Financial Risk Analysis Project

By Abdullah AL KHAMYASI

Executive Summary

- The focal point of the report is the utilization of classification models, namely Logistic Regression, Random Forest, and Linear Discriminatory Analysis, for the examination of company financials. The main goal is to determine the financial well-being of a specific company and forecast the probability of it experiencing default in the next year. These classification models function as instruments to classify companies according to their financial statuses, offering valuable information for decision-makers regarding potential risks related to default. The emphasis is on utilizing statistical and machine learning methods to improve comprehension of financial stability and risk within the scope of the scrutinized companies.

Problem Statement

- Businesses or corporations may face default if they struggle to meet their debt obligations. Defaulting can result in a diminished credit rating for the company, reducing its future credit prospects and potentially leading to higher interest rates on existing and new debts. From an investor's perspective, the appeal of investing in a company lies in its ability to manage financial commitments, facilitate rapid growth, and effectively handle expansion.
- A crucial financial document, the balance sheet offers a snapshot of a company's assets, liabilities, and shareholders' investments, serving as a vital tool for assessing business performance. The available data encompasses information extracted from the previous year's financial statements of the companies.
- Regarding the dependent variable, there is no need to create a new one, as the 'Default' variable is already present in the dataset and can be considered the dependent variable. For the test-train split, the data should be divided into training and testing datasets in a 67:33 ratio, with a random state set to 42 (random_state=42). The model-building process will be conducted using the training dataset, while model validation will be carried out on the testing dataset.

