Finance and Retail Analytics

(Part 1)

Issac Abraham

Table of Contents

Executive summary.....3
DATA DICTIONARY.....3
Data Ingestion.....4
Sample of the dataset.....5
Checking for Duplicate Rows.....5
Null Value Check 5
Imputing values.....5
Dropping Irrelevant Columns.....6
Top Correlated Features with Default.....6
Heatmap of top 10 corelated features.....7
Univariate Analysis: Distribution of Top Correlated Features with Default.....7
Bivariate Analysis: Boxplot of Top Correlated Features by Default.....9
Data Splitting for Model Training and Testing.....11
Logistic Regression Model Building Approach and Performance.....11
Random Forest Model Building Approach and Performance.....12
Linear Discriminant Analysis (LDA) Model Performance.....13
Performance Comparison of Classification Models.....14
Insights.....14
Recommendations.....15

List of Figures

Fig 1: Sample of the dataset.....5
Fig 2 : Boxplot.....6
Fig 3: Heatmap of top 10 corelated features.....7
Fig 4: Univariate Analysis Distribution plots.....8
Fig 5 : Bivariate Analysis Boxplots.....9
Fig 6: ROC Curve Logistic Regression.....12
Fig 7: ROC Curve Random Forest.....13
Fig 8: ROC Curve Linear Discriminant Analysis.....14

List of Tables

Table 1:Train test split.....11
Table 2: Performance Comparison of Classification Models.....14

Default Prediction and Financial Health Analysis: A Comprehensive Approach

Executive summary:

In the world of business, defaulting on debts can spell trouble. It lowers a company's credit rating, making it harder to borrow money and potentially increasing interest rates. Investors want to put their money into companies that can handle their finances well, grow steadily, and adapt to changes. This report dives into how we can predict and understand default risk, helping everyone involved make smarter decisions in today's fast-paced business world

DATA DICTIONARY:

1. **Co_Code & Co_Name:** Identification variables for the company.
2. **Operating_Expense_Rate:** Operating Expenses to Net Sales ratio.
3. **Research_and_development_expense_rate:** R&D Expenses to Net Sales ratio.
4. **Cash_flow_rate:** Cash Flow from Operating Activities to Current Liabilities ratio.
5. **Interest_bearing_debt_interest_rate:** Interest-Bearing Debt to Equity ratio.
6. **Tax_rate_A:** Effective Tax Rate on taxable income.
7. **Cash_Flow_Per_Share:** After-tax earnings plus depreciation per share.
8. **Per_Share_Net_profit_before_tax_Yuan:** Pretax Income per share.
9. **Realized_Sales_Gross_Profit_Growth_Rate:** Growth rate of realized sales gross profit.
10. **Operating_Profit_Growth_Rate:** Growth rate of operating income.
11. **Continuous_Net_Profit_Growth_Rate:** Growth rate of net income excluding one-time gains/losses.
12. **Total_Asset_Growth_Rate:** Growth rate of total assets.
13. **Net_Value_Growth_Rate:** Growth rate of total equity.
14. **Total_Asset_Return_Growth_Rate_Ratio:** Return on total asset growth rate ratio.
15. **Cash_Reinvestment_perc:** Percentage of annual cash flow reinvested back into the business.
16. **Current_Ratio:** Ratio of current assets to current liabilities.
17. **Quick_Ratio:** Acid test ratio indicating liquidity.
18. **Interest_Expense_Ratio:** Interest Expenses to Total Revenue ratio.
19. **Total_debt_to_Total_net_worth:** Total Debt to Total Net Worth ratio.
20. **Long_term_fund_suitability_ratio_A:** Ratio of long-term liability and equity to fixed assets.
21. **Net_profit_before_tax_to_Paid_in_capital:** Ratio of net profit before tax to paid-in capital.
22. **Total_Asset_Turnover:** Net Sales to Average Total Assets ratio.
23. **Accounts_Receivable_Turnover:** Net Credit Sales to Average Accounts Receivable ratio.
24. **Average_Collection_Days:** Days Receivable Outstanding.
25. **Inventory_Turnover_Rate_times:** Inventory Turnover Rate.

