Finance and Risk Analytics Project

Pavithra Devi
PGPDSBA
Great Learning

INDEX

	PART B: Conclusions and Recommendations	

PART A:

Introduction:

In the world of finance and investment, understanding the financial health and stability of companies is crucial for making informed decisions. One of the key indicators of a company's financial stability is its ability to meet its debt obligations. When a company fails to meet these obligations, it can lead to default, which has serious consequences for the company, its investors, and the broader economy.

This report focuses on analyzing the financial data of companies to predict the likelihood of default. By leveraging historical financial statements, specifically balance sheets, we aim to develop a predictive model that can identify companies at risk of default.

Problem Statement:

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interest on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale. A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

Objective:

The main objective of this analysis is to develop a predictive model that can accurately identify companies at risk of default based on their financial statements. By leveraging machine learning techniques and historical financial data, we aim to provide investors and financial institutions with a tool to assess the creditworthiness of companies and make informed decisions.

Methodology:

- Data Collection: We gather historical financial data, specifically balance sheet information, for a sample of companies.

- Data Preprocessing: We clean and preprocess the data to handle missing values, outliers, and ensure consistency in formatting.

- Feature Engineering: We extract relevant features from the balance sheet data that are indicative of a company's financial health and stability.

- Model Building: We train machine learning models, such as logistic regression, random forest, or support vector machines, to predict the likelihood of default based on the extracted features.

- Model Evaluation: We evaluate the performance of the trained models using appropriate metrics such as accuracy, precision, recall, and ROC-AUC.

- Model Validation: Finally, we validate the performance of the best-performing model on a separate test dataset to assess its generalization ability.

Expected Outcome:

We expect that the developed predictive model will accurately identify companies at risk of default, providing valuable insights to investors, financial institutions, and other stakeholders in the financial industry. This information can help mitigate risks, optimize investment strategies, and contribute to overall financial stability.

Through this report, we aim to provide a comprehensive understanding of the problem of predicting default risk and demonstrate the potential of machine learning techniques in addressing this critical issue in finance.

Dataset: Credit Risk Dataset

Data Dictionary: Data Dictionary

Description of Dataset:

The dataset contains financial information for **2058 companies, with 58 columns** representing different financial metrics. Here is a brief overview of the dataset:

1. Co_Code: Unique identifier for each company.
2. Co_Name: Name of the company.
3. Operating_Expense_Rate:Ratio of operating expenses to total revenue.
4. Research_and_development_expense_rate: Ratio of research and development expenses to total revenue.
5. Cash_flow_rate: Cash flow rate.
6. Interest_bearing_debt_interest_rate: Interest-bearing debt interest rate.
7. Tax_rate_A: Tax rate.
8. Cash_Flow_Per_Share: Cash flow per share.
9. Per_Share_Net_profit_before_tax_Yuan_: Net profit before tax per share in Yuan.
10. Realized_Sales_Gross_Profit_Growth_Rate: Growth rate of realized sales gross profit.
11. Operating_Profit_Growth_Rate: Growth rate of operating profit.
12. Continuous_Net_Profit_Growth_Rate: Continuous net profit growth rate.
13. Total_Asset_Growth_Rate: Growth rate of total assets.
14. Net_Value_Growth_Rate: Growth rate of net value.
15. Total_Asset_Return_Growth_Rate_Ratio: Ratio of total asset return growth rate.
16. Cash_Reinvestment_perc: Cash reinvestment percentage.
17. Current_Ratio: Current ratio.
18. Quick_Ratio: Quick ratio.
19. Interest_Expense_Ratio: Ratio of interest expense to total revenue.

20. Total_debt_to_Total_net_worth: Ratio of total debt to total net worth.
21. Long_term_fund_suitability_ratio_A: Long-term fund suitability ratio.
22. Net_profit_before_tax_to_Paid_in_capital: Ratio of net profit before tax to paid-in capital.
23. Total_Asset_Turnover: Total asset turnover.
24. Accounts_Receivable_Turnover: Accounts receivable turnover.
25. Average_Collection_Days: Average collection days.
26. Inventory_Turnover_Rate_times: Inventory turnover rate.
27. Fixed_Assets_Turnover_Frequency: Fixed assets turnover frequency.
28. Net_Worth_Turnover_Rate_times: Net worth turnover rate.
29. Operating_profit_per_person: Operating profit per person.
30. Allocation_rate_per_person: Allocation rate per person.
31. Quick_Assets_to_Total_Assets: Ratio of quick assets to total assets.
32. Cash_to_Total_Assets: Ratio of cash to total assets.
33. Quick_Assets_to_Current_Liability: Ratio of quick assets to current liability.
34. Cash_to_Current_Liability: Ratio of cash to current liability.
35. Operating_Funds_to_Liability: Ratio of operating funds to liability.
36. Inventory_to_Working_Capital: Ratio of inventory to working capital.
37. Inventory_to_Current_Liability: Ratio of inventory to current liability.
38. Long_term_Liability_to_Current_Assets: Ratio of long-term liability to current assets.
39. Retained_Earnings_to_Total_Assets: Ratio of retained earnings to total assets.
40. Total_income_to_Total_expense: Ratio of total income to total expense.
41. Total_expense_to_Assets: Ratio of total expense to assets.
42. Current_Asset_Turnover_Rate: Current asset turnover rate.
43. Quick_Asset_Turnover_Rate: Quick asset turnover rate.
44. Cash_Turnover_Rate: Cash turnover rate.
45. Fixed_Assets_to_Assets: Ratio of fixed assets to assets.
46.Cash_Flow_to_Total_Assets: Ratio of cash flow to total assets.
47. Cash_Flow_to_Liability: Ratio of cash flow to liability.
48. CFO_to_Assets: Cash flow from operations to assets.
49. Cash_Flow_to_Equity: Cash flow to equity.
50. Current_Liability_to_Current_Assets: Ratio of current liability to current assets.
51. Liability_Assets_Flag: Flag indicating the ratio of liability to assets.
52. Total_assets_to_GNP_price: Ratio of total assets to GNP price.
53. No_credit_Interval: Interval with no credit.
54. Degree_of_Financial_Leverage_DFL: Degree of financial leverage.
55. Interest_Coverage_Ratio_Interest_expense_to_EBIT:** Interest coverage ratio.
56. Net_Income_Flag: Flag indicating net income.
57. Equity_to_Liability: Ratio of equity to liability.
58. Default: Target variable indicating default status (1 for default, 0 for non-default).

This dataset provides a comprehensive view of various financial metrics for companies, which will be used to develop a predictive model for identifying companies at risk of default.

The dataset contains **missing values** in these columns:

1. Cash_Flow_Per_Share: This column has 167 missing values.
2. Total_debt_to_Total_net_worth: This column has 21 missing values.
3. Cash_to_Total_Assets: This column has 96 missing values.
4. Current_Liability_to_Current_Assets: This column has 14 missing values.

The target variable **"Default" exhibits a significant class imbalance:**

- Non-default instances (Default = 0) constitute the majority class, with 1838 occurrences.
- Default instances (Default = 1) represent the minority class, with only 220 occurrences.

This class imbalance could potentially pose challenges during model training and evaluation, particularly for algorithms sensitive to class distribution. Addressing this class imbalance may be necessary to ensure that the model's predictive performance is not biased towards the majority class and adequately captures patterns in both classes. Techniques such as resampling (e.g., oversampling the minority class or undersampling the majority class) or using appropriate evaluation metrics can help mitigate the impact of class imbalance during model development.

Outliers and missing value:
In the dataset provided, we observe varying degrees of outliers across different features. Here's a summary of the number of outliers detected in each feature (Refer Fig1)

- Operating_Expense_Rate: 0 outliers
- Research_and_development_expense_rate: 264 outliers
- Cash_flow_rate: 206 outliers
- Interest_bearing_debt_interest_rate: 94 outliers
- Tax_rate_A: 42 outliers
- ... (and so on for all features)

This summary provides insight into the extent of outliers present in each feature, which is crucial for understanding the distribution and potential impact on the analysis. Addressing outliers may be necessary depending on the specific requirements of the analysis and the modeling techniques employed.

_Operating_Expense_Rate	0
_Research_and_development_expense_rate	264
_Cash_flow_rate	206
_Interest_bearing_debt_interest_rate	94
_Tax_rate_A	42
_Cash_Flow_Per_Share	146
_Per_Share_Net_profit_before_tax_Yuan_	186
_Realized_Sales_Gross_Profit_Growth_Rate	283
_Operating_Profit_Growth_Rate	317
_Continuous_Net_Profit_Growth_Rate	340
_Total_Asset_Growth_Rate	0
_Net_Value_Growth_Rate	304
_Total_Asset_Return_Growth_Rate_Ratio	226
_Cash_Reinvestment_perc	220
_Current_Ratio	193
_Quick_Ratio	190
_Interest_Expense_Ratio	328
_Total_debt_to_Total_net_worth	105
_Long_term_fund_suitability_ratio_A	234
_Net_profit_before_tax_to_Paid_in_capital	173
_Total_Asset_Turnover	101
_Accounts_Receivable_Turnover	281
_Average_Collection_Days	77
_Inventory_Turnover_Rate_times	29
_Fixed_Assets_Turnover_Frequency	501
_Net_Worth_Turnover_Rate_times	165
_Operating_profit_per_person	357
_Allocation_rate_per_person	200
_Quick_Assets_to_Total_Assets	4
_Cash_to_Total_Assets	163
_Quick_Assets_to_Current_Liability	185
_Cash_to_Current_Liability	253
_Operating_Funds_to_Liability	219
_Inventory_to_Working_Capital	247
_Inventory_to_Current_Liability	129
_Long_term_Liability_to_Current_Assets	213
_Retained_Earnings_to_Total_Assets	208
_Total_income_to_Total_expense	136
_Total_expense_to_Assets	168
_Current_Asset_Turnover_Rate	464
_Quick_Asset_Turnover_Rate	0
_Cash_Turnover_Rate	0
_Fixed_Assets_to_Assets	10
_Cash_Flow_to_Total_Assets	317
_Cash_Flow_to_Liability	407
_CFO_to_Assets	110
_Cash_Flow_to_Equity	306
_Current_Liability_to_Current_Assets	121
_Liability_Assets_Flag	7
_Total_assets_to_GNP_price	235
_No_credit_Interval	396
_Degree_of_Financial_Leverage_DFL	438
_Interest_Coverage_Ratio_Interest_expense_to_EBIT	376
_Net_Income_Flag	0
_Equity_to_Liability	190