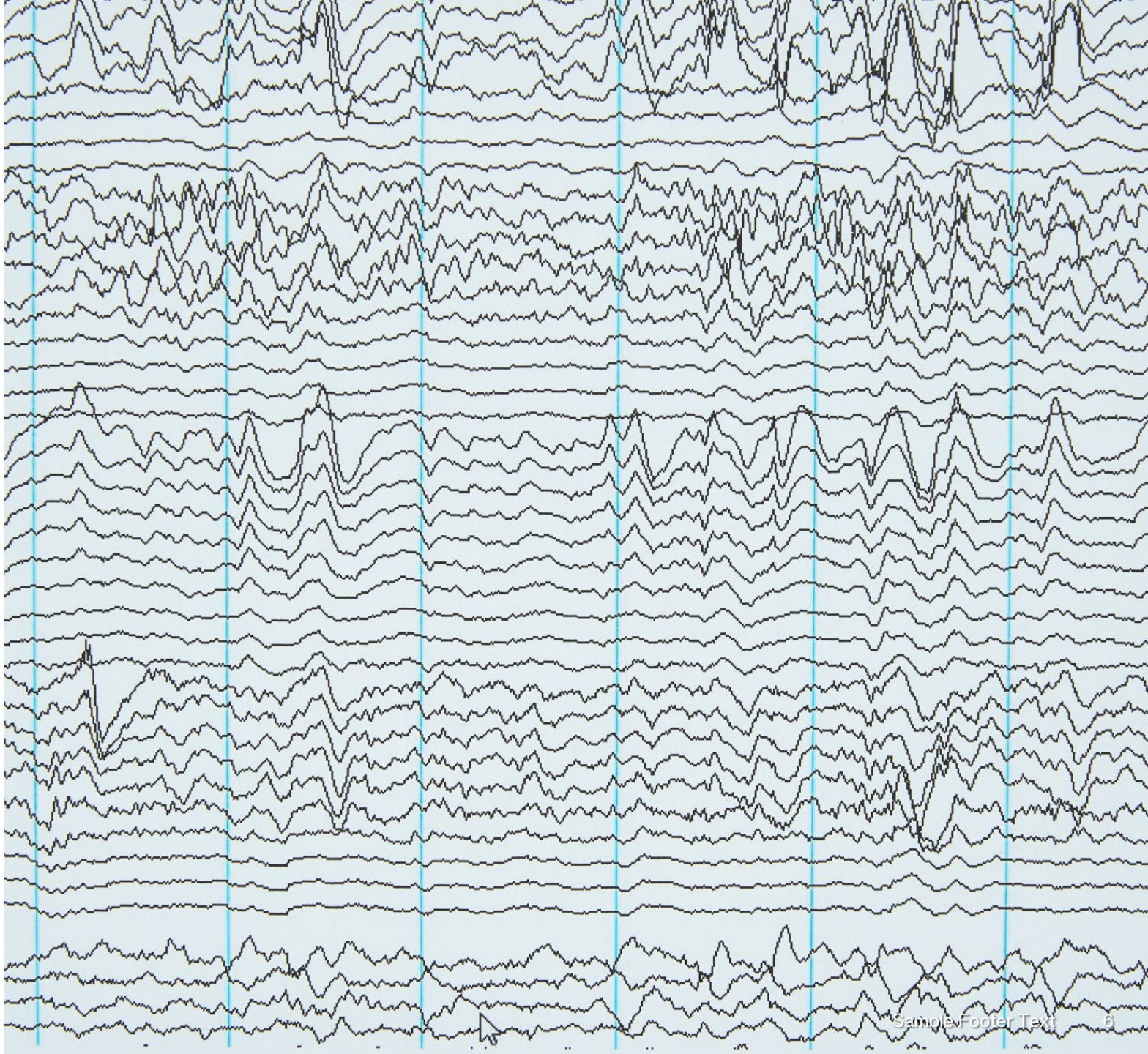
Data Dictionary

Sl. No	Column Name	Description
1	Co_Code	Company Code
2	Co_Name	Company Name
3	_Operating_Expense_Rate	Operating Expense Rate: Operating Expenses/Net Sales. The operating expense ratio (OER) is the cost to operate a piece of property compared to the income the property brings in.
4	_Research_and_development_expense_rate	Research and development expense rate: (Research and Development Expenses)/Net Sales. Research and development (R&D) expenses are direct expenditures relating to a company's efforts to develop, design, and enhance its products, services, technologies, or processes.
5	_Cash_flow_rate	Cash flow rate: Cash Flow from Operating/Current Liabilities. Cash flow is a measure of how much cash a business brought in or spent in total over a period of time.
6	_Interest_bearing_debt_interest_rate	Interest-bearing debt interest rate: Interest-bearing Debt/Equity
7	_Tax_rate_A	Tax rate (A): Effective Tax Rate. Effective tax rate represents the percentage of their taxable income that individuals pay in taxes. For corporations, the effective corporate tax rate is the rate they pay on their pre-tax profits.
8	_Cash_Flow_Per_Share	Cash Flow Per Share. It is the after-tax earnings plus depreciation on a per-share basis that functions as a measure of a firm's financial strength
9	_Per_Share_Net_profit_before_tax_Yuan_	Per Share Net profit before tax (Yuan ¥): Pretax Income Per Share. Pretax income, also known as earnings before tax or pretax earnings, is the net income earned by a business before taxes are subtracted/accounted for.
10	_Realized_Sales_Gross_Profit_Growth_Rate	Realized Sales Gross Profit Growth Rate.
11	_Operating_Profit_Growth_Rate	Operating Profit Growth Rate: Operating Income Growth. It is the rate of increase in operating income over the last year.
12	_Continuous_Net_Profit_Growth_Rate	Continuous Net Profit Growth Rate: Net Income-Excluding Disposal Gain or Loss Growth
13	_Total_Asset_Growth_Rate	Total Asset Growth Rate: Total Asset Growth. It is the rate at which how quickly the company has been growing its Assets
14	_Net_Value_Growth_Rate	Net Value Growth Rate: Total Equity Growth
15	_Total_Asset_Return_Growth_Rate_Ratio	Total Asset Return Growth Rate Ratio: Return on Total Asset Growth
