Predicting Company Defaults

A Data-Driven Approach to Financial Risk Management

Financial Risk Analysis
Project: Default Prediction

Shaishav Merchant
PGPDSBA | MDS

Table of contents

List of Tables

List of Figures

Executive Summary

The project aimed to predict company defaults using financial data to assist in credit risk management. The approach involved exploring and preprocessing the data, addressing class imbalance, and building various models. After comparing **Logistic Regression**, **Random Forest**, and **LDA**, the **Random Forest Tuned Model** was selected for its strong balance between recall and precision. Key insights were derived from feature importance, and actionable business recommendations were provided to mitigate default risk.

- **Problem Statement**: Predict company defaults using financial indicators to improve credit risk management.
- **Objectives**: Build predictive models that maximize recall, identify key financial drivers, and provide actionable insights.
- **EDA Summary**: Key variables like **Retained Earnings to Total Assets** and **Total Debt to Total Net Worth** showed strong correlations with default risk.
- **Data Preprocessing**: Addressed missing values, outliers, and used **SMOTE** to balance the dataset.
- **Model Building**: Developed and tuned **Logistic Regression**, **Random Forest**, and **LDA** models.
- **Model Comparison**: Random Forest (Tuned) achieved the best recall and balance between precision and F1-score.
- **Key Insights**: High leverage and weak retained earnings increase default risk, while strong cash flow reduces it.
- **Business Recommendations**: Focus on debt reduction, improving retained earnings, and proactive risk management using the Random Forest Tuned Model.

Problem Statement

Companies facing financial distress may default on their obligations, leading to a decline in credit ratings and limiting future access to credit. Defaults can result in increased interest rates on existing debt and create challenges for securing new financing. For investors and financial institutions, identifying companies at risk of default is crucial for managing credit risk and protecting financial stability. This project focuses on using financial data from companies' balance sheets to predict defaults and help stakeholders make informed decisions.

Objectives

- Develop machine learning models to identify companies at risk of default using their financial data, helping investors make informed decisions.
- Balance the data to ensure that the model captures as many potential defaults as possible, reducing missed risks.
- Prioritize identifying companies likely to default while minimizing the chances of false alarms.
- Highlight key financial factors that indicate a company’s likelihood of default, offering insights into financial health.

- Offer practical recommendations to help investors manage credit risks and enhance their financial strategies.

Solution Approach

The solution involved exploring financial data, addressing missing values and outliers, balancing the dataset, and building predictive models. The models were tuned and compared to select the most effective one for predicting defaults.

Solution Approach in Detail

- **Data Understanding:** Analyze key financial variables impacting defaults.
- **Exploratory Data Analysis (EDA):** Uncover patterns and correlations through univariate and bivariate analysis.
- **Data Preparation:** Address missing values, outliers, and scale the data.
- **Balancing Data:** Apply SMOTE to handle class imbalance.
- **Model Development:** Build and tune Logistic Regression, Random Forest, and LDA models.
- **Hyperparameter Tuning:** Optimize the Random Forest model for better performance.
- **Model Comparison:** Select the best-performing model based on key metrics.
- **Business Recommendations:** Provide insights and strategies based on model outcomes.

Data Overview

The dataset contains financial data of companies used to predict defaults, with various financial ratios and indicators as features.

Key Observations

- **Shape**: 2,058 rows and 58 columns.
- **Data Type**: Dataset contains 1 categorical column and 57 numerical columns.
- **Target Distribution**: Highly imbalanced; few defaults (class 1).
- **Outliers**: Present in financial ratios like debt and retained earnings.
- **Irrelevant Columns**: Company name and code are not predictive.
- **Missing Values**: Some financial metrics have missing data.
- **Scaling**: Required for features with different units/ranges.
- **Multicollinearity**: Some features are highly correlated, requiring reduction.

