which is expected, but the model still shows a better ability to identify defaults than the initial logistic regression model.

5.2 Optimised Random Forest Model

The Random Forest model is optimized using Hyperparameter tuning. It optimizes a Random Forest model by selecting the best configuration for parameters like number of trees, maximum depth, and feature selection. It can be done using Grid Search or Random Search to systematically test combinations for the highest accuracy. Cross-validation ensures robustness by evaluating different parameter sets across multiple data splits.

The current model will be optimized based on Grid Search. The best parameters identified are maximum depth = 5, min samples leaf = 7, min samples split = 2, no of estimators =200

Figure 17 - Confusion Matrix - Optimised Random Forest Model on Training Data



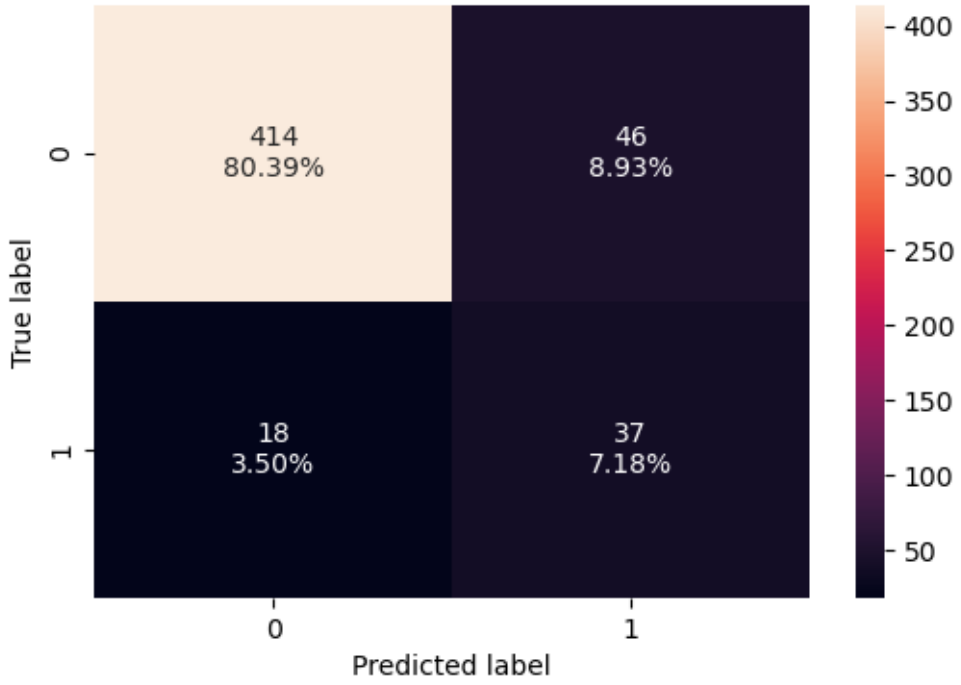
	Accuracy	Recall	Precision	F1
0	0.917045	0.921212	0.569288	0.703704

The optimized model predicts 81.85% of the non-defaulters accurately and predicts 9.85% of the defaulters. It predicts 0.84% of the defaulters incorrectly while predicting 7.45% of the non-defaulters incorrectly.

The model shows an accuracy score of 91.70%, suggesting good overall performance on the training set. The recall for the default class is excellent at 0.9212, indicating it's capturing almost all the actual defaults in the training data. However, the precision is lower at 0.5693,

meaning that when the model predicts a default, it's correct a little over half the time. The F1-score of 0.7037, which balances recall and precision, shows a solid performance in identifying defaults on the training data. This optimized Random Forest seems to have addressed the earlier overfitting issue by effectively identifying defaults in the training set, though with a moderate rate of false positives.

Figure 18 - Confusion Matrix - Optimised Random Forest Model on Test Data



	Accuracy	Recall	Precision	F1
0	0.875728	0.672727	0.445783	0.536232

The optimized model when performed on the test data predicts 80.39% of the non-defaulters accurately and predicts 7.18% of the defaulters. It predicts 3.5% of the defaulters incorrectly while predicting 8.93% of the non-defaulters incorrectly.