16	_Cash_Reinvestment_perc	Cash Reinvestment %: Cash Reinvestment Ratio. It is the valuation ratio that is used to measure the percentage of annual cash flow that the company invests back into the business as a new investment.
17	_Current_Ratio	Current Ratio. The current ratio describes the relationship between a company's assets and liabilities
18	_Quick_Ratio	Quick Ratio: Acid Test. Acid-test ratio (also known as quick ratio) is a measure of a company's liquidity, which is its ability to pay its short-term obligations using only its most liquid assets.
19	_Interest_Expense_Ratio	Interest Expense Ratio: Interest Expenses/Total Revenue
20	_Total_debt_to_Total_net_worth	Total debt/Total net worth: Total Liability/Equity Ratio
21	_Long_term_fund_suitability_ratio_A	Long-term fund suitability ratio (A): (Long-term Liability+Equity)/Fixed Assets
22	_Net_profit_before_tax_to_Paid_in_capital	Net profit before tax/Paid-in capital: Pretax Income/Capital
23	_Total_Asset_Turnover	Total Asset Turnover. Net Sales/Average Total Assets
24	_Accounts_Receivable_Turnover	Accounts Receivable Turnover. The accounts receivable turnover ratio, or receivables turnover, is used in business accounting to quantify how well companies are managing the credit that they extend to their customers by evaluating how long it takes to collect the outstanding debt throughout the accounting period.
25	_Average_Collection_Days	Average Collection Days: Days Receivable Outstanding
26	_Inventory_Turnover_Rate_times	Inventory Turnover Rate (times). The inventory turnover ratio is the number of times a company has sold and replenished its inventory over a specific amount of time. The formula can also be used to calculate the number of days it will take to sell the inventory on hand.
27	_Fixed_Assets_Turnover_Frequency	Fixed Assets Turnover Frequency. Fixed Asset Turnover (FAT) is an efficiency ratio that indicates how well or efficiently a business uses fixed assets to generate sales. This ratio divides net sales by net fixed assets, calculated over an annual period.
28	_Net_Worth_Turnover_Rate_times	Net Worth Turnover Rate (times): Equity Turnover. Equity turnover is a ratio that measures the proportion of a company's sales to its stockholders' equity. The intent of the measurement is to