Table 1: Dataset Information

Column Name	Data Type	Description
Co_Code	int64	Company Code
Co_Name	object	Company Name
_Operating_Expense_Rate	float64	Operating Expenses/Net Sales
_Research_and_development_expense_rate	float64	R&D Expenses/Net Sales
_Cash_flow_rate	float64	Cash Flow from Operating/Current Liabilities
_Interest_bearing_debt_interest_rate	float64	Interest-bearing Debt/Equity
_Tax_rate_A	float64	Effective Tax Rate
_Cash_Flow_Per_Share	float64	After-tax earnings per share
_Per_Share_Net_profit_before_tax_Yuan_	float64	Pretax Income Per Share
_Realized_Sales_Gross_Profit_Growth_Rate	float64	Realized Sales Gross Profit Growth Rate
_Operating_Profit_Growth_Rate	float64	Operating Income Growth Rate
_Continuous_Net_Profit_Growth_Rate	float64	Continuous Net Profit Growth Rate
_Total_Asset_Growth_Rate	float64	Total Asset Growth Rate
_Net_Value_Growth_Rate	float64	Total Equity Growth
_Total_Asset_Return_Growth_Rate_Ratio	float64	Return on Total Asset Growth
_Cash_Reinvestment_perc	float64	Cash Reinvestment %
_Current_Ratio	float64	Current Ratio (Assets/Liabilities)
_Quick_Ratio	float64	Acid-test Ratio
_Interest_Expense_Ratio	float64	Interest Expenses/Total Revenue
_Total_debt_to_Total_net_worth	float64	Total Debt/Net Worth
_Long_term_fund_suitability_ratio_A	float64	Long-term Liability + Equity / Fixed Assets
_Net_profit_before_tax_to_Paid_in_capital	float64	Pretax Income/Capital
_Total_Asset_Turnover	float64	Net Sales / Average Total Assets
_Accounts_Receivable_Turnover	float64	Receivables Turnover Ratio
_Average_Collection_Days	float64	Days Receivable Outstanding
_Inventory_Turnover_Rate_times	float64	Inventory Turnover Rate
_Fixed_Assets_Turnover_Frequency	float64	Fixed Asset Turnover
_Net_Worth_Turnover_Rate_times	float64	Equity Turnover
_Operating_profit_per_person	float64	Operation Income Per Employee
_Allocation_rate_per_person	float64	Fixed Assets Per Employee
_Quick_Assets_to_Total_Assets	float64	Quick Assets / Total Assets
_Cash_to_Total_Assets	float64	Cash / Total Assets
_Quick_Assets_to_Current_Liability	float64	Quick Assets / Current Liability
_Cash_to_Current_Liability	float64	Cash / Current Liability
_Operating_Funds_to_Liability	float64	Operating Funds to Liability
_Inventory_to_Working_Capital	float64	Inventory/Working Capital
_Inventory_to_Current_Liability	float64	Inventory/Current Liability
_Long_term_Liability_to_Current_Assets	float64	Long-term Liability / Current Assets
_Retained_Earnings_to_Total_Assets	float64	Retained Earnings / Total Assets
_Total_income_to_Total_expense	float64	Total Income / Total Expense
_Total_expense_to_Assets	float64	Total Expense / Assets
_Current_Asset_Turnover_Rate	float64	Current Assets / Sales
_Quick_Asset_Turnover_Rate	float64	Quick Assets / Sales
_Cash_Turnover_Rate	float64	Cash to Sales
_Fixed_Assets_to_Assets	float64	Fixed Assets / Total Assets
_Cash_Flow_to_Total_Assets	float64	Cash Flow to Total Assets
_Cash_Flow_to_Liability	float64	Cash Flow to Liability
_CFO_to_Assets	float64	Cash Flow from Operations / Assets
_Cash_Flow_to_Equity	float64	Cash Flow to Equity
_Current_Liability_to_Current_Assets	float64	Current Liability / Current Assets
_Liability_Assets_Flag	int64	1 if Total Liability > Assets, else 0
_Total_assets_to_GNP_price	float64	Total Assets to GNP Price
_No_credit_Interval	float64	No-credit Interval

_Degree_of_Financial_Leverage_DFL	float64	Financial Leverage
_Interest_Coverage_Ratio_Interest_expense_to_EBIT	float64	EBIT/Interest Expense
_Net_Income_Flag	int64	1 if Net Income Negative, else 0
_Equity_to_Liability	float64	Equity / Liability
Default	int64	1 = Default, 0 = Not Defaulted

Data Statistics (Numerical and Categorical)

- Many financial metrics, like Operating Expense Rate and Research & Development Expense Rate, show a broad range in values. For example:
    - The operating expense rate varies from small value to as high as $9.98 billion.
    - Tax rate ranges from 0 to 99.9%, indicating variability across companies.
- The dataset includes a combination of continuous financial ratios and binary flags (e.g., _Net_Income_Flag, Default).
- Columns irrelevant for model building can be removed, for e.g. Co_Code, Co_Name.