26. **Fixed_Assets_Turnover_Frequency:** Fixed Asset Turnover.
27. **Net_Worth_Turnover_Rate_times:** Equity Turnover.
28. **Operating_profit_per_person:** Operation Income Per Employee.
29. **Allocation_rate_per_person:** Fixed Assets Per Employee.
30. **Quick_Assets_to_Total_Assets:** Quick Assets to Total Assets ratio.
31. **Cash_to_Total_Assets:** Cash to Total Assets ratio.
32. **Quick_Assets_to_Current_Liability:** Quick Assets to Current Liability ratio.
33. **Cash_to_Current_Liability:** Cash to Current Liability ratio.
34. **Operating_Funds_to_Liability:** Operating Funds to Liability ratio.
35. **Inventory_to_Working_Capital:** Inventory to Working Capital ratio.
36. **Inventory_to_Current_Liability:** Inventory to Current Liability ratio.
37. **Long_term_Liability_to_Current_Assets:** Long-term Liability to Current Assets ratio.
38. **Retained_Earnings_to_Total_Assets:** Retained Earnings to Total Assets ratio.
39. **Total_income_to_Total_expense:** Total Income to Total Expense ratio.
40. **Total_expense_to_Assets:** Total Expense to Assets ratio.
41. **Current_Asset_Turnover_Rate:** Current Asset Turnover Rate.
42. **Quick_Asset_Turnover_Rate:** Quick Asset Turnover Rate.
43. **Cash_Turnover_Rate:** Cash Turnover Rate.
44. **Fixed_Assets_to_Assets:** Fixed Assets to Assets ratio.
45. **Cash_Flow_to_Total_Assets:** Cash Flow to Total Assets ratio.
46. **Cash_Flow_to_Liability:** Cash Flow to Liability ratio.
47. **CFO_to_Assets:** Cash Flow from Operations to Assets ratio.
48. **Cash_Flow_to_Equity:** Cash Flow to Equity ratio.
49. **Current_Liability_to_Current_Assets:** Current Liability to Current Assets ratio.
50. **Liability_Assets_Flag:** Indicator for Total Liability exceeding Total Assets.
51. **Total_assets_to_GNP_price:** Total Assets to GNP price ratio.
52. **No_credit_Interval:** Interval without Credit.
53. **Degree_of_Financial_Leverage_DFL:** Degree of Financial Leverage.
54. **Interest_Coverage_Ratio_Interest_expense_to_EBIT:** Interest Coverage Ratio.
55. **Net_Income_Flag:** Indicator for Negative Net Income in the last two years.
56. **Equity_to_Liability:** Equity to Liability Ratio.
57. **Default:** Indicator for Company Default (1 if Defaulted, 0 if Not Defaulted).

Data Ingestion:

We begin our study by loading the dataset and setting up the Python environment. This involves importing essential libraries such as pandas, numpy, and scikit-learn for data manipulation, preprocessing, and modeling tasks. The dataset is then loaded into a pandas DataFrame, laying the groundwork for subsequent analysis and exploration.

Sample of the dataset

	Co_Code	Co_Name	_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_rate
0	16974	Hind.Cables	8.820000e+09	0.000000e+00	0.462045	0.000352	0.001
1	21214	Tata Tele. Mah.	9.380000e+09	4.230000e+09	0.460116	0.000716	0.0000
2	14852	ABG Shipyard	3.800000e+09	8.150000e+08	0.449893	0.000496	0.0000
3	2439	GTL	6.440000e+09	0.000000e+00	0.462731	0.000592	0.0093
4	23505	Bharati Defence	3.680000e+09	0.000000e+00	0.463117	0.000782	0.4002

Fig 1: Sample of the dataset

The initial exploration of the dataset reveals 2058 observations and 58 variables. This comprehensive dataset will serve as the foundation for our in-depth analysis of financial metrics and default prediction.

Checking for Duplicate Rows

We examined the dataset for any duplicate entries to ensure data integrity and avoid potential biases in our analysis. By using the duplicated() function, we identified and counted duplicate rows, which returned a total of 0 duplicate rows in the dataset.

Null Value Check

The columns "_Cash_Flow_Per_Share", "_Total_debt_to_Total_net_worth", "_Cash_to_Total_Assets", and "_Current_Liability_to_Current_Assets" have null values in the dataset.

Imputing values

We addressed missing values in the dataset by imputing them with the mode value for specific columns. This was done for columns such as '_Cash_Flow_Per_Share', '_Total_debt_to_Total_net_worth', '_Cash_to_Total_Assets', and '_Current_Liability_to_Current_Assets'. Imputation helps ensure that the dataset is complete and suitable for analysis without losing valuable information.

BOXPLOTS and Outlier Treatment

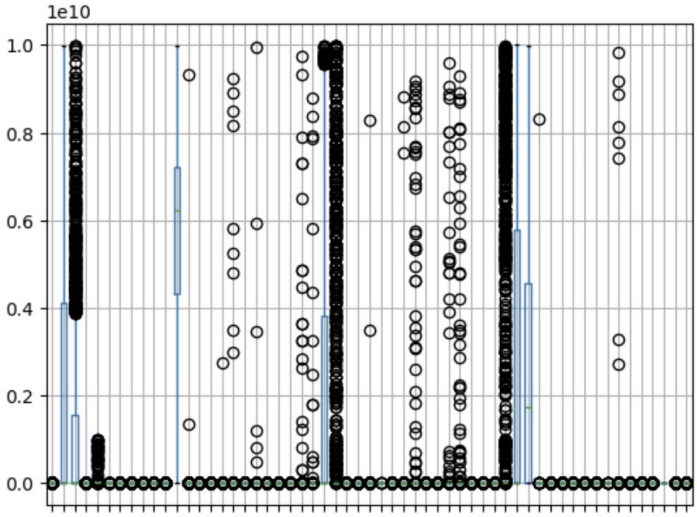


Fig 2 : Boxplot

We implemented outlier treatment by defining a function to detect and cap outliers using the interquartile range (IQR) method. This function iterates over each column, calculates the lower and upper bounds based on the IQR, and replaces outlier values beyond these bounds with the respective bound values. This approach ensures that extreme values do not disproportionately influence the analysis.

Dropping Irrelevant Columns

To focus on relevant features for the analysis, the 'Co_Code' and 'Co_Name' columns, which represent company identification, are removed from the dataset using the **drop()** method with **axis=1**. This ensures that only pertinent variables are retained for further analysis.