		determine the efficiency with which management is using equity to generate revenue.
29	_Operating_profit_per_person	Operating profit per person: Operation Income Per Employee
30	_Allocation_rate_per_person	Allocation rate per person: Fixed Assets Per Employee
31	_Quick_Assets_to_Total_Assets	Quick Assets/Total Assets
32	_Cash_to_Total_Assets	Cash/Total Assets
33	_Quick_Assets_to_Current_Liability	Quick Assets/Current Liability
34	_Cash_to_Current_Liability	Cash/Current Liability
35	_Operating_Funds_to_Liability	Operating Funds to Liability
36	_Inventory_to_Working_Capital	Inventory/Working Capital
37	_Inventory_to_Current_Liability	Inventory/Current Liability
38	_Long_term_Liability_to_Current_Assets	Long-term Liability to Current Assets
39	_Retained_Earnings_to_Total_Assets	Retained Earnings to Total Assets
40	_Total_income_to_Total_expense	Total income/Total expense
41	_Total_expense_to_Assets	Total expense/Assets
42	_Current_Asset_Turnover_Rate	Current Asset Turnover Rate: Current Assets to Sales. The current assets turnover ratio indicates how many times the current assets are turned over in the form of sales within a specific period of time. A higher asset turnover ratio means a better percentage of sales.
43	_Quick_Asset_Turnover_Rate	Quick Asset Turnover Rate: Quick Assets to Sales. The asset turnover ratio measures the efficiency of a company's assets in generating revenue or sales.
44	_Cash_Turnover_Rate	Cash Turnover Rate: Cash to Sales. The cash turnover ratio is an efficiency ratio that reveals the number of times that cash is turned over in an accounting period.
45	_Fixed_Assets_to_Assets	Fixed Assets to Assets. Fixed assets are also known as non-current assets—assets that can't be easily converted into cash.
46	_Cash_Flow_to_Total_Assets	Cash Flow to Total Assets. This ratio indicates the cash a company can generate in relation to its size.
47	_Cash_Flow_to_Liability	Cash Flow to Liability. The amount of money available to run business operations and complete transactions. This is calculated as current assets (cash or near-cash assets, like notes receivable) minus current liabilities (liabilities due during the upcoming accounting period)
48	_CFO_to_Assets	CFO to Assets. Cash flow on total assets is an efficiency ratio that rates cash flows to the company
		assets without being affected by income recognition or income measurements.
49	_Cash_Flow_to_Equity	Cash Flow to Equity. cash flow to equity is a measure of how much cash is available to the equity shareholders of a company after all expenses, reinvestment, and debt are paid.
50	_Current_Liability_to_Current_Assets	Current Liability to Current Assets. Current liabilities are a company's financial commitments that are due and payable within a year, Current assets are projected to be consumed, sold, or converted into cash within a year or within the operational cycle.
51	_Liability_Assets_Flag	Liability-Assets Flag: 1 if Total Liability exceeds Total Assets, 0 otherwise
52	_Total_assets_to_GNP_price	Total assets to GNP price. Gross National Product (GNP) is the total value of all finished goods and services produced by a country's citizens in a given financial year, irrespective of their location.
53	_No_credit_Interval	No-credit Interval
54	_Degree_of_Financial_Leverage_DFL	Degree of Financial Leverage (DFL). The degree of financial leverage is a financial ratio that measures the sensitivity in fluctuations of a company's overall profitability to the volatility of its operating income caused by changes in its capital structure.
55	_Interest_Coverage_Ratio_Interest_expense_to_EBIT	Interest Coverage Ratio (Interest expense to EBIT). The interest coverage ratio is a debt and profitability ratio used to determine how easily a company can pay interest on its outstanding debt. The interest coverage ratio is calculated by dividing a company's earnings before interest and taxes (EBIT) by its interest expense during a given period.
56	_Net_Income_Flag	Net Income Flag: 1 if Net Income is Negative for the last two years, 0 otherwise
57	_Equity_to_Liability	Equity to Liability Ratio.
58	Default	Whether the Company has Default (Bankrupted) or not? 1 - Defaulted, 0 - Not Defaulted.