Table 2: Data Statistics - Numerical Features

Feature	count	mean	std	min	25%	50%	75%	max
Co_Code	2058	1.76E+04	2.19E+04	4	3.67E+03	6.24E+03	2.43E+04	7.25E+04
_Operating_Expense_Rate	2058	2.05E+09	3.25E+09	0.0001	1.58E-04	3.33E-04	4.11E+09	9.98E+09
_Research_and_development_expense_rate	2058	1.21E+09	2.14E+09	0	0.00E+00	1.99E-04	1.55E+09	9.98E+09
_Cash_flow_rate	2058	4.65E-01	2.27E-02	0	4.60E-01	4.63E-01	4.68E-01	1.00E+00
_Interest_bearing_debt_interest_rate	2058	1.11E+07	9.04E+07	0	2.76E-04	4.54E-04	6.63E-04	9.90E+08
_Tax_rate_A	2058	1.15E-01	1.52E-01	0	0.00E+00	3.71E-02	2.16E-01	1.00E+00
_Cash_Flow_Per_Share	1891	3.20E-01	1.53E-02	0.169449	3.15E-01	3.21E-01	3.26E-01	4.62E-01
_Per_Share_Net_profit_before_tax_Yuan_	2058	1.77E-01	3.02E-02	0	1.67E-01	1.76E-01	1.86E-01	7.92E-01
_Realized_Sales_Gross_Profit_Growth_Rate	2058	2.28E-02	2.17E-02	0.004282	2.21E-02	2.21E-02	2.22E-02	1.00E+00
_Operating_Profit_Growth_Rate	2058	8.48E-01	4.59E-03	0.73643	8.48E-01	8.48E-01	8.48E-01	1.00E+00
_Continuous_Net_Profit_Growth_Rate	2058	2.17E-01	5.68E-03	0	2.18E-01	2.18E-01	2.18E-01	2.33E-01
_Total_Asset_Growth_Rate	2058	5.29E+09	2.91E+09	0	4.32E+09	6.23E+09	7.22E+09	9.98E+09
_Net_Value_Growth_Rate	2058	5.19E+06	2.08E+08	0	4.36E-04	4.55E-04	4.88E-04	9.33E+09
_Total_Asset_Return_Growth_Rate_Ratio	2058	2.64E-01	2.42E-03	0.25162	2.64E-01	2.64E-01	2.64E-01	3.59E-01
_Cash_Reinvestment_perc	2058	3.77E-01	2.74E-02	0.025828	3.71E-01	3.79E-01	3.86E-01	1.00E+00
_Current_Ratio	2058	1.34E+06	6.06E+07	0	6.57E-03	8.95E-03	1.35E-02	2.75E+09
_Quick_Ratio	2058	2.78E+07	4.45E+08	0	2.95E-03	5.28E-03	8.90E-03	9.23E+09
_Interest_Expense_Ratio	2058	6.31E-01	6.79E-03	0.525126	6.31E-01	6.31E-01	6.32E-01	8.12E-01
_Total_debt_to_Total_net_worth	2037	1.07E+07	2.70E+08	0	3.92E-03	7.27E-03	1.31E-02	9.94E+09
_Long_term_fund_suitability_ratio_A	2058	8.97E-03	3.49E-02	0.004129	5.16E-03	5.52E-03	6.42E-03	1.00E+00
_Net_profit_before_tax_to_Paid_in_capital	2058	1.75E-01	2.62E-02	0	1.66E-01	1.75E-01	1.84E-01	7.92E-01
_Total_Asset_Turnover	2058	1.29E-01	1.01E-01	0	6.15E-02	1.03E-01	1.68E-01	9.19E-01
_Accounts_Receivable_Turnover	2058	4.16E+07	5.05E+08	0	7.45E-04	1.08E-03	1.85E-03	9.74E+09
_Average_Collection_Days	2058	2.63E+07	4.11E+08	0	3.58E-03	6.00E-03	8.64E-03	8.80E+09
_Inventory_Turnover_Rate_times	2058	2.03E+09	3.08E+09	0	1.91E-04	1.91E+07	3.82E+09	9.99E+09
_Fixed_Assets_Turnover_Frequency	2058	1.23E+09	2.65E+09	0	2.28E-04	6.00E-04	8.42E-03	9.99E+09
_Net_Worth_Turnover_Rate_times	2058	3.96E-02	4.24E-02	0.008871	2.05E-02	2.87E-02	4.44E-02	1.00E+00
_Operating_profit_per_person	2058	4.04E-01	5.36E-02	0	3.91E-01	3.95E-01	4.01E-01	1.00E+00
_Allocation_rate_per_person	2058	5.73E+06	1.98E+08	0	4.67E-03	1.06E-02	2.46E-02	8.28E+09