Top Correlated Features with Default

1. Total Debt to Total Net Worth: 0.341
2. Current Liability to Current Assets: 0.307
3. Total Expense to Assets: 0.184
4. Long-term Liability to Current Assets: 0.132
5. Fixed Assets Turnover Frequency: 0.123
6. Average Collection Days: 0.118
7. Fixed Assets to Assets: 0.086
8. Allocation Rate per Person: 0.080
9. Total Assets to GNP Price: 0.068
10. Research and Development Expense Rate: 0.064

These features demonstrate varying degrees of correlation with the target variable 'Default', suggesting their potential importance in predicting default occurrences

Heatmap of top 10 corelated features

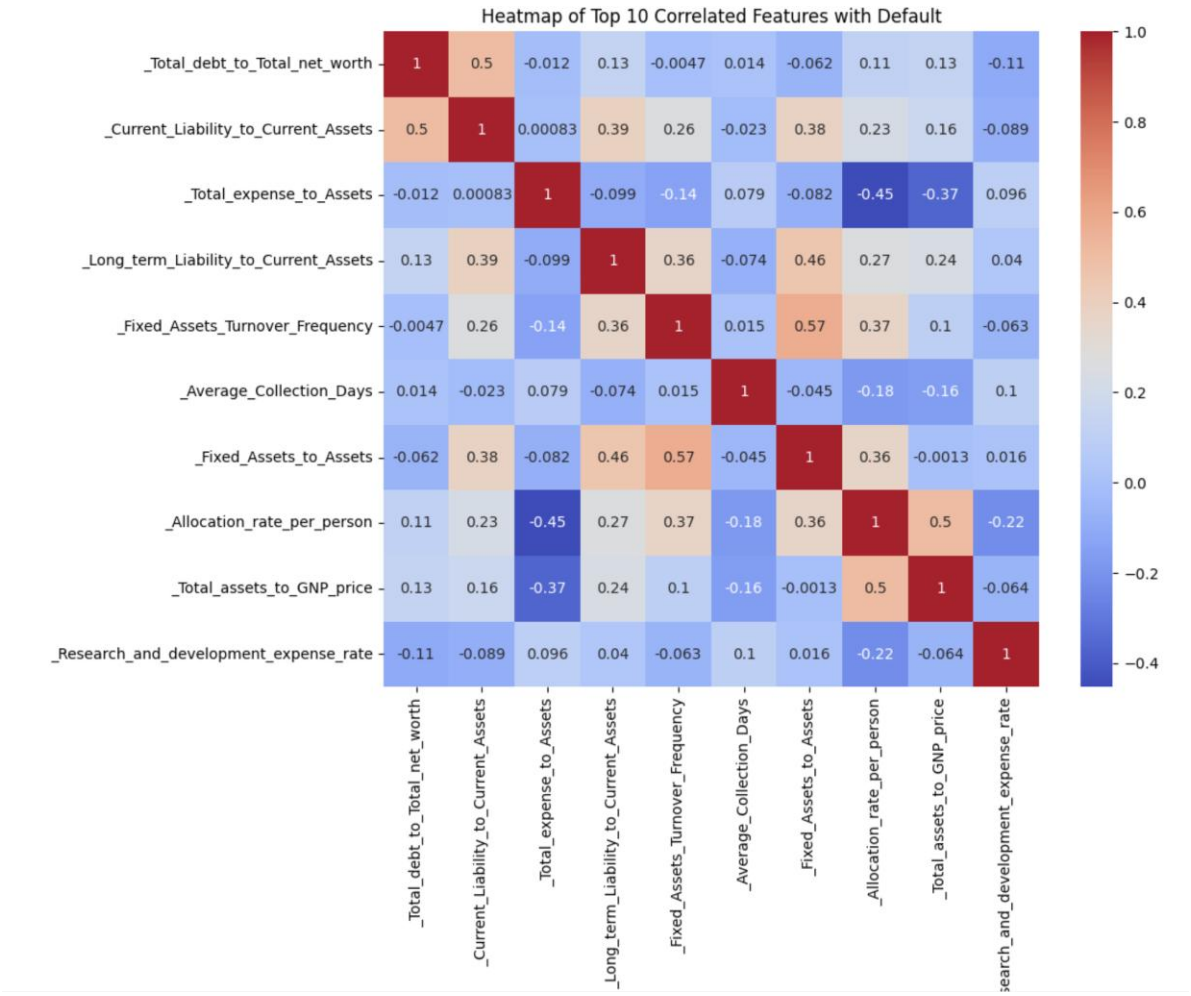
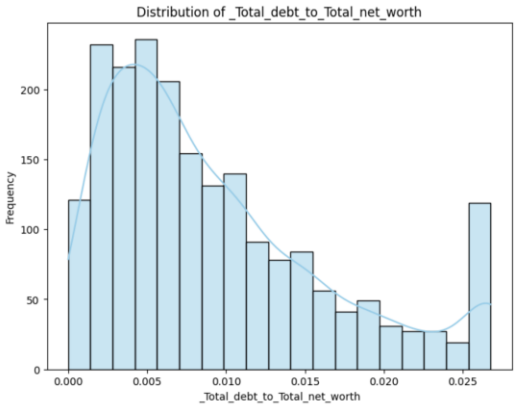
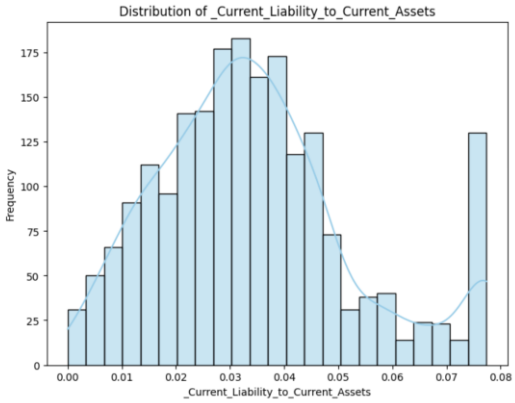