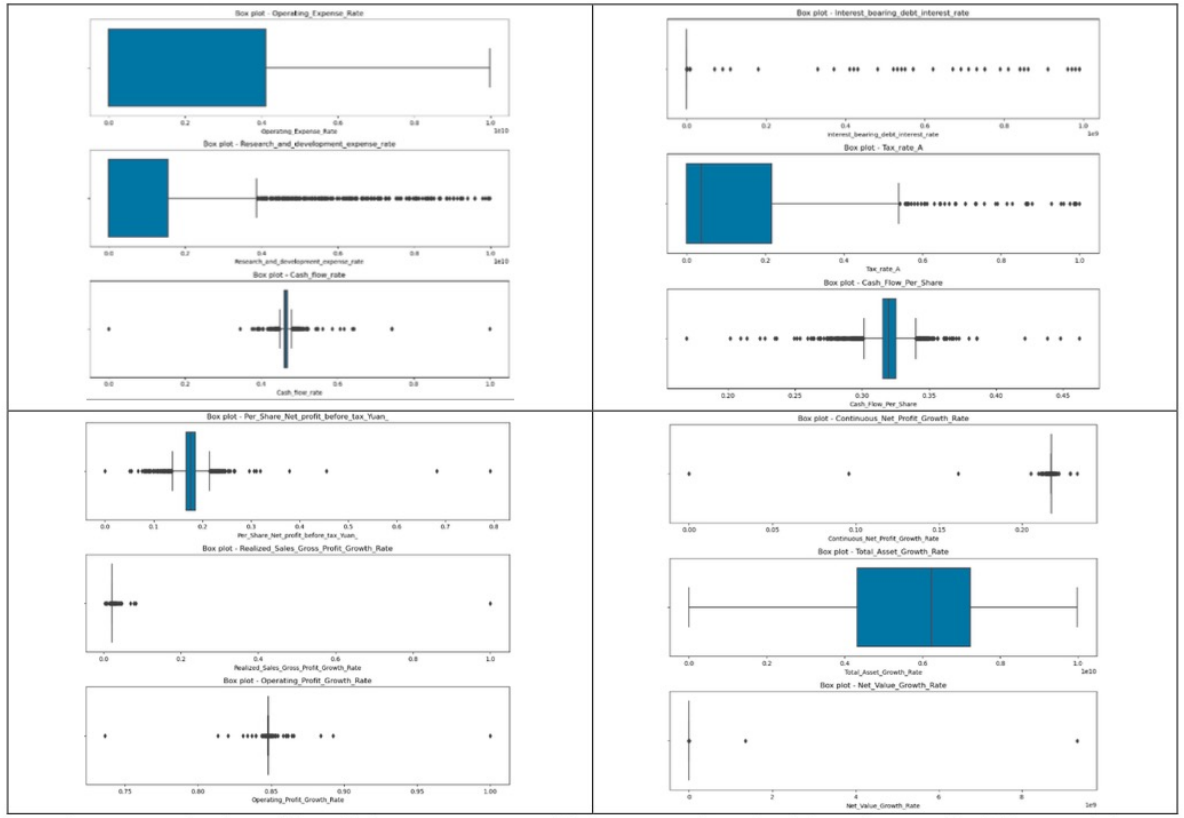
Exploratory data analysis

- **Basic Analysis:**
- Total Companies: 2058
- Total Variables: 58
- Target Variable: 'Default'
- Duplicate Entries: 0
- Default Instances in Dataset: 10.69% (Unbalanced Dataset)
- **Missing Value Handling:**
- Imputation Method: Mean used for replacing missing values
- **Outlier Handling:**
- Detection and Treatment: Employed IQR methods for identifying and addressing outliers