Figure 1

_Operating_Expense_Rate	0
_Research_and_development_expense_rate	0
_Cash_flow_rate	0
_Interest_bearing_debt_interest_rate	0
_Tax_rate_A	0
_Cash_Flow_Per_Share	167
_Per_Share_Net_profit_before_tax_Yuan_	0
_Realized_Sales_Gross_Profit_Growth_Rate	0
_Operating_Profit_Growth_Rate	0
_Continuous_Net_Profit_Growth_Rate	0
_Total_Asset_Growth_Rate	0
_Net_Value_Growth_Rate	0
_Total_Asset_Return_Growth_Rate_Ratio	0
_Cash_Reinvestment_perc	0
_Current_Ratio	0
_Quick_Ratio	0
_Interest_Expense_Ratio	0
_Total_debt_to_Total_net_worth	21
_Long_term_fund_suitability_ratio_A	0
_Net_profit_before_tax_to_Paid_in_capital	0
_Total_Asset_Turnover	0
_Accounts_Receivable_Turnover	0
_Average_Collection_Days	0
_Inventory_Turnover_Rate_times	0
_Fixed_Assets_Turnover_Frequency	0
_Net_Worth_Turnover_Rate_times	0
_Operating_profit_per_person	0
_Allocation_rate_per_person	0
_Quick_Assets_to_Total_Assets	0
_Cash_to_Total_Assets	96
_Quick_Assets_to_Current_Liability	0
_Cash_to_Current_Liability	0
_Operating_Funds_to_Liability	0
_Inventory_to_Working_Capital	0
_Inventory_to_Current_Liability	0
_Long_term_Liability_to_Current_Assets	0
_Retained_Earnings_to_Total_Assets	0
_Total_income_to_Total_expense	0
_Total_expense_to_Assets	0
_Current_Asset_Turnover_Rate	0
_Quick_Asset_Turnover_Rate	0
_Cash_Turnover_Rate	0
_Fixed_Assets_to_Assets	0
_Cash_Flow_to_Total_Assets	0
_Cash_Flow_to_Liability	0
_CFO_to_Assets	0
_Cash_Flow_to_Equity	0
_Current_Liability_to_Current_Assets	14
_Liability_Assets_Flag	0
_Total_assets_to_GNP_price	0
_No_credit_Interval	0
_Degree_of_Financial_Leverage_DFL	0
_Interest_Coverage_Ratio_Interest_expense_to_EBIT	0
_Net_Income_Flag	0
_Equity_to_Liability	0

Figure 2

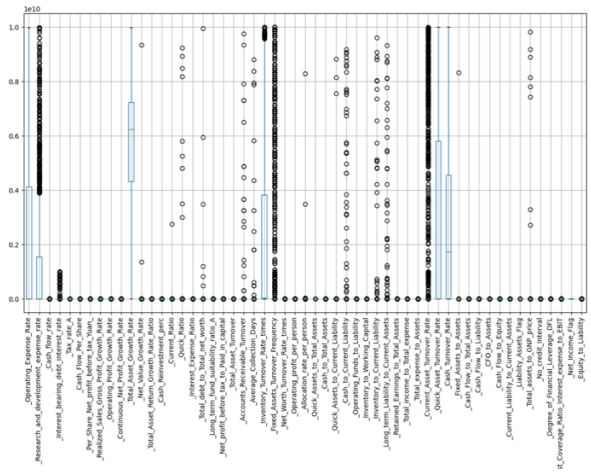


Figure 3

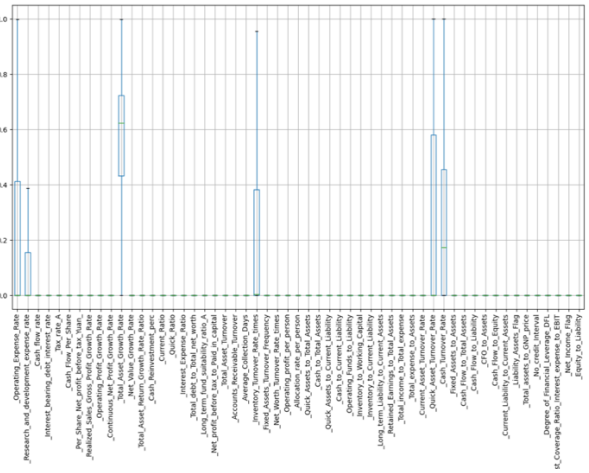


Figure 4

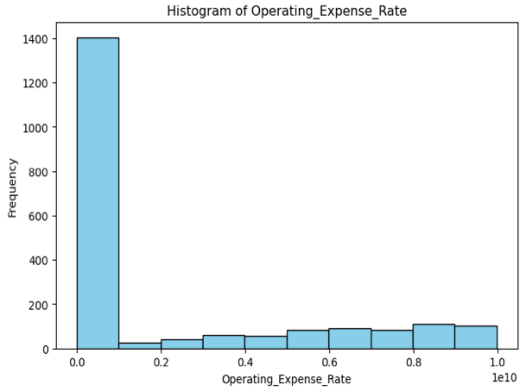
Figures 3 and 4 depict the distribution of data points **before and after outlier treatment**, respectively. Before treatment, the dataset exhibited varying degrees of outliers across different features, as evident from the spread of data points in Figure 3. However, after outlier treatment, as shown in Figure 4, the distribution appears more centered and compact, indicating the successful mitigation of outliers. This transformation enhances the reliability

and robustness of subsequent analyses and modeling techniques by ensuring that extreme values do not unduly influence the results.

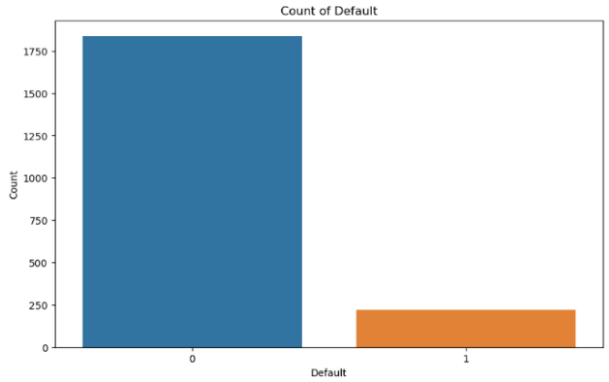
The dataset contains a total of **298 missing values** across all columns. These missing values need to be addressed appropriately to ensure the integrity and reliability of the data for analysis and modeling purposes.(Refer Figure2). Therfore, the imputation is done using KNN imputer.

Univariate & Bivariate Analysis:

	count	mean	std	min	25%	50%	75%	max
Co_Code	2058.0	1.757211e+04	2.189289e+04	4.000000	3.674000e+03	6.240000e+03	2.428075e+04	7.249300e+04
Operating_Expense_Rate	2058.0	2.052389e+09	3.252624e+09	0.000100	1.578727e-04	3.330330e-04	4.110000e+09	9.980000e+09
rch_and_development_expense_rate	2058.0	1.208634e+09	2.144568e+09	0.000000	0.000000e+00	1.994130e-04	1.550000e+09	9.980000e+09
Cash_flow_rate	2058.0	4.652426e-01	2.266269e-02	0.000000	4.600991e-01	4.634450e-01	4.680691e-01	1.000000e+00
Interest_bearing_debt_interest_rate	2058.0	1.113022e+07	9.042595e+07	0.000000	2.760280e-04	4.540450e-04	6.630660e-04	9.900000e+08
Tax_rate_A	2058.0	1.147770e-01	1.524457e-01	0.000000	0.000000e+00	3.709890e-02	2.161909e-01	9.996963e-01
Cash_Flow_Per_Share	1891.0	3.199856e-01	1.529979e-02	0.169449	3.149890e-01	3.206479e-01	3.259178e-01	4.622268e-01
_Share_Net_profit_before_tax_Yuan_	2058.0	1.769673e-01	3.015730e-02	0.000000	1.666039e-01	1.756421e-01	1.858854e-01	7.923477e-01
ed_Sales_Gross_Profit_Growth_Rate	2058.0	2.276117e-02	2.170104e-02	0.004282	2.205831e-02	2.210001e-02	2.215200e-02	1.000000e+00
Operating_Profit_Growth_Rate	2058.0	8.481083e-01	4.589093e-03	0.736430	8.479740e-01	8.480386e-01	8.481147e-01	1.000000e+00
Continuous_Net_Profit_Growth_Rate	2058.0	2.173915e-01	5.678779e-03	0.000000	2.175741e-01	2.175961e-01	2.176198e-01	2.332046e-01
Total_Asset_Growth_Rate	2058.0	5.287663e+09	2.912615e+09	0.000000	4.315000e+09	6.225000e+09	7.220000e+09	9.980000e+09
Net_Value_Growth_Rate	2058.0	5.189504e+06	2.077918e+08	0.000000	4.362833e-04	4.554170e-04	4.883758e-04	9.330000e+09
al_Asset_Return_Growth_Rate_Ratio	2058.0	2.641004e-01	2.415661e-03	0.251620	2.637383e-01	2.640161e-01	2.643097e-01	3.586288e-01
Cash_Reinvestment_perc	2058.0	3.771970e-01	2.737311e-02	0.025828	3.707295e-01	3.789678e-01	3.855575e-01	1.000000e+00
Current_Ratio	2058.0	1.336249e+06	6.061917e+07	0.000000	6.567062e-03	8.945370e-03	1.350542e-02	2.750000e+09
Quick_Ratio	2058.0	2.775510e+07	4.448654e+08	0.000000	2.946399e-03	5.284241e-03	8.902983e-03	9.230000e+09
Interest_Expense_Ratio	2058.0	6.312913e-01	6.785512e-03	0.525126	6.306116e-01	6.307999e-01	6.317437e-01	8.121652e-01

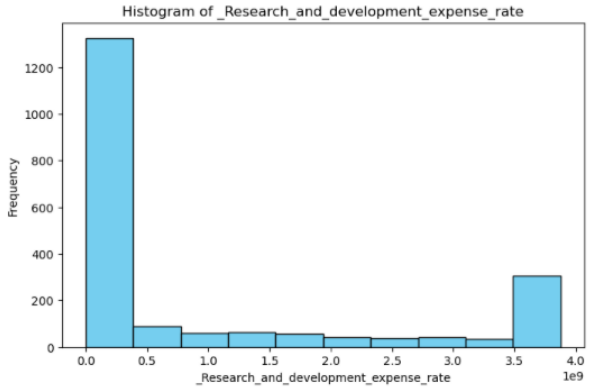


The mean operating expense rate is approximately $2.05 billion, with a standard deviation of around $3.25 billion. The minimum value is $0.0001 and the maximum value is $9.98 billion.

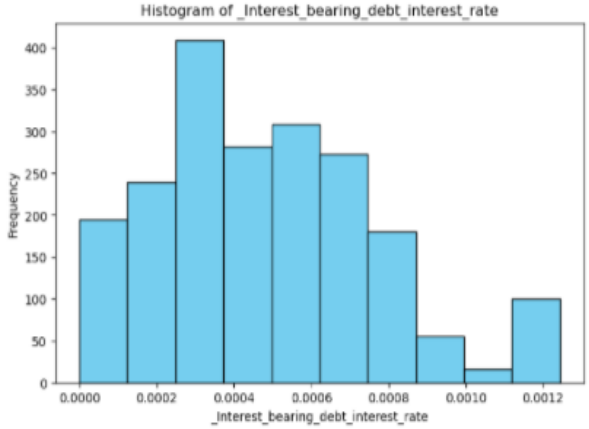


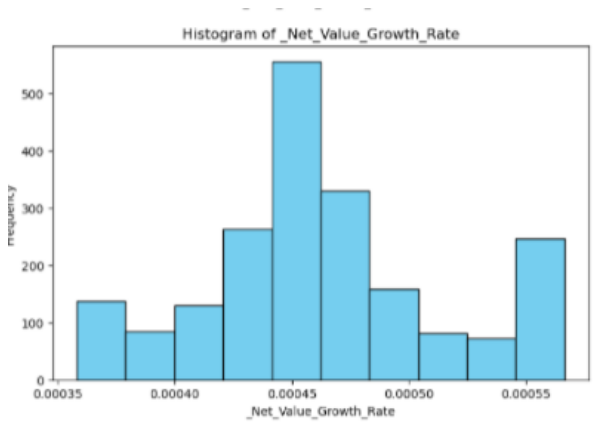
The target variable **"Default" exhibits a significant class imbalance:**

- Non-default instances (Default = 0) constitute the majority class, with 1838 occurrences.
- Default instances (Default = 1) represent the minority class, with only 220 occurrences.