_Quick_Assets_to_Total_Assets	2058	3.42E-01	2.10E-01	0	1.73E-01	3.06E-01	4.85E-01	9.89E-01
_Cash_to_Total_Assets	1962	7.99E-02	9.86E-02	0.000184	2.06E-02	4.56E-02	9.77E-02	9.25E-01
_Quick_Assets_to_Current_Liability	2058	1.19E+07	3.12E+08	0	3.62E-03	5.97E-03	9.61E-03	8.82E+09
_Cash_to_Current_Liability	2058	9.28E+07	7.85E+08	0.000101	1.09E-03	2.68E-03	7.54E-03	9.17E+09
_Operating_Funds_to_Liability	2058	3.48E-01	3.84E-02	0.026274	3.38E-01	3.45E-01	3.54E-01	1.00E+00
_Inventory_to_Working_Capital	2058	2.78E-01	1.84E-02	0	2.77E-01	2.77E-01	2.78E-01	1.00E+00
_Inventory_to_Current_Liability	2058	5.79E+07	6.28E+08	0	2.89E-03	6.78E-03	1.28E-02	9.60E+09
_Long_term_Liability_to_Current_Assets	2058	7.34E+07	6.69E+08	0	0.00E+00	2.59E-03	1.05E-02	9.31E+09
_Retained_Earnings_to_Total_Assets	2058	9.30E-01	2.98E-02	0	9.28E-01	9.35E-01	9.41E-01	9.73E-01
_Total_income_to_Total_expense	2058	2.36E-03	4.64E-04	0	2.19E-03	2.30E-03	2.43E-03	1.03E-02
_Total_expense_to_Assets	2058	3.11E-02	3.87E-02	0.000853	1.27E-02	2.09E-02	3.53E-02	1.00E+00
_Current_Asset_Turnover_Rate	2058	1.27E+09	2.84E+09	0	1.50E-04	2.46E-04	1.26E-03	9.99E+09
_Quick_Asset_Turnover_Rate	2058	2.57E+09	3.45E+09	0	1.51E-04	3.79E-04	5.79E+09	1.00E+10
_Cash_Turnover_Rate	2058	2.65E+09	2.82E+09	0.0001	1.74E-03	1.73E+09	4.55E+09	9.99E+09
_Fixed_Assets_to_Assets	2058	4.04E+06	1.83E+08	0	9.65E-02	2.14E-01	4.15E-01	8.32E+09
_Cash_Flow_to_Total_Assets	2058	6.44E-01	4.51E-02	0	6.33E-01	6.43E-01	6.54E-01	1.00E+00
_Cash_Flow_to_Liability	2058	4.60E-01	3.29E-02	0.032583	4.57E-01	4.59E-01	4.62E-01	9.05E-01
_CFO_to_Assets	2058	5.80E-01	6.38E-02	0	5.50E-01	5.83E-01	6.12E-01	9.75E-01
_Cash_Flow_to_Equity	2058	3.15E-01	1.28E-02	0	3.13E-01	3.15E-01	3.17E-01	5.69E-01
_Current_Liability_to_Current_Assets	2044	3.94E-02	4.80E-02	0	2.18E-02	3.27E-02	4.39E-02	1.00E+00
_Liability_Assets_Flag	2058	3.40E-03	5.82E-02	0	0.00E+00	0.00E+00	0.00E+00	1.00E+00
_Total_assets_to_GNP_price	2058	2.78E+07	4.72E+08	0	9.12E-04	2.48E-03	7.00E-03	9.82E+09
_No_credit_Interval	2058	6.24E-01	1.16E-02	0.408682	6.23E-01	6.24E-01	6.24E-01	9.56E-01
_Degree_of_Financial_Leverage_DFL	2058	2.79E-02	1.38E-02	0.012845	2.68E-02	2.68E-02	2.70E-02	4.64E-01
_Interest_Coverage_Ratio_Interest_expense_to_EBIT	2058	5.65E-01	1.15E-02	0.172065	5.65E-01	5.65E-01	5.66E-01	6.67E-01
_Net_Income_Flag	2058	1.00E+00	0.00E+00	1	1.00E+00	1.00E+00	1.00E+00	1.00E+00
_Equity_to_Liability	2058	4.25E-02	5.95E-02	0.003946	2.04E-02	2.85E-02	4.34E-02	1.00E+00
Default	2058	1.07E-01	3.09E-01	0	0.00E+00	0.00E+00	0.00E+00	1.00E+00