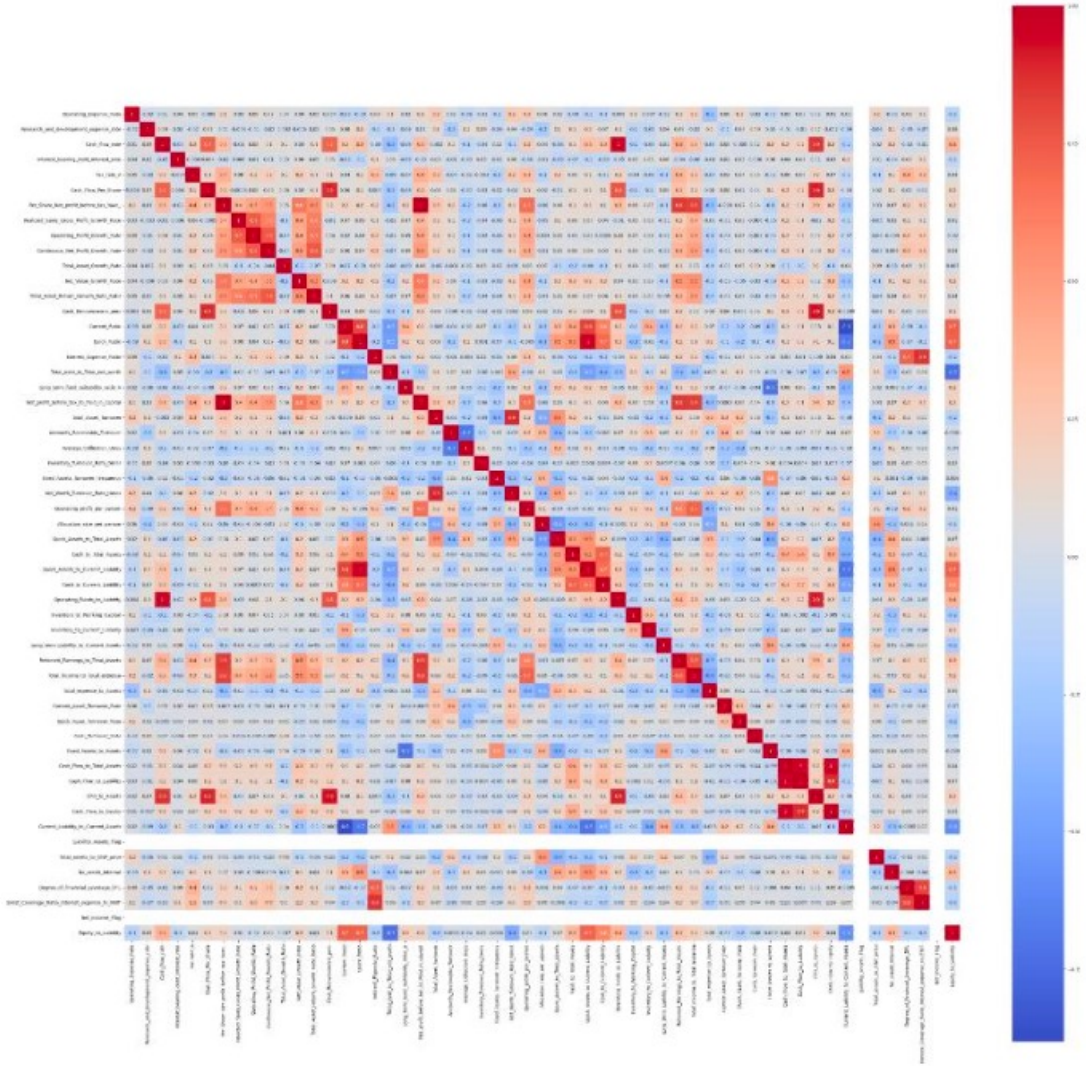
Univariate Analysis

- Box plots were utilized to assess all numeric variables, revealing the presence of outliers in nearly all of them. Consequently, outlier treatment is deemed necessary for these variables.