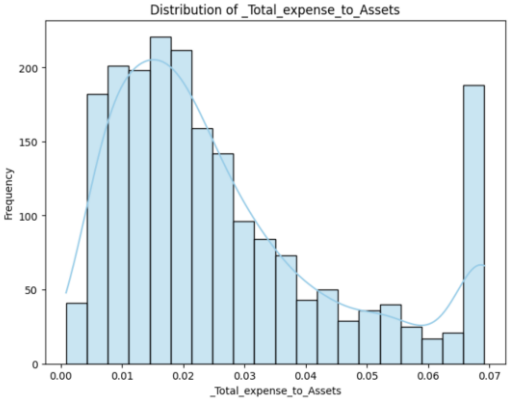
Fig 3: Heatmap of top 10 corelated features

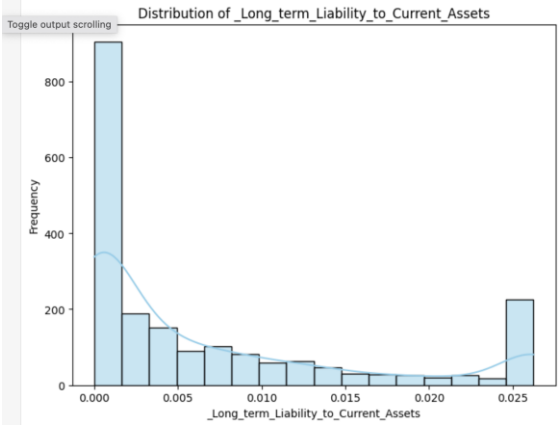
Univariate Analysis: Distribution of Top Correlated Features with Default

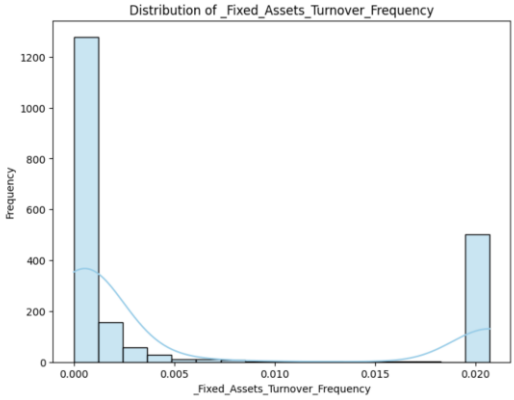
The following histograms display the distribution of each top correlated feature with the target variable 'Default'. These visualizations provide insights into the distribution patterns and potential outliers within each feature.

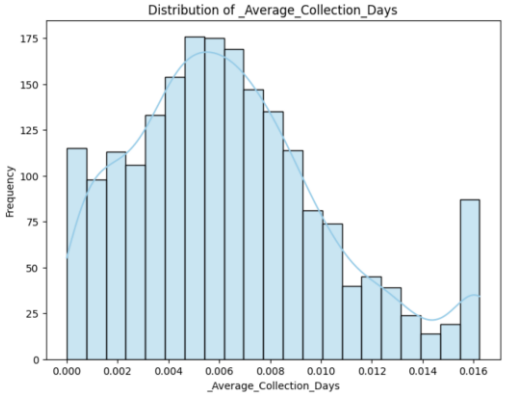

Distribution of _Total_debt_to_Total_net_worth
Frequency
_Total_debt_to_Total_net_worth


Distribution of _Current_Liability_to_Current_Assets
Frequency
_Current_Liability_to_Current_Assets


Distribution of _Total_expense_to_Assets
Frequency
_Total_expense_to_Assets


Toggle output scrolling
Distribution of _Long_term_Liability_to_Current_Assets
Frequency
_Long_term_Liability_to_Current_Assets


Distribution of _Fixed_Assets_Turnover_Frequency
Frequency
_Fixed_Assets_Turnover_Frequency