Missing Value Check and Treatment

- Some columns, such as _Cash_Flow_Per_Share (167), _Cash_to_Total_Assets (96), _Total_debt_to_Total_net_worth (21) and _Current_Liability_to_Current_Assets (14), have missing entries, though most columns are fully populated.
- Handling these missing values will be necessary as part of data treatment.
- We have replaced missing values with median of the respective column (numerical).

Table 3: Missing Values

Column Name	Missing Value Count
Cash_Flow_Per_Share	167
Total_debt_to_Total_net_worth	21
Cash_to_Total_Assets	96
Current_Liability_to_Current_Assets	14

Outlier Check and Treatment

- We used Boxplots to identify outliers for given numerical features.
- Boxplots reveal substantial outliers.
- Outliers were treated by applying Upper and Lower bound values for respective upper and lower outliers.

Exploratory Data Analysis (EDA)

EDA involved analyzing the dataset's structure, visualizing distributions, identifying correlations, handling outliers, and assessing missing values. Key insights include class imbalance, outlier detection, and the strong predictive potential of financial ratios.

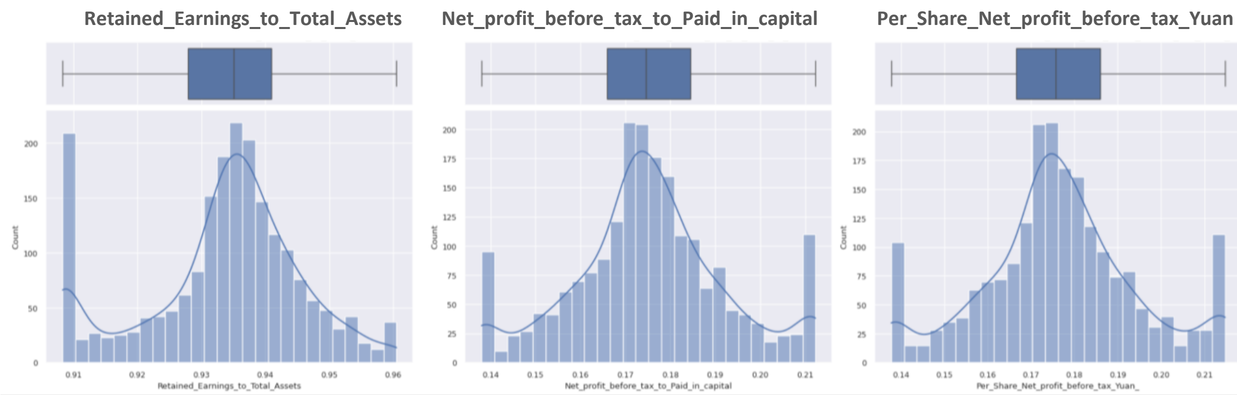
EDA Summary and Key Insights

- Most financial metrics exhibit **moderate variance**, with some features showing right-skewed distributions, such as **debt-to-assets ratios** and **current liabilities**.
- Features like **retained earnings**, **net profit margins**, and **liquidity ratios** generally indicate that the majority of companies are financially stable, though some outliers indicate higher risk.
- **Debt-related metrics** (e.g., **Total_debt_to_Total_net_worth**, **Liabilities_to_Assets**) show a strong distinction between defaulting and non-defaulting companies, with higher debt levels linked to a higher likelihood of default.
- **Profitability ratios** (e.g., **Net_profit_before_tax_to_Paid_in_capital**) reveal that companies with lower profitability are more prone to default.
- **Liquidity metrics** (e.g., **Current_Liability_to_Current_Assets**) highlight that companies struggling with liquidity are at a higher risk of default.
- The **target variable** (Default) shows an **imbalanced distribution**, with a larger proportion of non-defaulting companies compared to defaulting ones. This may necessitate balancing techniques during model building to avoid bias.
- **Debt and liquidity** metrics are the strongest predictors of default risk, with companies carrying high debt and poor liquidity being significantly more likely to default.
- **Profitability** also plays a crucial role, but its predictive power is secondary to debt and liquidity.
- Addressing the **class imbalance** in the target variable will be important to ensure model robustness and fairness.