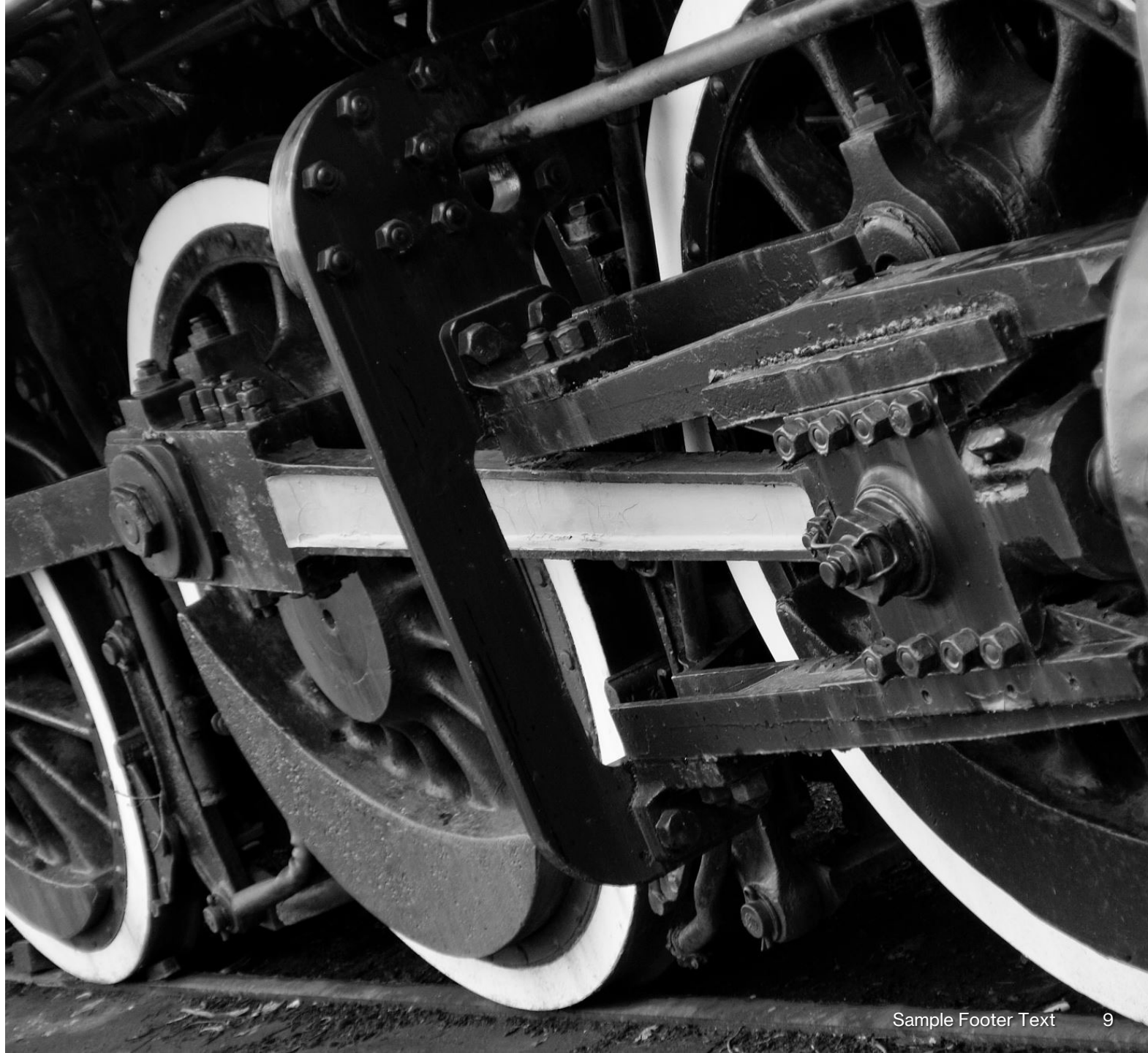
Bivariate Analysis

- To conduct bivariate analysis, a correlation heatmap was employed. The heatmap highlighted numerous red patches, indicating high correlations between variables. This raises concerns about multi-collinearity. To address this, a check for multi-collinearity was performed using the Variance Inflation Factor (VIF), resulting in the reduction of independent variables to 43.



Train Test Split

- The data was divided into a training set (67%) and a testing (validation) set (33%), with a random seed set at random_state=42. The training set comprised 1378 observations, while the testing set had 680 observations.



Model Building

Logistic Regression: We construct several Logistic Regression models employing various approaches and strategies. Each model undergoes testing on the test set, with iterative adjustments made to enhance the Recall and Precision metrics specifically for default=1. The development and evaluation of these models involve the utilization of both StatsModel and SciKitLearn libraries.

Performance Metrics of Logistic Regression on Train Dataset is given below:

	Precision	Recall	F1-score	Support
0	0.89	0.99	0.94	1225
1	0.07	0.01	0.01	153
Accuracy			0.88	1378
Macro Avg	0.48	0.50	0.47	1378
Weighted Avg	0.80	0.88	0.83	1378

Performance Metrics of Logistic Regression on Test Dataset is given below

	Precision	Recall	F1-score	Support
0	0.90	0.99	0.94	613
1	0.00	0.00	0.00	67
Accuracy			0.89	680
Macro Avg	0.45	0.49	0.47	680
Weighted Avg	0.81	0.89	0.85	680

The LR Model is not a good model. The recalls are not evident in this model.

Random Forest

- Performance Metrics of Random Forest on Train Dataset is given below:

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	1225
1	1.00	1.00	1.00	153
Accuracy			1.00	1378
Macro Avg	1.00	1.00	1.00	1378
Weighted Avg	1.00	1.00	1.00	1378

- Performance Metrics of Random Forest on Test Dataset is given below:

	Precision	Recall	F1-score	Support
0	0.94	0.98	0.96	613
1	0.67	0.43	0.53	67
Accuracy			0.92	680
Macro Avg	0.81	0.70	0.74	680
Weighted Avg	0.91	0.92	0.92	680

The accuracy of this model is 92%. It is a good model and should be considered to be used for this dataset.