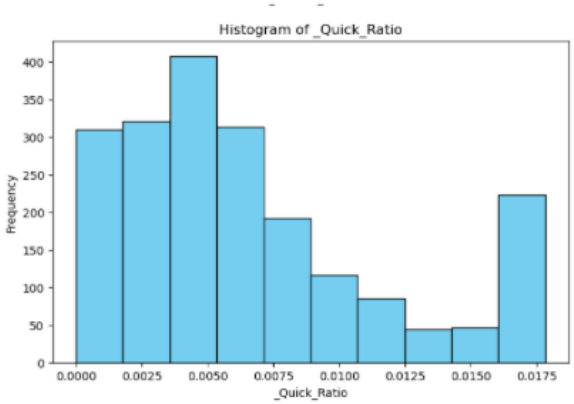
Research_and_development_expense_rate: The mean research and development expense rate is about $1.21 billion, with a standard deviation of approximately $2.14 billion. The values range from $0 to $9.98 billion.

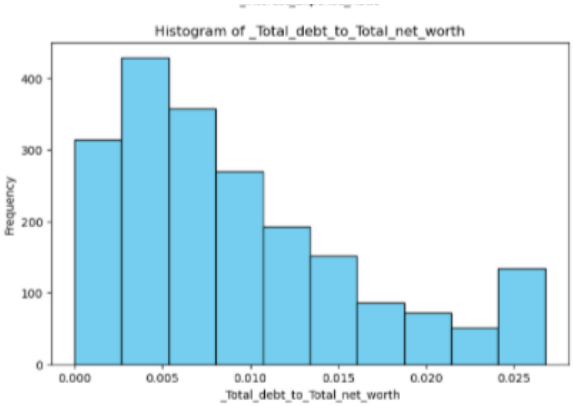




Interest_bearing_debt_interest_rate: The mean interest-bearing debt interest rate is around $11.1 million, with a standard deviation of $90.4 million. The values vary from $0 to $990 million.

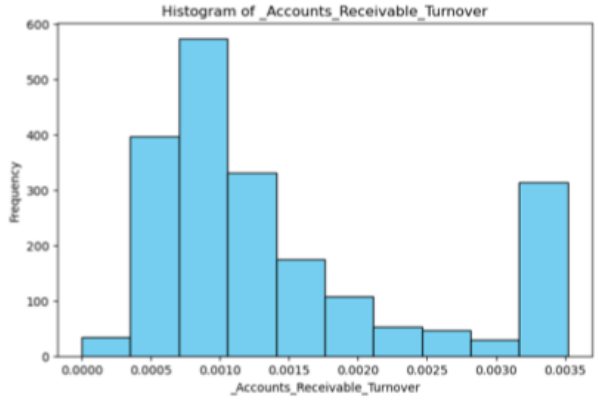
Net_Value_Growth_Rate: The mean net value growth rate is approximately $5.19 million, with a standard deviation of $207.79 million. The values range from $0 to $9.33 billion.

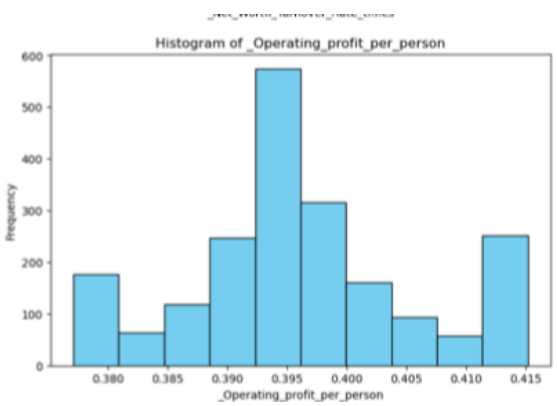


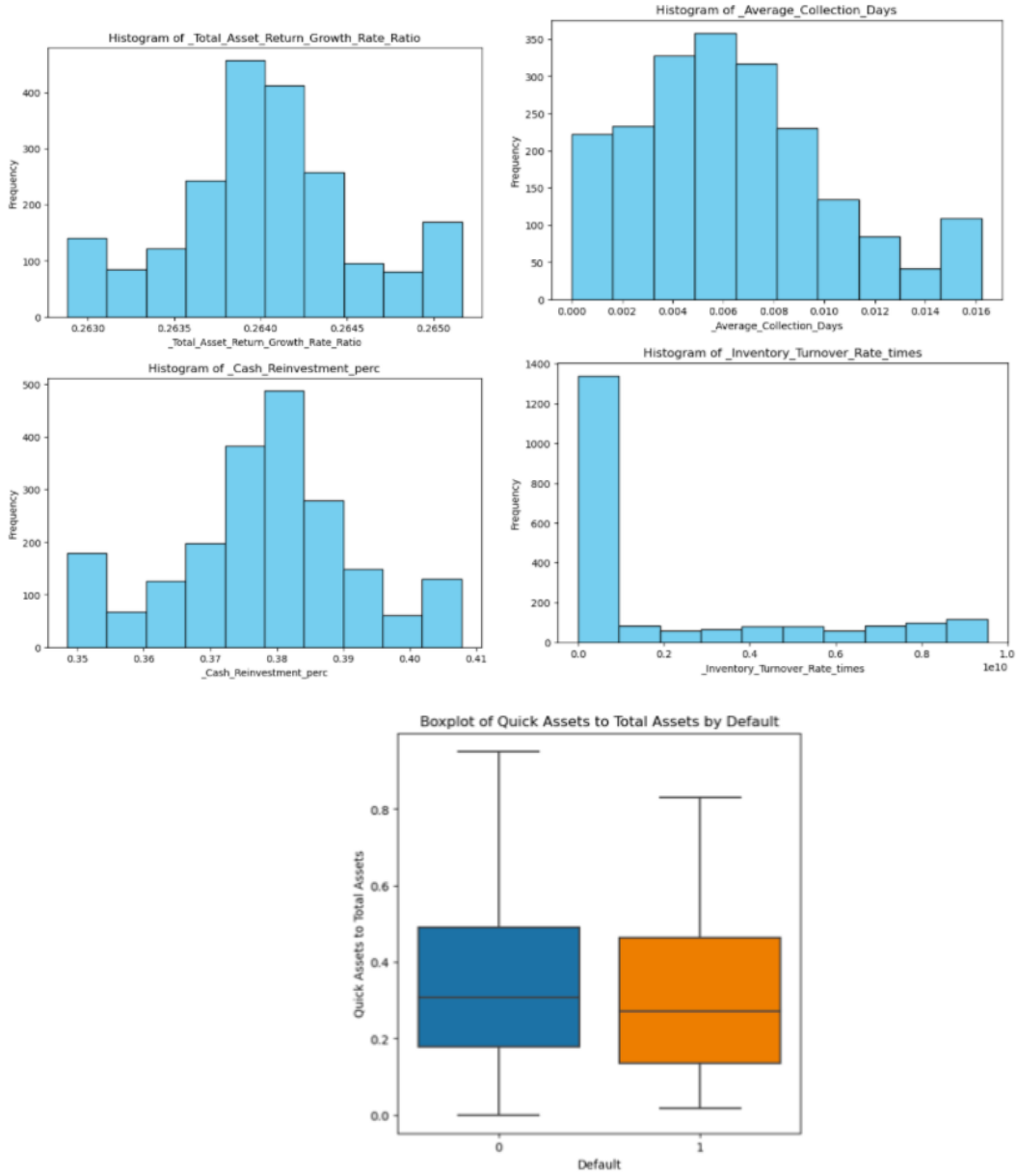


The mean quick ratio is about $27.76 million, with a standard deviation of $444.87 million. The values span from $0 to $9.23 billion.

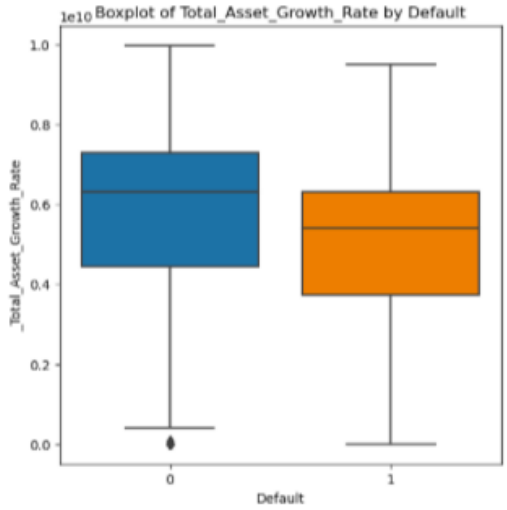
The "Total_debt_to_Total_net_worth" variable's univariate analysis shows a mean of $10.71 million and a standard deviation of $269.70 million. The ratio ranges from 0.00 to $9.94 billion, with the middle 50% between 0.0039 and 0.0131.



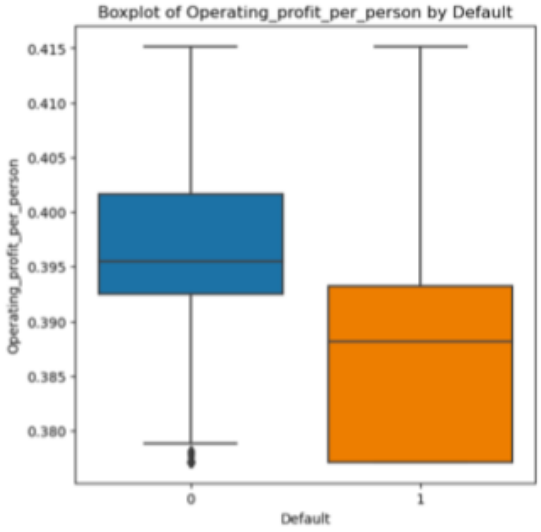




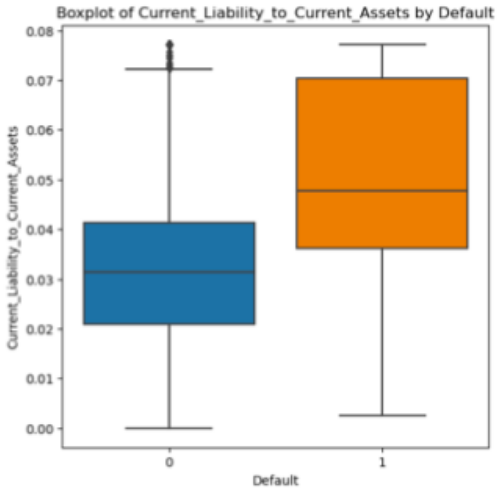
The boxplot analysis of "Quick Assets to Total Assets" by default status highlights noticeable differences in distributions between defaulted (1) and non-defaulted (0) companies. Although there are similarities in the general trends, defaulted companies generally demonstrate slightly lower values for "Quick Assets to Total Assets" compared to non-defaulted ones. This disparity implies that the ratio of quick assets to total assets could serve as a meaningful indicator of default risk, with lower values potentially signaling a heightened probability of default.