Distribution of _Average_Collection_Days
Frequency
_Average_Collection_Days

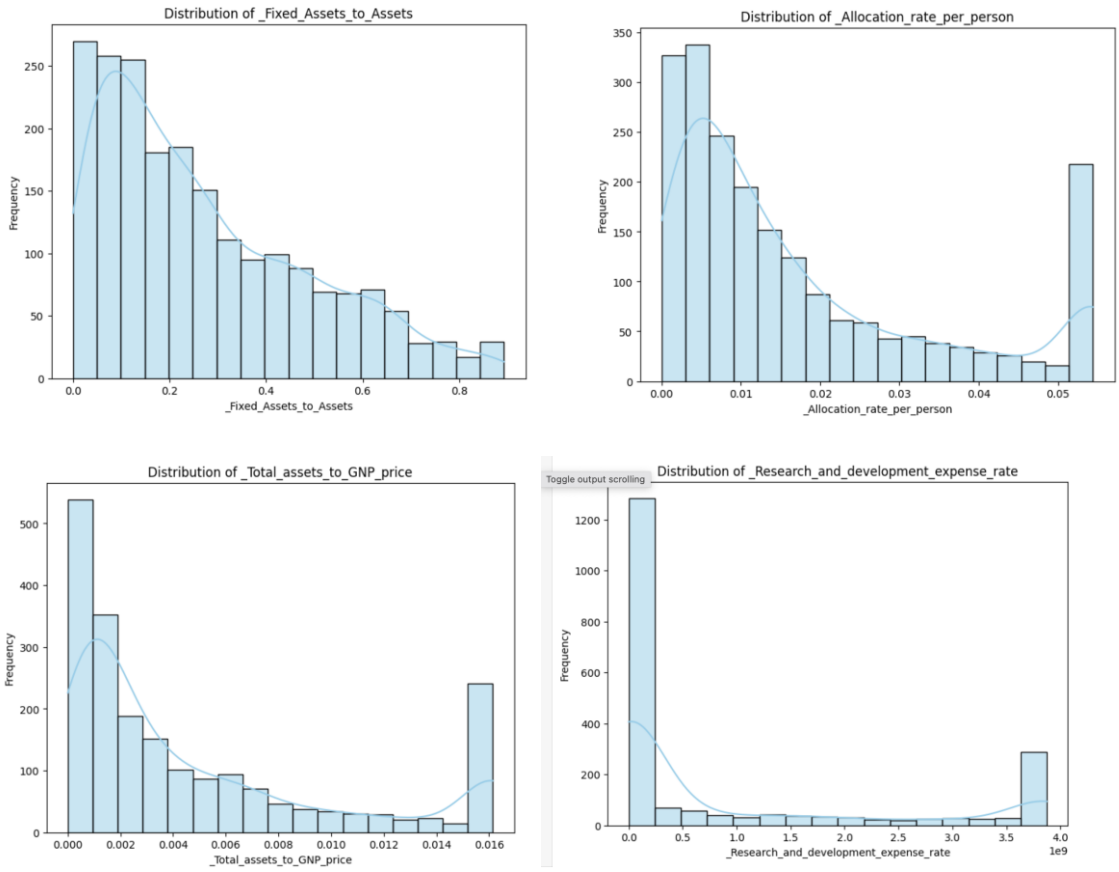
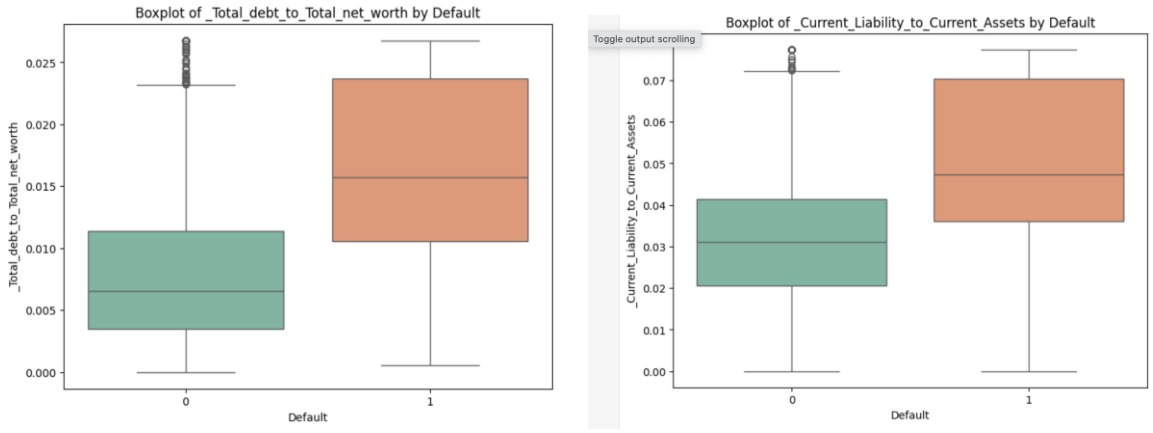
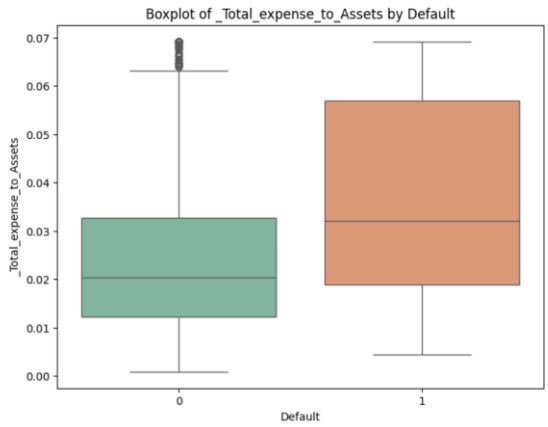