Linear Discriminant Analysis

- We build a Linear Discriminant Analysis model with different approaches and strategies.
- We test each model on Test set and fine-tune to improve Recall and Precision of default=1

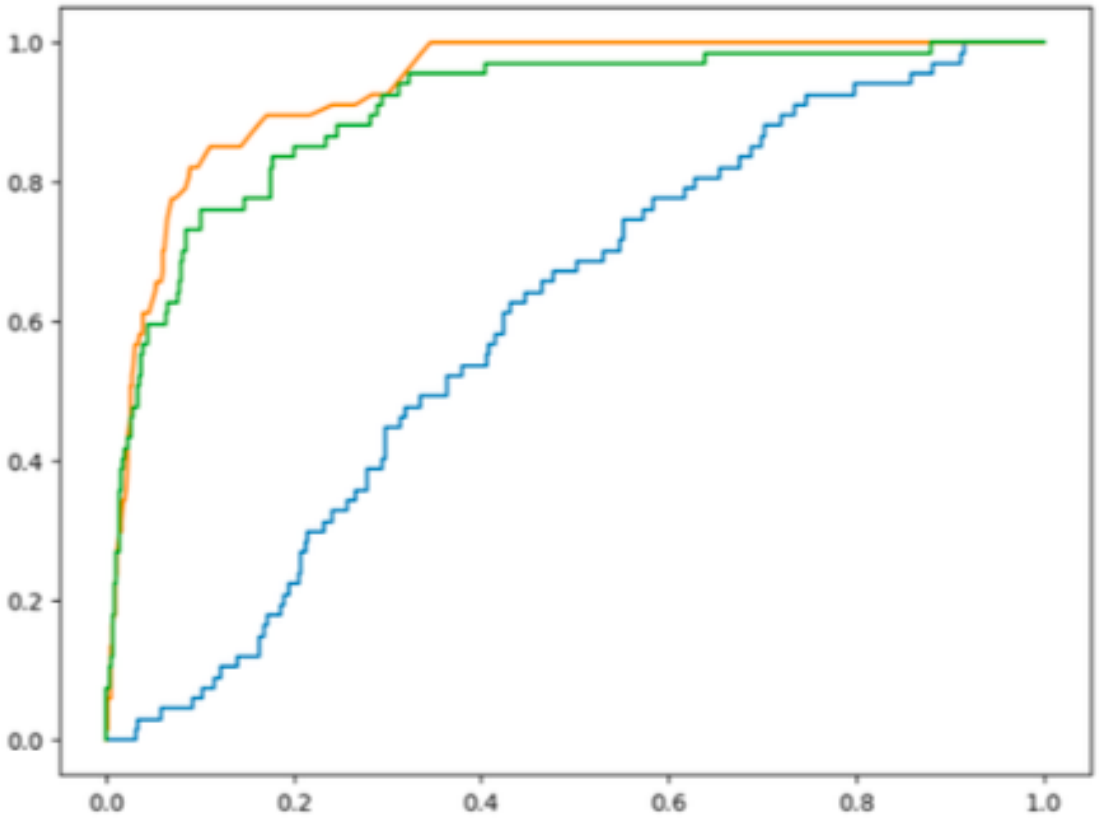
Performance Metrics of Linear Discriminant Analysis on Train Dataset is given below:

	Precision	Recall	F1-score	Support
0	0.95	0.96	0.95	1225
1	0.62	0.56	0.59	153
Accuracy			0.91	1378
Macro Avg	0.79	0.76	0.77	1378
Weighted Avg	0.91	0.91	0.91	1378

- Performance Metrics of Linear Discriminant Analysis on Test Dataset is given below:

	Precision	Recall	F1-score	Support
0	0.96	0.94	0.95	613
1	0.54	0.60	0.57	67
Accuracy			0.91	680
Macro Avg	0.75	0.77	0.76	680
Weighted Avg	0.91	0.91	0.91	680

LDA is also a good model with 91% accuracy. However the recalls are lesser than RF models.



Best Model

- The most effective model is the random forest, achieving a prediction accuracy of 92%.