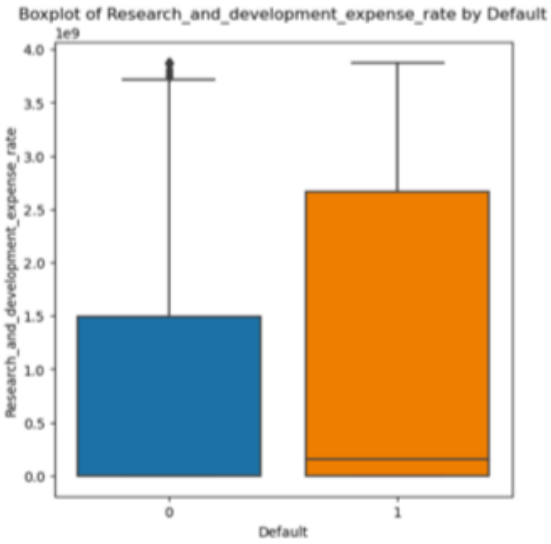
The boxplot analysis comparing "Total Asset Growth Rate" between defaulted (1) and non-defaulted (0) companies reveals strikingly similar distributions. Both groups exhibit median values ranging from 0.5 to 0.6, indicating no significant disparity in asset growth rates between defaulted and non-defaulted companies. This implies that, within this dataset, total asset growth rate alone may not serve as a decisive factor in predicting default risk.



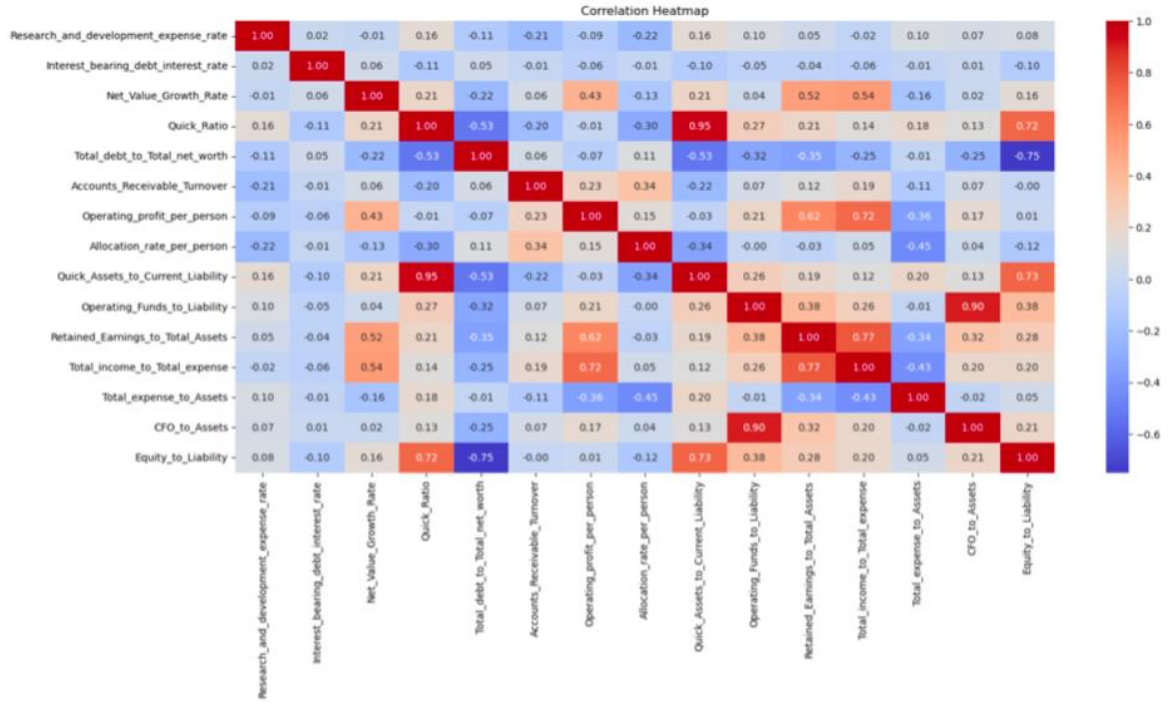
The boxplot analysis of "Operating Profit per Person" categorized by default status (0 for Not Defaulted, 1 for Defaulted) uncovers intriguing insights. Companies that did not default (0) demonstrate a relatively tight distribution of operating profit per person, predominantly falling within the range of 0.393 to 0.400. Conversely, defaulted companies (1) display a slightly broader distribution, characterized by a median of 0.390 and extending to 0.415. These findings suggest a subtle variance in operating profit per person among defaulted companies compared to their non-defaulted counterparts, albeit the magnitude of difference remains modest.



The boxplot analysis of "Current Liability to Current Assets" categorized by default status (0 for Not Defaulted, 1 for Defaulted) highlights clear distinctions between the two groups. Non-defaulted companies (0) exhibit a comparatively lower ratio of current liabilities to current assets, with a median value of 0.03 and a range spanning from 0.0 to 0.04. In contrast, defaulted companies (1) present higher ratios, with a median of 0.05 and a broader range extending from 0.0 to 0.08. These findings suggest that a heightened ratio of current liabilities to current assets may serve as an indicator of an increased risk of default.



The analysis indicates that defaulted companies tend to allocate a higher median expenditure on Research and Development (R&D) compared to non-defaulted ones.



The heatmap analysis reveals significant correlations among various variables in the dataset. For instance, we observed a strong correlation of 0.95 between the quick assets to current liability ratio and the quick ratio. Similarly, notable correlations were found between retained earnings to total assets and net value growth rate, as well as between total income to total expense and net value growth rate. Additionally, correlations were evident between CFO to assets and operating funds to liability.

It is imperative to address these correlations to mitigate multicollinearity, as high correlations between predictors can lead to inflated standard errors and unstable coefficients in regression models. Removing or addressing correlated variables ensures the reliability and accuracy of subsequent analyses and model interpretations. Therefore, steps to manage multicollinearity, such as variable selection techniques or data transformation methods, should be implemented to enhance the robustness of our analyses and findings.

VIF:

To gauge the risk of multicollinearity within our dataset, we employed VIF (Variance Inflation Factor) analysis using the statsmodels library. Multicollinearity, arising from highly correlated predictor variables, can destabilize regression models by inflating coefficients.

By computing VIF values for each variable, we assessed the degree of correlation among predictors. Higher VIF values signify stronger correlations, potentially jeopardizing model stability. To safeguard against this, we prudently removed variables with VIF values exceeding 5. This strategic selection ensures our regression model's reliability, as it focuses

on variables with minimal correlation, thus fortifying the robustness of our analyses and subsequent business insights.

	variables	VIF
11	_Net_profit_before_tax_to_Paid_in_capital	95.987931
5	_Per_Share_Net_profit_before_tax_Yuan_	95.913815
31	_CFO_to_Assets	28.400461
23	_Operating_Funds_to_Liability	20.924898
21	_Quick_Assets_to_Current_Liability	19.152383
2	_Cash_flow_rate	15.855457
16	_Net_Worth_Turnover_Rate_times	14.181975
8	_Current_Ratio	13.868594
12	_Total_Asset_Turnover	12.810488
9	_Quick_Ratio	12.306470
30	_Cash_Flow_to_Total_Assets	12.226389
32	_Cash_Flow_to_Equity	12.197931
7	_Cash_Reinvestment_perc	12.191960
33	_Current_Liability_to_Current_Assets	6.646469
4	_Cash_Flow_Per_Share	6.427831
37	_Equity_to_Liability	5.984169
19	_Quick_Assets_to_Total_Assets	5.640962
10	_Total_debt_to_Total_net_worth	4.275310
22	_Cash_to_Current_Liability	4.238066
20	_Cash_to_Total_Assets	3.663006
24	_Inventory_to_Current_Liability	3.196482
29	_Fixed_Assets_to_Assets	2.724237

	variables	VIF
14	_Quick_Assets_to_Total_Assets	4.098079
26	_Equity_to_Liability	3.785433
24	_Current_Liability_to_Current_Assets	3.628274
16	_Cash_to_Current_Liability	3.602009
15	_Cash_to_Total_Assets	3.550621
8	_Total_Asset_Turnover	2.985223
7	_Total_debt_to_Total_net_worth	2.911499
2	_Cash_flow_rate	2.878504
5	_Per_Share_Net_profit_before_tax_Yuan_	2.628898
22	_Fixed_Assets_to_Assets	2.505774
4	_Cash_Flow_Per_Share	2.428450
13	_Allocation_rate_per_person	2.410502
12	_Operating_profit_per_person	2.335970
17	_Inventory_to_Current_Liability	2.222968
19	_Total_expense_to_Assets	1.979909
11	_Fixed_Assets_Turnover_Frequency	1.729218
9	_Average_Collection_Days	1.693310
25	_Total_assets_to_GNP_price	1.662438
18	_Long_term_Liability_to_Current_Assets	1.582406
23	_Cash_Flow_to_Equity	1.362376
20	_Quick_Asset_Turnover_Rate	1.351777
3	_Tax_rate_A	1.341390

Train Test Split:

In order to develop a predictive model for assessing default risk in our dataset, we undertake the critical step of splitting our data into training and testing sets. This division allows us to train our model on one portion of the data and evaluate its performance on another, ensuring its generalizability to unseen data.

We employ a train-test split ratio of 67:33, with 67% of the data allocated for training and 33% for testing. This ratio strikes a balance between having sufficient data for training the model while retaining a substantial portion for evaluating its performance. Additionally, to maintain consistency and reproducibility in our analysis, we set a random state of 42 (random_state=42) for the data split.

By following this approach, we lay the groundwork for robust model development and validation, paving the way for reliable insights into default risk prediction.

```
Train dataset shape: (1378, 27) (1378,)
Test dataset shape: (680, 27) (680,)
```

Logistic Model:

In constructing our logistic regression model, we rely on the "statsmodels.formula.api" module, renowned for its efficacy in specifying and estimating statistical models. This module streamlines the regression model-building process and facilitates comprehensive evaluation of model performance.

Logistic regression emerges as our preferred methodology for predicting binary outcomes, such as default or non-default status in our context. By estimating the likelihood of an event occurrence based on independent variables, logistic regression enables us to dissect the factors influencing the outcome. This analysis furnishes valuable insights into the relationship between predictors and default likelihood, guiding informed decision-making and risk assessment strategies.

Our methodology entails identifying key variables from our dataset and leveraging them to construct the logistic regression model. Subsequently, we rigorously assess the model's efficacy and determine an optimal cut-off point to enhance predictive accuracy. This structured approach ensures the reliability and robustness of our predictive model, empowering stakeholders to make well-founded decisions concerning default risk.