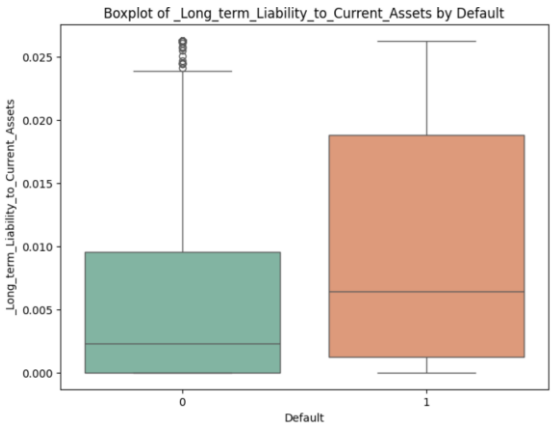
Fig 4: Univariate Analysis Distribution plots

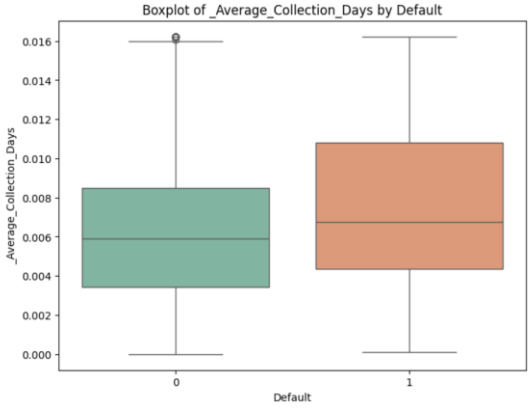
Bivariate Analysis: Boxplot of Top Correlated Features by Default

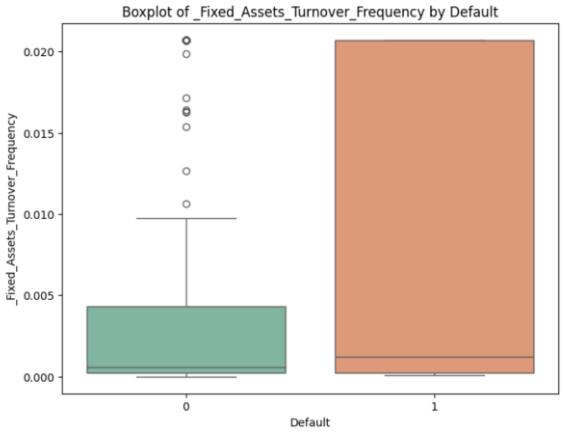
The boxplots above illustrate the distribution of each top correlated feature with respect to the 'Default' variable. This analysis helps identify potential differences in feature distributions between defaulted and non-defaulted companies.

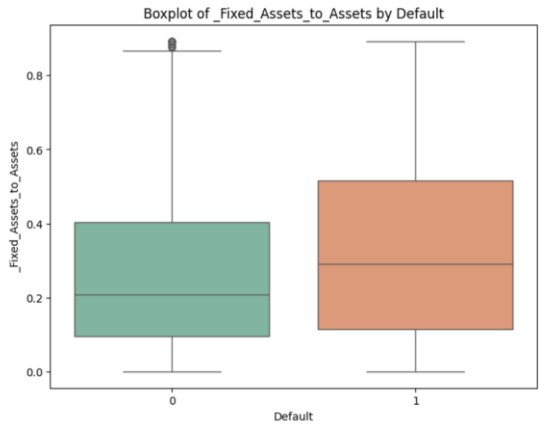


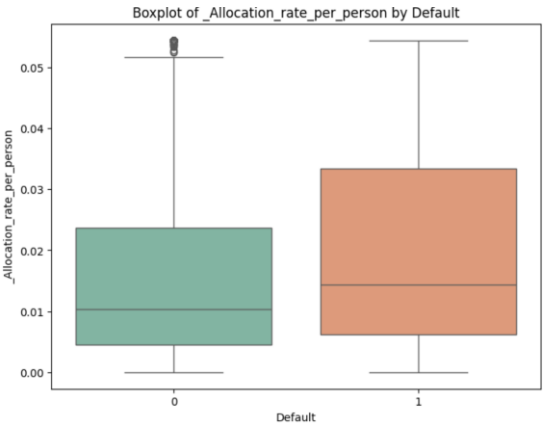

Boxplot of _Total_expense_to_Assets by Default
_Total_expense_to_Assets
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
1
Default


Boxplot of _Long_term_Liability_to_Current_Assets by Default
_Long_term_Liability_to_Current_Assets
0.000
0.005
0.010
0.015
0.020
0.025
0
1
Default


Boxplot of _Average_Collection_Days by Default
_Average_Collection_Days
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.016
0
1
Default


Boxplot of _Fixed_Assets_Turnover_Frequency by Default
_Fixed_Assets_Turnover_Frequency
0.000
0.005
0.010
0.015
0.020
0
1
Default


Boxplot of _Fixed_Assets_to_Assets by Default
_Fixed_Assets_to_Assets
0.0
0.2
0.4
0.6
0.8
0
1
Default


Boxplot of _Allocation_rate_per_person by Default
_Allocation_rate_per_person
0.00
0.01
0.02
0.03
0.04
0.05
0
1
Default