Univariate Analysis

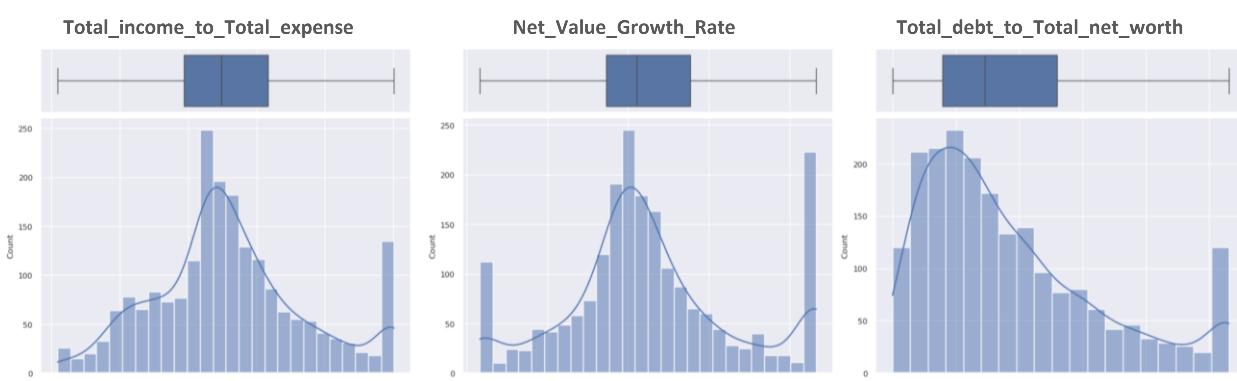
Examined individual feature distributions, identifying skewness and outliers in financial ratios like debt and earnings.

Figure 1: Univariate Analysis - 1 of 4



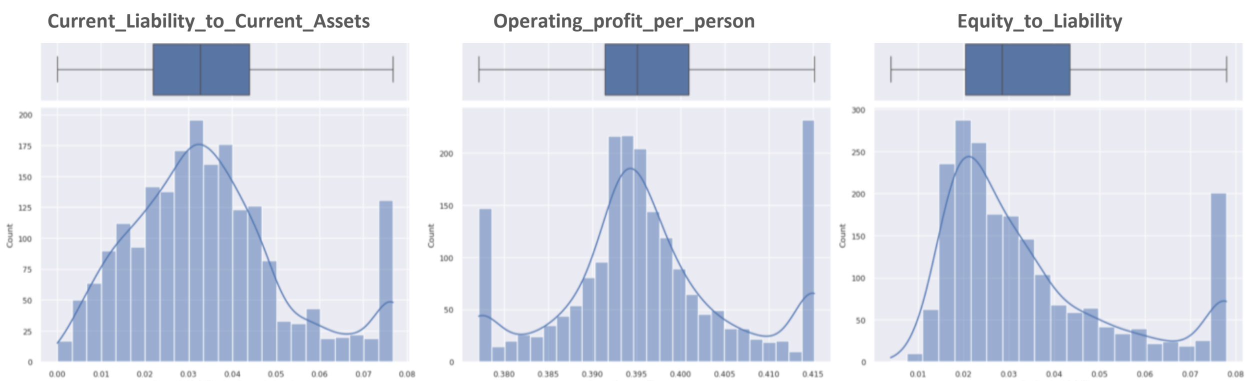
- **Retained_Earnings_to_Total_Assets:**
  - Mean: 0.9334, with a small standard deviation of 0.0124, indicating that most companies have a high proportion of retained earnings relative to total assets.
  - Range: The values are tightly clustered between 0.91 and 0.96, suggesting that this ratio might be a strong differentiator for predicting defaults.
- **Net_profit_before_tax_to_Paid_in_capital:**
  - Mean: 0.1752, with a moderate spread (std: 0.0172).
  - Insight: This feature shows that companies generally have positive net profits relative to their paid-in capital, but variations among companies could be significant.
- **Per_Share_Net_profit_before_tax_Yuan_:**
  - Mean: 0.1764, closely aligned with Net_profit_before_tax_to_Paid_in_capital.
  - Insight: Slight variations in profitability per share could be a potential indicator of financial stability, but the relatively narrow range suggests limited differences between most companies.