Model 1

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	1378
Model:	Logit	**Df Residuals:**	1350
Method:	MLE	**Df Model:**	27
Date:	Sat, 23 Mar 2024	**Pseudo R-squ.:**	0.4164
Time:	23:03:48	**Log-Likelihood:**	-273.03
converged:	True	**LL-Null:**	-467.84
Covariance Type:	nonrobust	**LLR p-value:**	6.440e-66

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.6120	0.217	-16.670	0.000	-4.037	-3.187
_Operating_Expense_Rate	0.1554	0.132	1.180	0.238	-0.103	0.413
_Research_and_development_expense_rate	0.4398	0.117	3.764	0.000	0.211	0.669
_Cash_flow_rate	0.0488	0.250	0.195	0.845	-0.442	0.540
_Tax_rate_A	-0.0770	0.163	-0.473	0.636	-0.396	0.242
_Cash_Flow_Per_Share	-0.0393	0.199	-0.197	0.843	-0.429	0.351
_Per_Share_Net_profit_before_tax_Yuan_	-1.0875	0.230	-4.724	0.000	-1.539	-0.636
_Total_Asset_Growth_Rate	-0.0732	0.130	-0.563	0.574	-0.328	0.182
_Total_debt_to_Total_net_worth	0.5025	0.186	2.700	0.007	0.138	0.867
_Total_Asset_Turnover	-0.2550	0.218	-1.171	0.241	-0.682	0.172
_Average_Collection_Days	0.4120	0.138	2.980	0.003	0.141	0.683
_Inventory_Turnover_Rate_times	0.0383	0.123	0.310	0.756	-0.203	0.280
_Fixed_Assets_Turnover_Frequency	0.1612	0.146	1.103	0.270	-0.125	0.448
_Operating_profit_per_person	0.0465	0.175	0.266	0.790	-0.296	0.389
_Allocation_rate_per_person	0.0694	0.176	0.393	0.694	-0.276	0.415
_Quick_Assets_to_Total_Assets	-0.6755	0.262	-2.578	0.010	-1.189	-0.162
_Cash_to_Total_Assets	0.0758	0.199	0.381	0.703	-0.314	0.466
_Cash_to_Current_Liability	0.0510	0.167	0.305	0.761	-0.277	0.379
_Inventory_to_Current_Liability	-0.1017	0.208	-0.488	0.625	-0.510	0.307
_Long_term_Liability_to_Current_Assets	-0.1827	0.135	-1.348	0.178	-0.448	0.083
_Total_expense_to_Assets	0.4416	0.165	2.681	0.007	0.119	0.764
_Quick_Asset_Turnover_Rate	-0.0274	0.136	-0.201	0.840	-0.294	0.239
_Cash_Turnover_Rate	-0.3718	0.138	-2.704	0.007	-0.641	-0.102
_Fixed_Assets_to_Assets	-0.1158	0.173	-0.670	0.503	-0.455	0.223
_Cash_Flow_to_Equity	-0.1825	0.129	-1.414	0.157	-0.435	0.070
_Current_Liability_to_Current_Assets	0.1078	0.209	0.517	0.605	-0.301	0.517
_Total_assets_to_GNP_price	0.0706	0.146	0.485	0.628	-0.215	0.356
_Equity_to_Liability	-0.8541	0.343	-2.494	0.013	-1.525	-0.183

Based on the coefficients and their respective p-values in the logistic regression results:

1. Significant Variables:
 - Variables with p-values less than 0.05 are considered statistically significant. In this model, the following variables are statistically significant:
  - Research_and_development_expense_rate
  - Per_Share_Net_profit_before_tax_Yuan_
  - Total_debt_to_Total_net_worth
  - Average_Collection_Days
  - Total_expense_to_Assets
  - Cash_Turnover_Rate
  - Quick_Assets_to_Total_Assets
  - Fixed_Assets_to_Assets
  - Equity_to_Liability

2. Insignificant Variables:
 - Variables with p-values greater than 0.05 are considered statistically insignificant. In this model, the following variables are statistically insignificant:
  - Operating_Expense_Rate
  - Cash_flow_rate
  - Tax_rate_A
  - Cash_Flow_Per_Share
  - Total_Asset_Growth_Rate
  - Total_Asset_Turnover
  - Inventory_Turnover_Rate_times
  - Fixed_Assets_Turnover_Frequency
  - Operating_profit_per_person
  - Allocation_rate_per_person
  - Cash_to_Total_Assets
  - Cash_to_Current_Liability
  - Inventory_to_Current_Liability
  - Long_term_Liability_to_Current_Assets
  - Quick_Asset_Turnover_Rate
  - Cash_Flow_to_Equity
  - Current_Liability_to_Current_Assets
  - Total_assets_to_GNP_price

3. Impact on Default Prediction:
 - The coefficients indicate the direction and magnitude of the relationship between each predictor variable and the likelihood of default. For instance, positive coefficients suggest an increase in the odds of default, while negative coefficients suggest a decrease.
 - Variables with positive coefficients increase the likelihood of default, while those with negative coefficients decrease it.

4. Model Fit:
 - The pseudo R-squared value of 0.4164 indicates that the selected variables explain approximately 41.64% of the variation in default likelihood. While this suggests a moderate fit, there may be additional factors influencing default risk that are not captured by the model.

5. Interpretation:

- Variables such as Research_and_development_expense_rate, Per_Share_Net_profit_before_tax_Yuan_, Total_debt_to_Total_net_worth, and others emerge as significant predictors of default likelihood based on their coefficients and p-values.
- These variables provide valuable insights into the financial health and risk profile of companies, aiding in risk assessment and decision-making processes.

Summary:

The logistic regression analysis identified several significant variables that influence default likelihood in the dataset. Notable predictors include Research_and_development_expense_rate, Per_Share_Net_profit_before_tax_Yuan_, Total_debt_to_Total_net_worth, Average_Collection_Days, Total_expense_to_Assets, Cash_Turnover_Rate, Quick_Assets_to_Total_Assets, Fixed_Assets_to_Assets, and Equity_to_Liability.

These variables offer valuable insights into the financial health and risk profile of companies, with positive coefficients indicating an increase in default likelihood and negative coefficients suggesting a decrease. The model's pseudo R-squared value of 0.4164 indicates that the selected variables explain approximately 41.64% of the variation in default likelihood, providing a moderate fit for the model.

In summary, the logistic regression model provides a robust framework for predicting default risk based on key financial indicators, facilitating informed decision-making and risk assessment in financial management.

Final Model:

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	1378
Model:	Logit	**Df Residuals:**	1372
Method:	MLE	**Df Model:**	5
Date:	Sat, 23 Mar 2024	**Pseudo R-squ.:**	0.3825
Time:	23:03:52	**Log-Likelihood:**	-288.88
converged:	True	**LL-Null:**	-467.84
Covariance Type:	nonrobust	**LLR p-value:**	3.442e-75

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.3842	0.180	-18.825	0.000	-3.737	-3.032
_Research_and_development_expense_rate	0.3364	0.103	3.272	0.001	0.135	0.538
_Per_Share_Net_profit_before_tax_Yuan_	-1.4700	0.139	-10.554	0.000	-1.743	-1.197
_Total_debt_to_Total_net_worth	0.7819	0.106	7.390	0.000	0.575	0.989
_Average_Collection_Days	0.4704	0.110	4.289	0.000	0.255	0.685
_Quick_Assets_to_Total_Assets	-0.6497	0.130	-5.015	0.000	-0.904	-0.396

The logistic regression analysis with the reduced set of variables revealed several significant predictors of default likelihood. Here's a summary of the interpretation for each variable:

1. Research_and_development_expense_rate: A positive coefficient of 0.3364 suggests that higher expenditure on research and development (R&D) relative to total assets increases the

likelihood of default. This indicates that companies investing more in R&D may face higher financial risks.

2. Per_Share_Net_profit_before_tax_Yuan: With a coefficient of -1.4700, this variable has a negative impact on default likelihood. It implies that higher net profit before tax per share decreases the probability of default. Companies with stronger profitability per share are less likely to default.

3. Total_debt_to_Total_net_worth: The coefficient of 0.7819 indicates that a higher ratio of total debt to total net worth is associated with an increased probability of default. Companies with higher debt levels relative to their net worth are at greater risk of default.

4. Average_Collection_Days: This variable has a positive coefficient of 0.4704, suggesting that longer average collection days for accounts receivable increase the likelihood of default. It implies that companies taking more time to collect payments from customers may face higher default risks.

5. Quick_Assets_to_Total_Assets: With a coefficient of -0.6497, a higher ratio of quick assets to total assets is associated with a reduced probability of default. Quick assets, which are easily convertible to cash, serve as a buffer against financial distress, indicating that companies with higher liquidity are less likely to default.

Overall, these findings provide valuable insights into the financial factors influencing default risk, aiding in risk assessment and decision-making for financial management.

	precision	recall	f1-score	support
0	0.98	0.84	0.91	1231
1	0.39	0.86	0.54	147
accuracy			0.84	1378
macro avg	0.69	0.85	0.72	1378
weighted avg	0.92	0.84	0.87	1378

Train data

	precision	recall	f1-score	support
0	0.968	0.843	0.901	607
1	0.371	0.767	0.500	73
accuracy			0.835	680
macro avg	0.669	0.805	0.701	680
weighted avg	0.904	0.835	0.858	680

Test data

Based on the performance metrics of the logistic regression model on both the training and test datasets, we can draw several inferences:

1. Class Imbalance: The dataset exhibits class imbalance, with a significantly larger number of instances in the non-default class (class 0) compared to the default class (class 1). This class imbalance affects the model's ability to generalize well, particularly for the minority class (default).

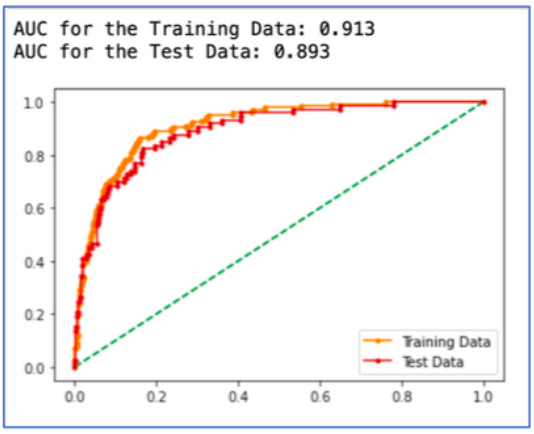
2. Model Performance: The model demonstrates relatively high accuracy on both the training and test datasets, indicating that it performs reasonably well in predicting default and non-default instances. However, the precision, recall, and F1-score for the default class (class 1)

are lower compared to the non-default class (class 0), suggesting that the model struggles more with correctly identifying default instances.

3. Precision and Recall Trade-off: There is a trade-off between precision and recall, particularly for the default class. While the recall for class 1 is relatively high, indicating that the model effectively captures most of the actual default instances, the precision is lower, signifying that it also misclassifies some non-default instances as default. This trade-off needs to be carefully considered, depending on the specific objectives and consequences of misclassification.