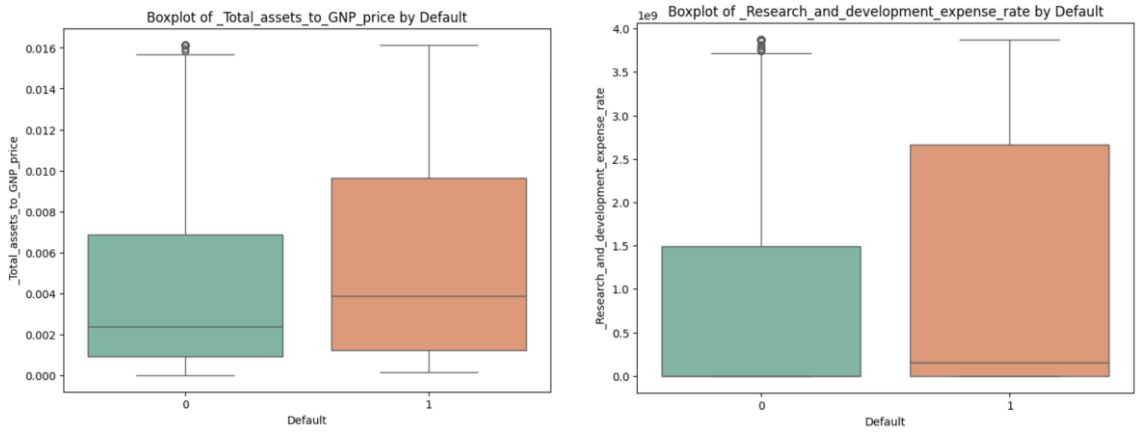


Fig 5 : Bivariate Analysis Boxplots

Data Splitting for Model Training and Testing

The dataset has been divided into training and testing sets with a ratio of 67:33, respectively. This splitting facilitates model training on the training data and subsequent evaluation on unseen testing data. The shapes of the training and testing sets are displayed above.

Dataset	Number of Observations	Number of Features
X_train	1378	10
X_test	680	10
y_train	1378	-
y_test	680	-

Table 1:Train test split

Logistic Regression Model Building Approach and Performance

Model Building Approach:

1. **Data Preparation:** Features were selected based on their correlation with the target variable.
2. **Train-Test Split:** The dataset was split into training and testing sets with a ratio of 67:33.
3. **Model Training:** Logistic regression model was fitted on the training data using **statsmodels** library.
4. **Threshold Optimization:** The optimal threshold for classification was determined using the ROC curve on the training set.
5.

Model Performance on Training Set:

- **Optimal Threshold:** 0.088
- **Classification Report:**

- **Precision (Defaulted):** 0.28
- **Recall (Defaulted):** 0.82
- **F1-score (Defaulted):** 0.42
- **Accuracy:** 0.77
- **Area Under ROC Curve (AUC):** 0.86

This logistic regression model demonstrates a good balance between precision and recall, with an accuracy of 77% on the training set.

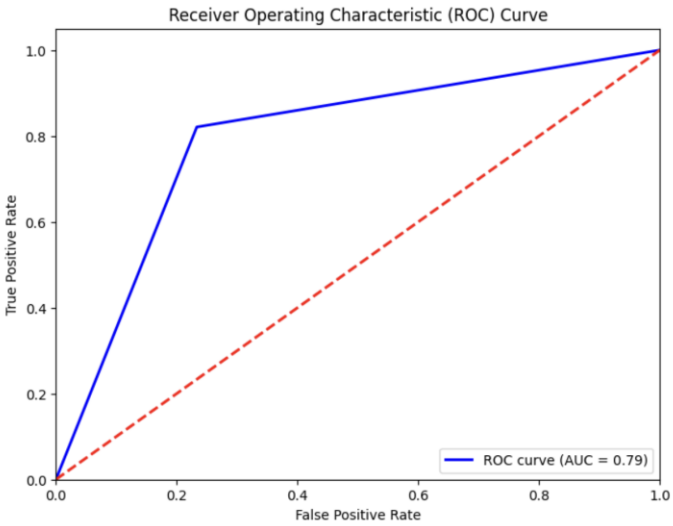


Fig 6: ROC Curve Logistic Regression

Random Forest Model Building Approach and Performance

Model Building Approach:

1. **Data Preparation:** Features were selected based on their correlation with the target variable.
2. **Train-Test Split:** The dataset was split into training and testing sets with a ratio of 67:33.
3. **Model Training:** Random Forest classifier was instantiated and fitted on the training data.

Model Performance on Test Set:

- **Accuracy:** 0.92
- **Classification Report:**
  - **Precision (Defaulted):** 0.74
  - **Recall (Defaulted):** 0.25
  - **F1-score (Defaulted):** 0.38
  - **Area Under ROC Curve (AUC):** 0.82

The Random Forest model achieved an accuracy of 92% on the test set. However, the precision and recall for detecting defaulted cases are relatively low, indicating the need for further model optimization.

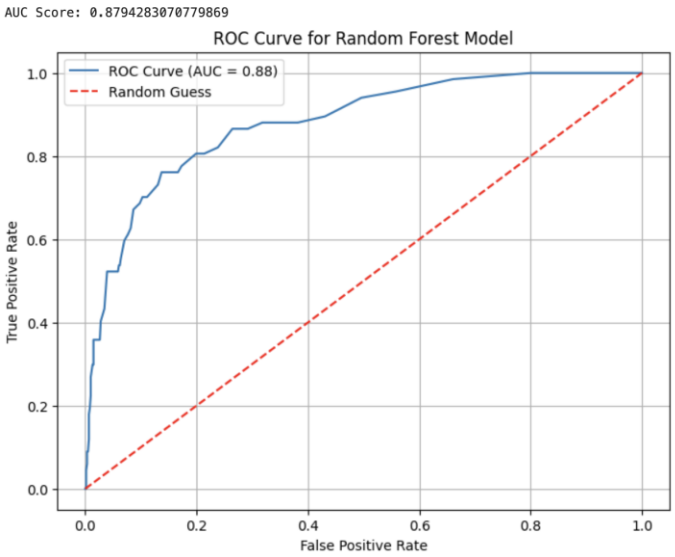


Fig 7: ROC Curve Random Forest

Linear Discriminant Analysis (LDA) Model Performance

Model Approach:

1. **Initialization:** The Linear Discriminant Analysis (LDA) model was initialized.
2. **Model Training:** The LDA model was trained using the training data.
3. **Prediction:** Predictions were made on the test dataset.