The model shows an accuracy score of 87.57%, indicating a good overall classification rate on unseen data. The recall for the default class is 0.6727, suggesting it's identifying a fair portion of the actual defaults in the test set. The precision is 0.4458, meaning that when the model predicts a default, it's correct less than half the time, leading to a number of false positives. The F1-score of 0.5362, balancing recall and precision, provides a moderate

measure of the model's ability to identify defaults on new data. While the recall is decent, the lower precision indicates a need to potentially refine the model to reduce false alarms.

The optimized Random forest model performs better compared to the random forest model without optimizations.

MODEL PERFORMANCE COMPARISION AND FINAL MODEL SELECTION

Figure 19 - Training Performance Comparison

Training performance comparison:

	Logistic Regression	Tuned Logistic Regression	Random Forest	Tuned Random Forest
Accuracy	0.922229	0.867790	1.0	0.917045
Recall	0.460606	0.854545	1.0	0.921212
Precision	0.710280	0.439252	1.0	0.569288
F1	0.558824	0.580247	1.0	0.703704

Figure 20 - Test Performance Comparison

Testing performance comparison:

	Logistic Regression	Tuned Logistic Regression	Random Forest	Tuned Random Forest
Accuracy	0.916505	0.858252	0.932039	0.875728
Recall	0.436364	0.690909	0.472727	0.672727
Precision	0.666667	0.404255	0.812500	0.445783
F1	0.527473	0.510067	0.597701	0.536232

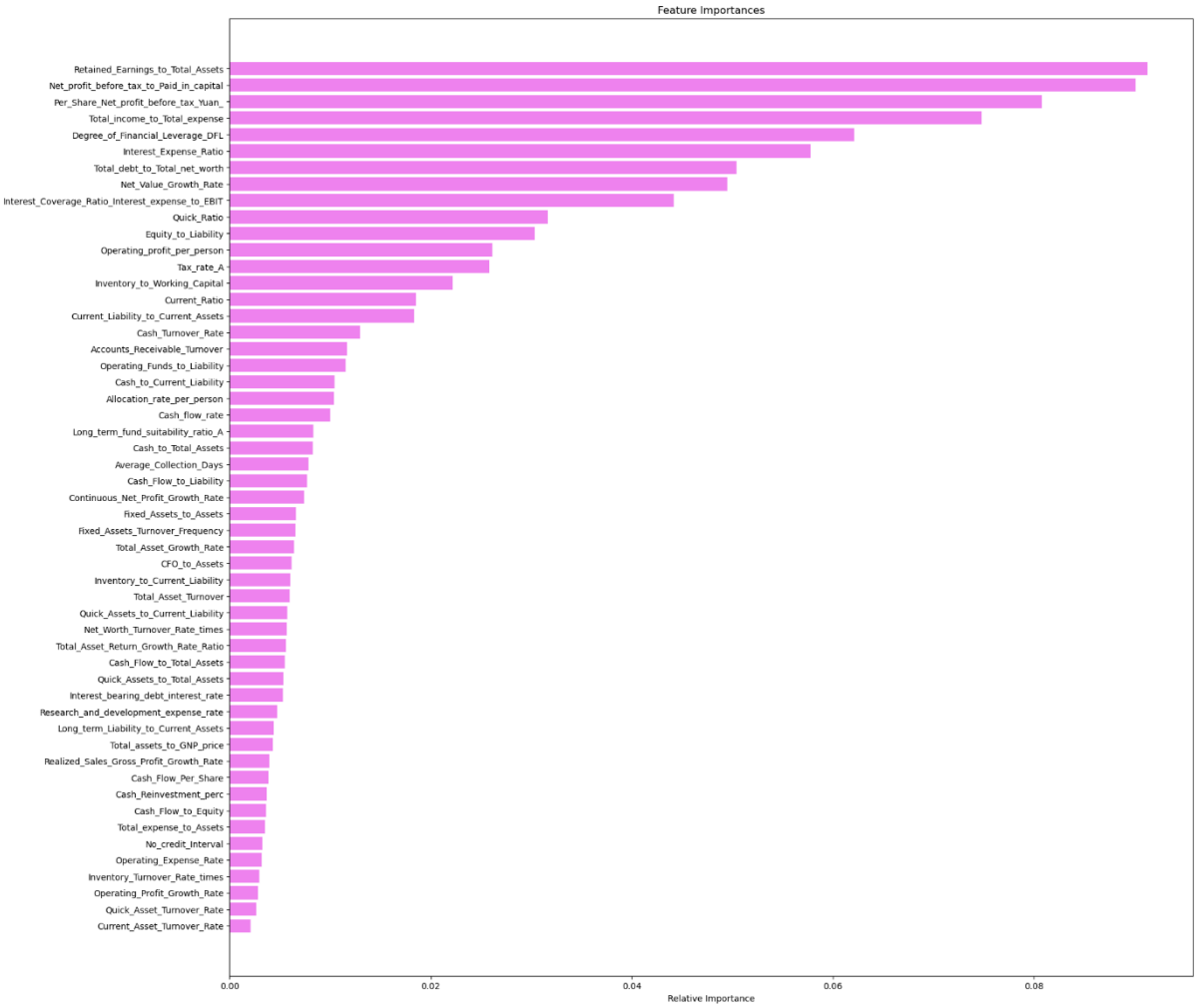
Observations:

- Random Forest has perfect training scores which suggests potential overfitting. However, its testing Accuracy is the highest, and it has a strong Precision and F1-score, meaning it generalizes relatively well.
- Tuned Random Forest has slightly lower training scores compared to its untuned counterpart, but it balances Recall and Precision well in testing, though it has a lower testing Accuracy.

- Logistic Regression has good testing Accuracy and maintains a good balance between Recall and Precision. However, its training Recall is relatively low.
- Tuned Logistic Regression improves training Recall but sacrifices Precision. Its testing Accuracy is on the lower side, and the F1-score suggests less overall robustness compared to the Random Forest models.

Since the aim is to predict if a company is a defaulter, The model needs to have a good balance of Recall and Precision. It is best to choose the Tuned Random Forest Model as the most suitable model to predict whether a given company will default on its debt repayments in the next two quarters.

Figure 21 - Feature Importances



- Retained Earnings to Total Assets stands out as the most influential factor in the model's predictions, having a significantly higher relative importance compared to all other features.

- The variables such as Net Profit Before Tax to Paid-in Capital, Per Share Net Profit Before Tax, and Total Income to Total Expense, also rank high in importance, suggesting profitability is a key indicator of default risk.
- The variables like asset turnover, cash flow, and working capital, have relatively lower importance scores, implying they contribute less individually to the model's predictive power compared to the top-ranking features.

ACTIONABLE INSIGHTS AND RECOMMENDATIONS

These are some of the key actionable insights and recommendations for the organization.

- Strengthen financial foundation by enhancing equity through fresh funding or retained earnings to lower dependence on liabilities and enhance financial stability.
- Refine debt strategy by negotiating improved repayment terms, such as extended deadlines or reduced interest rates, to ease financial strain and sustain liquidity.
- Enhance cost control by focusing on trimming non-essential expenditures while preserving operational efficiency to maintain financial health.
- Expand revenue avenues by exploring new market opportunities, refine product offerings, or strengthen marketing initiatives to create a more resilient income flow.
- Safeguard liquidity position by Optimizing cash flow strategies, defer non-urgent costs, and streamline working capital management to ensure stable short-term financial health.
- Allocate resources wisely and leverage cost-efficient methods like industry partnerships or grants to drive sustainable growth.
- Establish a proactive monitoring framework to continuously assess financial indicators and preemptively address potential threats.
- Boost predictive model accuracy by broadening data exposure, enhance feature optimization.