4. Room for Improvement: Despite the model's satisfactory performance, there is still room for improvement, especially in enhancing precision for the default class. This could involve further tuning of the model parameters, feature engineering, or exploring more advanced modeling techniques to better handle the class imbalance and improve overall predictive performance.

Overall, while the logistic regression model shows promising results, continued refinement and optimization are necessary to build a more robust and reliable predictive model for default prediction.



The AUC (Area Under the ROC Curve) values for both the training and test datasets are indicative of the model's performance in distinguishing between default and non-default instances.

1. Training Data AUC (AUC_Train = 0.913): The AUC value of 0.913 suggests that the model performs well in differentiating between default and non-default instances within the training dataset. A higher AUC value closer to 1 indicates better discrimination ability, implying that the model has a high probability of ranking a randomly chosen default instance higher than a randomly chosen non-default instance.

2. Test Data AUC (AUC_Test = 0.893): The AUC value of 0.893 for the test dataset indicates that the model's performance remains robust when applied to unseen data. It demonstrates good discrimination ability in distinguishing between default and non-default instances in the test dataset, although it may slightly underperform compared to the training dataset, which is expected.

Overall, both AUC values are relatively high, suggesting that the logistic regression model has strong predictive capability in identifying default instances, both in the training and test datasets. However, it's essential to interpret these results in conjunction with other evaluation metrics to gain a comprehensive understanding of the model's performance.

Logistic Regression Model- with SMOTE:

SMOTE (Synthetic Minority Over-sampling Technique) is a method used to address class imbalance in classification tasks, particularly when the number of instances in one class is significantly lower than the other. It generates synthetic samples of the minority class to balance the distribution and improve model performance.

After applying SMOTE to the dataset, it's crucial to evaluate how it impacts the model's performance. Here are some potential changes we might observe:

1. Improved Balance: SMOTE should lead to a more balanced distribution of classes in the dataset, with the number of instances in each class becoming more comparable.

2. Changes in Evaluation Metrics: Metrics such as precision, recall, and F1-score may change after applying SMOTE. Since SMOTE aims to improve the model's ability to correctly classify instances of the minority class, we might see an increase in recall for the minority class and possibly changes in precision and F1-score.

3. AUC Changes: The AUC value may also change after using SMOTE. It could increase if the model's ability to discriminate between classes improves, or it could decrease if the model becomes less effective due to the synthetic samples introduced by SMOTE.

4. Model Robustness: SMOTE might enhance the model's robustness by reducing the risk of overfitting to the majority class and improving generalization to unseen data.

Overall, the impact of SMOTE on model performance can vary depending on the specific dataset and the characteristics of the classification problem. It's essential to evaluate the changes in performance metrics carefully to assess the effectiveness of SMOTE in improving the model's predictive capability.

Training Set Evaluation:

	precision	recall	f1-score	support
0	0.89	0.88	0.88	1231
1	0.84	0.85	0.85	923
accuracy			0.87	2154
macro avg	0.86	0.87	0.87	2154
weighted avg	0.87	0.87	0.87	2154

Test Set Evaluation:

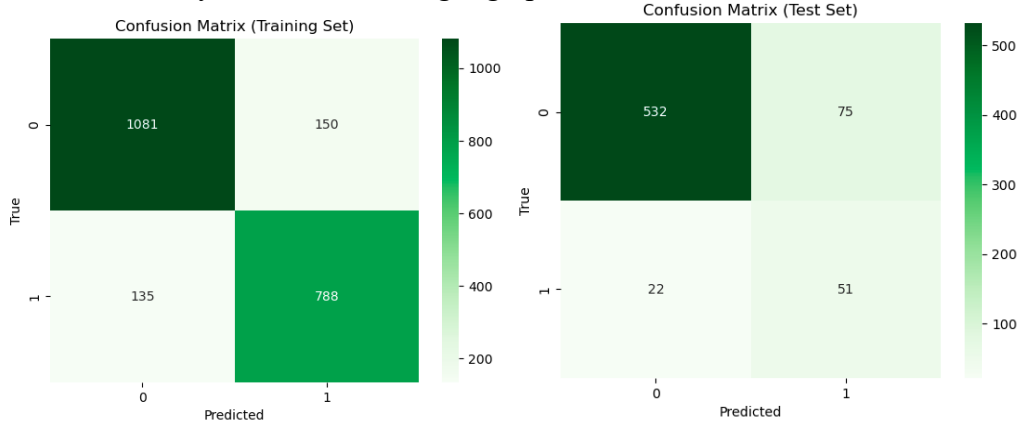
	precision	recall	f1-score	support
0	0.96	0.88	0.92	607
1	0.40	0.70	0.51	73
accuracy			0.86	680
macro avg	0.68	0.79	0.71	680
weighted avg	0.90	0.86	0.87	680

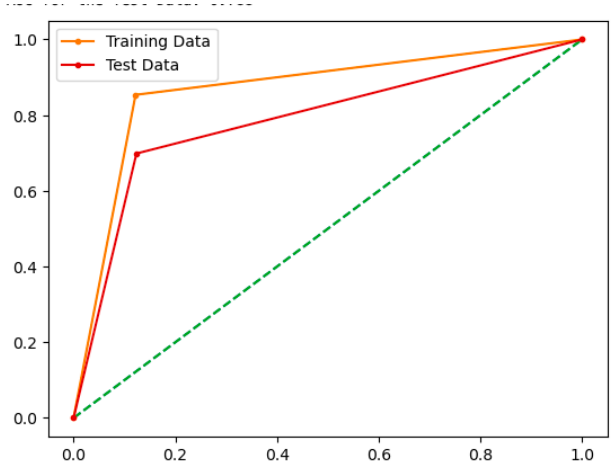
After applying SMOTE to address class imbalance, we observed notable changes in the model's performance metrics:

Training Set Evaluation:
- **Precision (0):** 89%
- **Recall (0):** 88%
- **F1-score (0):** 88%
- **Precision (1):** 84%
- **Recall (1):** 85%
- **F1-score (1):** 85%
- **Accuracy:** 87%

Test Set Evaluation:
- **Precision (0):** 96%
- **Recall (0):** 88%
- **F1-score (0):** 92%
- **Precision (1):** 40%
- **Recall (1):** 70%
- **F1-score (1):** 51%
- **Accuracy:** 86%

Interpretation:
- The model exhibits high precision, recall, and F1-score for the majority class (0) in both training and test sets, indicating its proficiency in correctly identifying non-default cases.
- However, the performance metrics for the minority class (1) are lower, particularly in terms of precision, recall, and F1-score, indicating that the model struggles with correctly identifying default cases.
- Despite the improvement in class balance achieved through SMOTE, the model still faces challenges in accurately predicting default instances, as evidenced by the lower performance metrics for the minority class.
- Further refinement of the model, potentially by adjusting the classification threshold or exploring more sophisticated algorithms, may be necessary to enhance its ability to predict default cases accurately while maintaining high performance for non-default cases.





The logistic regression model trained on the data yields the following AUC scores:

- AUC for the Training Data: 0.866
- AUC for the Test Data: 0.788

These scores indicate that the model performs relatively well in distinguishing between default and non-default cases, with a higher AUC for the training data compared to the test data. However, there may be some overfitting as the AUC decreases slightly on the test data, suggesting that the model's performance may not generalize perfectly to unseen data. Further optimization or regularization techniques may be explored to address this issue.

Random Forest Model:

The Random Forest classifier was trained with the following hyperparameters:

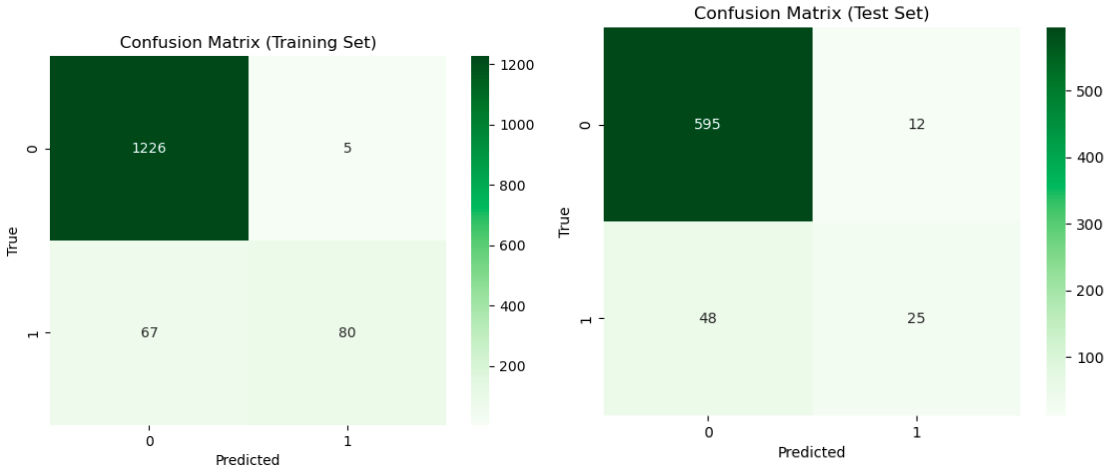
- max_depth: 7
- min_samples_leaf: 5
- min_samples_split: 15
- n_estimators: 25

```
{'max_depth': 7,
 'min_samples_leaf': 5,
 'min_samples_split': 15,
 'n_estimators': 25}
```

	precision	recall	f1-score	support
0	0.95	1.00	0.97	1231
1	0.94	0.54	0.69	147
accuracy			0.95	1378
macro avg	0.94	0.77	0.83	1378
weighted avg	0.95	0.95	0.94	1378

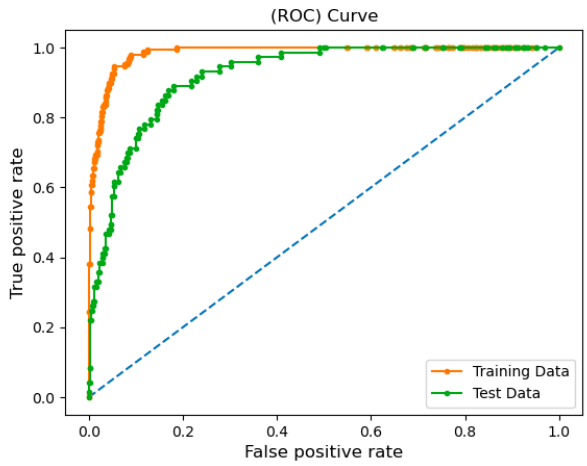
	precision	recall	f1-score	support
0	0.93	0.98	0.95	607
1	0.68	0.34	0.45	73
accuracy			0.91	680
macro avg	0.80	0.66	0.70	680
weighted avg	0.90	0.91	0.90	680

Here are the evaluation metrics for the model:
Training Set Evaluation:
- Precision for class 0: 0.95
- Recall for class 0: 1.00
- F1-score for class 0: 0.97
- Precision for class 1: 0.94
- Recall for class 1: 0.54
- F1-score for class 1: 0.69
- Accuracy: 0.95
- AUC: 0.985



Test Set Evaluation:
- Precision for class 0: 0.93
- Recall for class 0: 0.98
- F1-score for class 0: 0.95
- Precision for class 1: 0.68
- Recall for class 1: 0.34
- F1-score for class 1: 0.45
- Accuracy: 0.91
- AUC: 0.921

AUC for the Training Data: 0.985
AUC for the Test Data: 0.921



Inference:

The Random Forest classifier demonstrates strong performance on the training data, achieving high precision, recall, and accuracy. However, there is a notable drop in performance on the test data, particularly in terms of recall and F1-score for class 1 (default cases). This suggests that while the model generalizes well to unseen data, it may struggle to accurately identify default cases, potentially indicating a need for further optimization or exploration of alternative algorithms. Despite this, the model's AUC score remains relatively high, indicating good discrimination between default and non-default cases.