Model Performance:

- **Accuracy:** 0.91
- **Classification Report:**
  - **Precision (Defaulted):** 0.60
  - **Recall (Defaulted):** 0.37
  - **F1-score (Defaulted):** 0.46
- **Confusion Matrix:**
  - True Negatives (TN): 596
  - False Positives (FP): 17
  - False Negatives (FN): 42
  - True Positives (TP): 25
- **Area Under ROC Curve (AUC):** 0.80

The LDA model achieved an accuracy of 91% on the test set. The precision and recall for detecting defaulted cases are relatively better than the Random Forest model, indicating its potential effectiveness in identifying default cases. However, there is still room for improvement in terms of model performance.

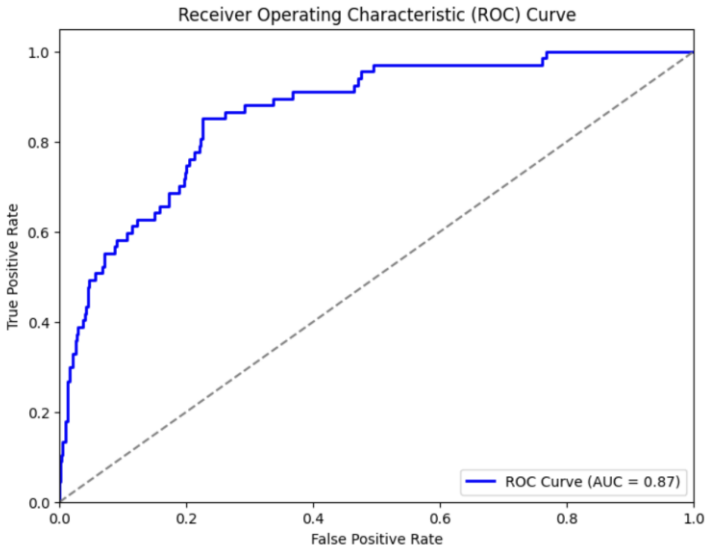


Fig 8: ROC Curve Linear Discriminant Analysis

Performance Comparison of Classification Models:

Model	Accuracy	AUC
Logistic Regression	0.77	0.86
Random Forest	0.92	0.77
Linear Discriminant Analysis	0.91	0.80

Table 2: Performance Comparison of Classification Models

Insights:

- **Accuracy:** Random Forest model achieved the highest accuracy (92%), followed by LDA (91%) and Logistic Regression (77%).
- **AUC (Area Under ROC Curve):** Logistic Regression model has the highest AUC (0.86), indicating its better discriminatory power compared to Random Forest (0.77) and LDA (0.80).

Conclusions:

1. Model Performance: After evaluating Logistic Regression, Random Forest, and Linear Discriminant Analysis (LDA) models, we found that each model had its strengths and weaknesses.

2. Logistic Regression: While it demonstrated a moderate accuracy of 77%, it exhibited the highest discriminatory power with an AUC of 0.86. This suggests that it is effective at distinguishing between default and non-default cases.
3. Random Forest: With an accuracy of 92%, the Random Forest model outperformed the other models in terms of overall prediction accuracy. However, its AUC of 0.77 indicates that it may not be as effective at correctly identifying positive cases as Logistic Regression.
4. Linear Discriminant Analysis (LDA): LDA achieved an accuracy of 91%, making it comparable to Random Forest. However, its AUC of 0.80 suggests that it may not perform as well as Logistic Regression in terms of discriminatory power.

Recommendations:

1. Utilize Logistic Regression for Discriminatory Power: Given its high AUC, Logistic Regression is recommended when the emphasis is on correctly identifying positive cases (defaulted companies). It may be particularly useful when the cost of false negatives (misclassifying a defaulted company as non-defaulted) is high.
2. Consider Random Forest for Overall Accuracy: If the primary goal is to maximize overall prediction accuracy, Random Forest is a suitable choice. Its robustness to outliers and non-linearity can lead to better performance in complex datasets.
3. Explore Ensemble Methods: Ensemble methods, such as combining the predictions of multiple models (e.g., Logistic Regression, Random Forest, and LDA), could potentially improve overall performance by leveraging the strengths of each model.
4. Regular Model Evaluation: Continuous monitoring and evaluation of model performance are crucial. As business dynamics evolve, retraining models and reassessing their performance regularly can ensure that they remain effective and relevant.
5. Further Investigation: Investigate features contributing to model predictions. Understanding the relationship between predictor variables and the target variable (default) can provide valuable insights into the underlying factors driving default risk.

END OF REPORT