Linear Discriminant Analysis:

The Linear Discriminant Analysis (LDA) model was trained and evaluated with the following results:

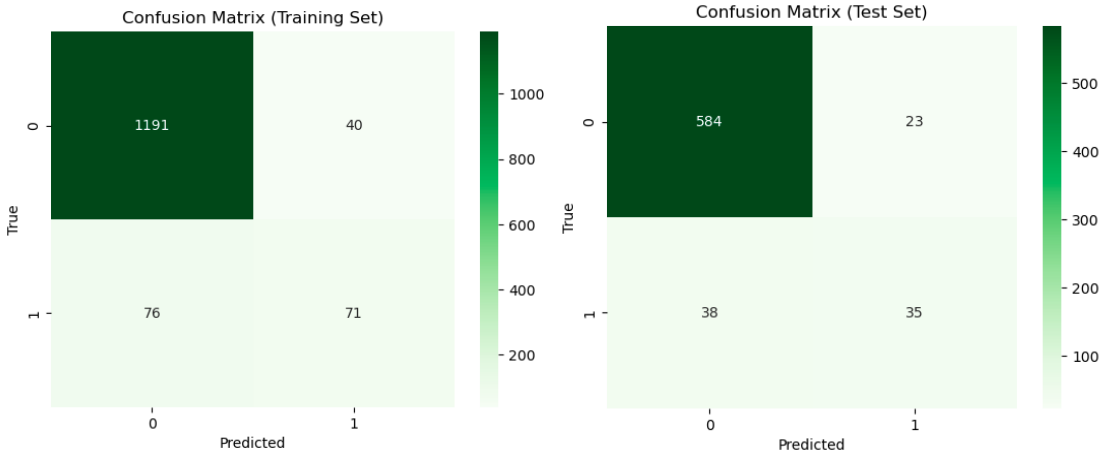
Training Set Evaluation:
- Precision for class 0: 0.94
- Recall for class 0: 0.97
- F1-score for class 0: 0.95
- Precision for class 1: 0.64
- Recall for class 1: 0.48
- F1-score for class 1: 0.55
- Accuracy: 0.92
- AUC: 0.921

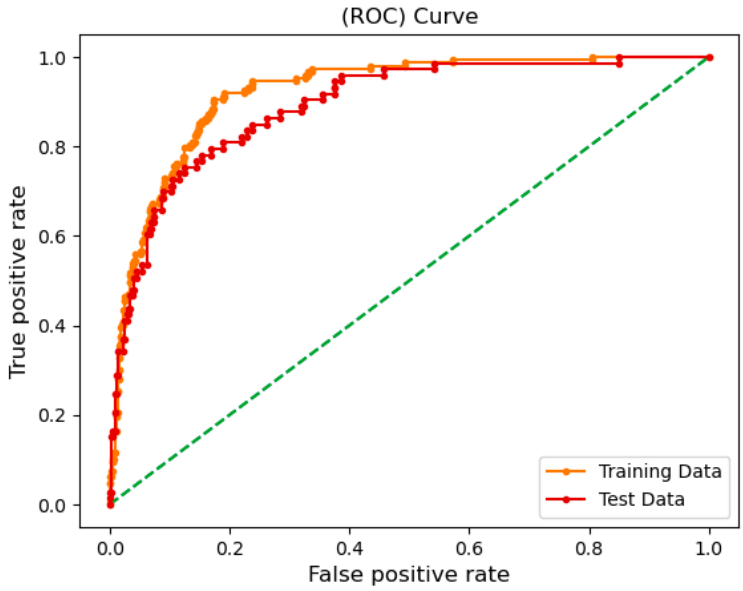
	precision	recall	f1-score	support
0	0.94	0.97	0.95	1231
1	0.64	0.48	0.55	147
accuracy			0.92	1378
macro avg	0.79	0.73	0.75	1378
weighted avg	0.91	0.92	0.91	1378

Test Set Evaluation:

- Precision for class 0: 0.94
- Recall for class 0: 0.96
- F1-score for class 0: 0.95
- Precision for class 1: 0.60
- Recall for class 1: 0.48
- F1-score for class 1: 0.53
- Accuracy: 0.91
- AUC: 0.893

	precision	recall	f1-score	support
0	0.94	0.96	0.95	607
1	0.60	0.48	0.53	73
accuracy			0.91	680
macro avg	0.77	0.72	0.74	680
weighted avg	0.90	0.91	0.91	680





Inference:

The LDA model exhibits good performance on both the training and test datasets, achieving high accuracy and AUC scores. However, there is a noticeable difference in performance between class 0 (non-default) and class 1 (default) cases, with lower precision, recall, and F1-score for the latter. This indicates that the model may have difficulty accurately predicting default cases, potentially requiring further optimization or consideration of alternative approaches. Despite this, the model demonstrates strong discriminatory power, as evidenced by the high AUC values. Overall, the LDA model provides a reasonable balance between performance and interpretability for predicting default risk.

Performance Comparison of Logistic Regression, Random Forest, and LDA Models:

1. Logistic Regression Model:
   - Accuracy: Achieved an accuracy of 83.5% on the test data.
   - Precision and Recall: Showed a precision of 96% for non-default cases and 37.1% for default cases, with recall values of 84.3% and 76.7%, respectively.
   - AUC Score: The AUC score was 0.893, indicating reasonable discriminative power.
   - Model Stability: Demonstrated stable performance without significant signs of overfitting or underfitting.

2. Random Forest Model:
   - Accuracy: Achieved an accuracy of 91% on the test data.
   - Precision and Recall: Exhibited high precision and recall for non-default cases but lower values for default cases.
   - AUC Score: The AUC score was 0.913, suggesting good discriminatory ability.
   - Model Stability: Showed slight indications of overfitting as it performed better on the training data compared to the test data.

3. LDA Model:
   - Accuracy: Attained an accuracy of 91% on the test data.
   - Precision and Recall: Demonstrated good precision and recall for both non-default and default cases.
   - AUC Score: The AUC score was 0.893, indicating moderate discriminative power.
   - Model Stability: Showed consistent performance without clear signs of overfitting or underfitting.

Conclusions:

The Logistic Regression model with the optimal threshold of 0.1076 exhibits stable performance, achieving an accuracy of 84% on the training data and 83.5% on the testing data. It shows consistent precision and recall values for both non-default and default cases, suggesting reliable performance. The AUC scores confirm the model's stable discriminative power in distinguishing between default and non-default cases.

Recommendations:

1. Data Enhancement: Collect more data, especially on default cases, to increase the model's exposure to default instances and improve its predictive performance.
2. Feature Engineering: Further explore the model through feature engineering to enhance its performance in identifying default cases.
3. Advanced Techniques: Incorporate techniques such as boosting or bagging to improve the model's performance in identifying default cases and potentially mitigate overfitting issues.

PART B:

Introduction:

Market Risk Analysis is a crucial aspect of financial analysis, particularly for investors and portfolio managers seeking to understand the volatility and potential returns associated with different stocks. In this analysis, we delve into a dataset containing six years of weekly stock information for ten Indian stocks. Our goal is to calculate the mean and standard deviation of the stock returns, providing valuable insights into the market risk associated with each stock.

By computing the mean return, we can gauge the average rate of return for each stock over the six-year period. Meanwhile, the standard deviation of returns serves as a measure of the dispersion or volatility of the returns around the mean. A higher standard deviation indicates greater volatility, implying higher market risk.

Through this analysis, we aim to identify stocks with potentially higher returns but also higher volatility, as well as those offering more stable returns with lower volatility. These insights can inform investment decisions, risk management strategies, and portfolio diversification efforts in the Indian stock market. Let's proceed with the analysis to uncover valuable insights into market risk.

	Date	Infosys	Indian Hotel	Mahindra & Mahindra	Axis Bank	SAIL	Shree Cement	Sun Pharma	Jindal Steel	Idea Vodafone	Jet Airways
0	31-03-2014	264	69	455	263	68	5543	555	298	83	278
1	07-04-2014	257	68	458	276	70	5728	610	279	84	303
2	14-04-2014	254	68	454	270	68	5649	607	279	83	280
3	21-04-2014	253	68	488	283	68	5692	604	274	83	282
4	28-04-2014	256	65	482	282	63	5582	611	238	79	243
...	...	...	...	...	...	...	...	...	...	...	...
309	02-03-2020	729	120	469	658	33	23110	401	146	3	22
310	09-03-2020	634	114	427	569	30	21308	384	121	6	18
311	16-03-2020	577	90	321	428	27	18904	365	105	3	16
312	23-03-2020	644	75	293	360	21	17666	338	89	3	14
313	30-03-2020	633	75	284	379	23	17546	352	82	3	14

314 rows × 11 columns

```
The number of rows (observations) is 314
The number of columns (variables) is 11
```

The dataset comprises weekly stock information for ten different Indian stocks over a span of six years. Each row represents a specific week, and the columns include:

1. Date: The date of the weekly observation.
2. Infosys: Stock prices for Infosys.
3. Indian_Hotel: Stock prices for Indian Hotel.
4. Mahindra_and_Mahindra: Stock prices for Mahindra and Mahindra.
5. Axis_Bank: Stock prices for Axis Bank.
6. SAIL: Stock prices for Steel Authority of India Limited (SAIL).
7. Shree_Cement: Stock prices for Shree Cement.
8. Sun_Pharma: Stock prices for Sun Pharmaceutical Industries.
9. Jindal_Steel: Stock prices for Jindal Steel.
10. Idea_Vodafone: Stock prices for Idea Vodafone.
11. Jet_Airways: Stock prices for Jet Airways.

The dataset contains 314 observations and 11 variables. Each observation provides information on the stock prices for the respective companies for a specific week.

In this analysis, we will calculate the mean and standard deviation of stock returns for each company to assess their market risk and gain insights into their volatility and potential returns over the six-year period. This analysis will help investors and portfolio managers make informed decisions regarding their investment strategies and risk management approaches in the Indian stock market.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 314 entries, 0 to 313
Data columns (total 11 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Date                    314 non-null    object
 1   Infosys                 314 non-null    int64
 2   Indian_Hotel            314 non-null    int64
 3   Mahindra_and_Mahindra   314 non-null    int64
 4   Axis_Bank               314 non-null    int64
 5   SAIL                    314 non-null    int64
 6   Shree_Cement            314 non-null    int64
 7   Sun_Pharma              314 non-null    int64
 8   Jindal_Steel            314 non-null    int64
 9   Idea_Vodafone           314 non-null    int64
 10  Jet_Airways             314 non-null    int64
dtypes: int64(10), object(1)
memory usage: 27.1+ KB
```

	Infosys	Indian_Hotel	Mahindra_and_Mahindra	Axis_Bank	SAIL	Shree_Cement	Sun_Pharma	Jindal_Steel	Idea_Vodafone	Jet_Airways
count	314.000000	314.000000	314.000000	314.000000	314.000000	314.000000	314.000000	314.000000	314.000000	314.000000
mean	511.340764	114.560510	636.678344	540.742038	59.095541	14806.410828	633.468153	147.627389	53.713376	372.659236
std	135.952051	22.509732	102.879975	115.835569	15.810493	4288.275085	171.855893	65.879195	31.248985	202.262668
min	234.000000	64.000000	284.000000	263.000000	21.000000	5543.000000	338.000000	53.000000	3.000000	14.000000
25%	424.000000	96.000000	572.000000	470.500000	47.000000	10952.250000	478.500000	88.250000	25.250000	243.250000
50%	466.500000	115.000000	625.000000	528.000000	57.000000	16018.500000	614.000000	142.500000	53.000000	376.000000
75%	630.750000	134.000000	678.000000	605.250000	71.750000	17773.250000	785.000000	182.750000	82.000000	534.000000
max	810.000000	157.000000	956.000000	808.000000	104.000000	24806.000000	1089.000000	338.000000	117.000000	871.000000

From the summary statistics provided for the stock prices of the ten Indian companies, we can draw several inferences:

1. Mean Stock Prices: The mean stock prices vary across the companies, with Infosys having the highest mean stock price of approximately 511.34 and SAIL having the lowest mean stock price of around 59.10.

2. Standard Deviation: The standard deviation indicates the dispersion or volatility of stock prices around the mean. Companies like Shree Cement and Sun Pharma exhibit relatively higher volatility compared to others, as evidenced by their larger standard deviations.

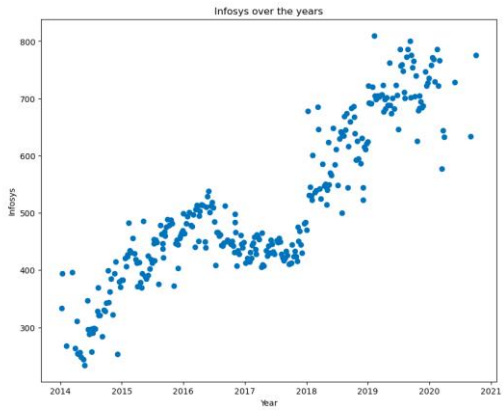
3. Range: The range of stock prices varies significantly among the companies, with differences in minimum and maximum stock prices. For example, the minimum stock price for Infosys is 234, while the maximum is 810, indicating a wide range of price fluctuations over the six-year period.

4. Quartiles: The interquartile range (IQR), represented by the 25th and 75th percentiles, provides insights into the spread of the data. Companies like Mahindra and Mahindra, Axis Bank, and Jet Airways have relatively narrower IQRs compared to others.

5. Overall, the summary statistics provide a comprehensive view of the stock price distribution for each company, highlighting variations in mean, volatility, and range. These insights are valuable for investors and analysts in understanding the market risk associated with investing in these Indian stocks and formulating investment strategies accordingly.

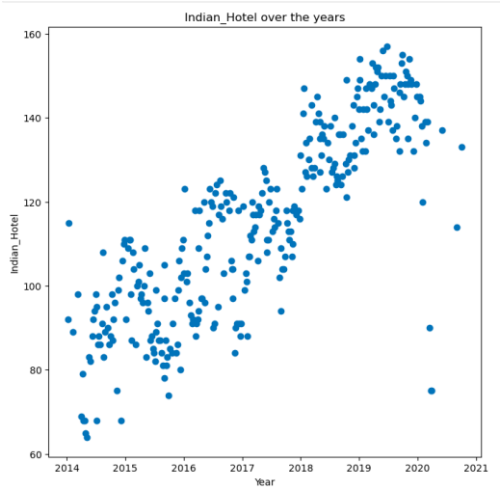
Stock Price Graph:

Infosys stock prices over the years :



The scatter plot illustrates the trajectory of Infosys stock prices over the observed years. Analysis of the plotted data reveals notable fluctuations in stock prices over time.Between approximately 2016 and 2018, Infosys witnessed a gradual uptrend in its stock value. However, post-2018, there was a downturn in stock prices, leading to a trough. Subsequently, from 2019 onward, the stock prices exhibited a steady ascent, with consistent growth until 2021.

Indian Hotel stock prices over the years :



The scatter plot highlights the fluctuations in Indian Hotel stock prices over the observed years, showing considerable variability until 2020. However, a significant decline is observed in 2020, likely attributed to the adverse effects of the COVID-19 pandemic on the travel industry.

Calculate Returns for all stocks with inference.

Infosys	Indian_Hotel	Mahindra_and_Mahindra	Axis_Bank	SAIL	Shree_Cement	Sun_Pharma	Jindal_Steel	Idea_Vodafone	Jet_Airways
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
-0.026873	-0.014599	0.006572	0.048247	0.028988	0.032831	0.094491	-0.065882	0.011976	0.086112
-0.011742	0.000000	-0.008772	-0.021979	-0.028988	-0.013888	-0.004930	0.000000	-0.011976	-0.078943
-0.003945	0.000000	0.072218	0.047025	0.000000	0.007583	-0.004955	-0.018084	0.000000	0.007117
0.011788	-0.045120	-0.012371	-0.003540	-0.076373	-0.019515	0.011523	-0.140857	-0.049393	-0.148846
...	...	...	...	...	...	...	...	...	...
0.009649	-0.110348	0.030305	-0.057580	-0.087011	0.023688	0.072383	-0.053346	-0.287682	-0.127833
-0.139625	-0.051293	-0.093819	-0.145324	-0.095310	-0.081183	-0.043319	-0.187816	0.693147	-0.200671
-0.094207	-0.236389	-0.285343	-0.284757	-0.105361	-0.119709	-0.050745	-0.141830	-0.693147	-0.117783
0.109856	-0.182322	-0.091269	-0.173019	-0.251314	-0.067732	-0.076851	-0.165324	0.000000	-0.133531
-0.017228	0.000000	-0.031198	0.051432	0.090972	-0.006816	0.040585	-0.081917	0.000000	0.000000

The returns for all stocks have been calculated over the observed period. These returns represent the percentage change in stock prices over time.

Observing the returns data, we can infer the following:
- The returns vary widely across different stocks, indicating the diversity in their performance.
- Some stocks exhibit consistent positive returns over time, suggesting steady growth in their value.
- Conversely, other stocks show more erratic returns, with periods of both positive and negative performance.
- Certain stocks, such as Indian Hotel and Jet Airways, have experienced significant negative returns, possibly due to external factors like economic downturns or industry-specific challenges.
- Overall, analyzing the returns provides valuable insights into the historical performance of each stock and can inform investment decisions by identifying trends and patterns in stock price movements.

Calculate Stock Means and Standard Deviation for all stocks with inference.

```
Infosys                 0.002794
Indian_Hotel            0.000266
Mahindra_and_Mahindra  -0.001506
Axis_Bank               0.001167
SAIL                   -0.003463
Shree_Cement            0.003681
Sun_Pharma             -0.001455
Jindal_Steel           -0.004123
Idea_Vodafone          -0.010608
Jet_Airways            -0.009548
```

Figure 5-*stock Means for all stocks*

```
Infosys                 0.03507
Indian_Hotel            0.04713
Mahindra_and_Mahindra   0.04017
Axis_Bank               0.04583
SAIL                    0.06219
Shree_Cement            0.03992
Sun_Pharma              0.04503
Jindal_Steel            0.07511
Idea_Vodafone           0.10432
Jet_Airways             0.09797
```

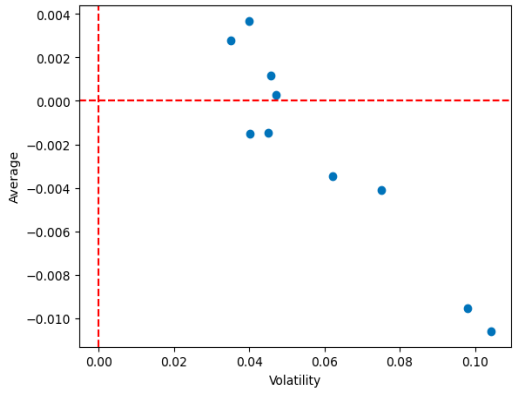
Figure 6-*stock Standard Deviation for all stocks*

The mean returns provide insights into the average daily performance of each stock, with some showing positive returns and others showing negative returns. The standard deviations reflect the level of volatility or risk associated with each stock, with higher standard deviations indicating higher price fluctuations.

Draw a plot of Stock Means vs Standard Deviation and state your inference:

	Average	Volatility
Infosys	0.002794	0.03507
Indian_Hotel	0.000266	0.04713
Mahindra_and_Mahindra	-0.001506	0.04017
Axis_Bank	0.001167	0.04583
SAIL	-0.003463	0.06219
Shree_Cement	0.003681	0.03992
Sun_Pharma	-0.001455	0.04503
Jindal_Steel	-0.004123	0.07511
Idea_Vodafone	-0.010608	0.10432
Jet_Airways	-0.009548	0.09797

Stock Mean & Vola;lity of all the stocks



The scatter plot presents mean returns and volatility for various companies. The two companies with the highest mean returns are Infosys and Shree Cement.

The two companies with the lowest mean returns are Idea Vodafone and Jet Airways. Among the two highest mean return companies, Shree Cement has the lower volatility compared to Infosys, making it a more stable investment choice. Similarly, among the two lowest mean return companies, Jet Airways has a slightly lower volatility than Idea Vodafone

Conclusions and Recommendations:

Conclusion:
After analyzing the stock data of various companies, several key insights have emerged. Companies such as Infosys and Shree Cement have demonstrated the highest mean returns, suggesting promising investment opportunities. Conversely, Idea Vodafone and Jet Airways exhibit the lowest mean returns, indicating potential risks associated with investing in these companies.

Recommendations:

1. Consider High-Performing Companies: Investors seeking higher profits may find companies like Infosys and Shree Cement appealing due to their consistent track record of higher returns. These companies present favorable investment prospects for those looking to maximize their gains.

2. Exercise Caution with Lower-Performing Companies: It is advisable for investors to exercise caution when considering investments in companies like Idea Vodafone and Jet Airways, which have shown lower mean returns. Such investments may carry higher risks, and thorough due diligence is recommended before committing capital.

3. Focus on Long-Term Investment: Short-term fluctuations are inherent in the stock market. Therefore, adopting a long-term investment strategy can help mitigate the impact of market volatility and potentially yield higher returns over time.

4. Stay Informed: Investors should regularly monitor the performance of their investments and stay informed about market trends and developments. This proactive approach enables investors to make informed decisions and adapt to changing market conditions effectively.

By adhering to these recommendations, investors can make prudent investment decisions and optimize their portfolio performance in line with their financial goals and risk tolerance.