Figure 2: Univariate Analysis - 2 of 4



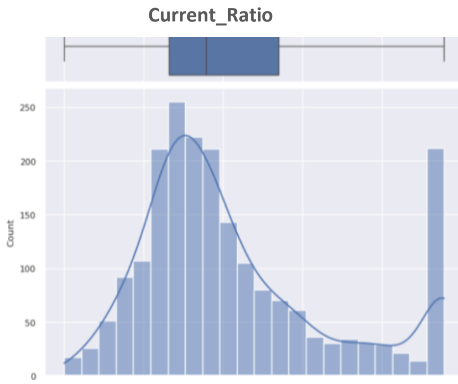
- **Total_income_to_Total_expense**:
    - **Mean**: 0.0023, indicating that income barely exceeds expenses for many companies.
    - **Insight**: This could be a critical feature, as companies operating on thin margins may be more prone to default. Outliers at the higher end could indicate more robust companies.
- **Net_Value_Growth_Rate**:
    - **Mean**: 0.000463, with very small variations.
    - **Insight**: This low rate suggests that most companies have modest growth in their net value, and deviations from this could be important in distinguishing high-risk companies.
- **Total_debt_to_Total_net_worth**:
    - **Mean**: 0.00933, with a **wide standard deviation** (0.0070).
    - **Insight**: This is a key indicator for predicting defaults, as companies with higher debt relative to net worth are at a higher risk. The range (0 to 0.02656) suggests significant variation across companies.

Figure 3: Univariate Analysis – 3 of 4



- **Current_Liability_to_Current_Assets**:
    - **Mean**: 0.0347, but with a large standard deviation (0.0180).
    - **Insight**: The ratio of liabilities to assets can highlight liquidity issues. Companies with high liabilities relative to assets may struggle to meet short-term obligations, increasing their default risk.
- **Operating_profit_per_person**:
    - **Mean**: 0.3962, with minimal variation (std: 0.0100).
    - **Insight**: Operating profit per employee could indicate operational efficiency. However, since most values are close to the mean, this may not be the strongest predictor of default.
- **Equity_to_Liability**:
    - **Mean**: 0.0350, with significant variation (std: 0.0194).
    - **Insight**: This ratio is critical in determining the leverage of a company. Companies with lower equity relative to liabilities are at higher risk, making this a potentially important feature.

Figure 4: Univariate Analysis – 4 of 4

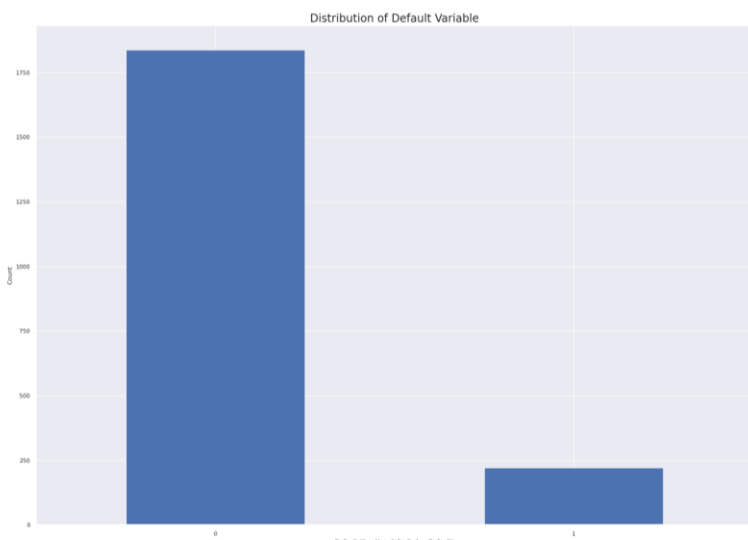


- **Current_Ratio**:
  - **Mean**: 0.0107, with a significant range (0 to 0.0239).
  - **Insight**: This ratio, which compares current assets to current liabilities, is another key indicator of short-term liquidity. Companies with low current ratios may face liquidity crises, which can lead to defaults.

Univariate Analysis – Target Variable Distribution (Default)

The target variable (default) showed a significant class imbalance, with far fewer defaults than non-defaults.

Figure 5: Univariate Analysis - Target Variable Distribution



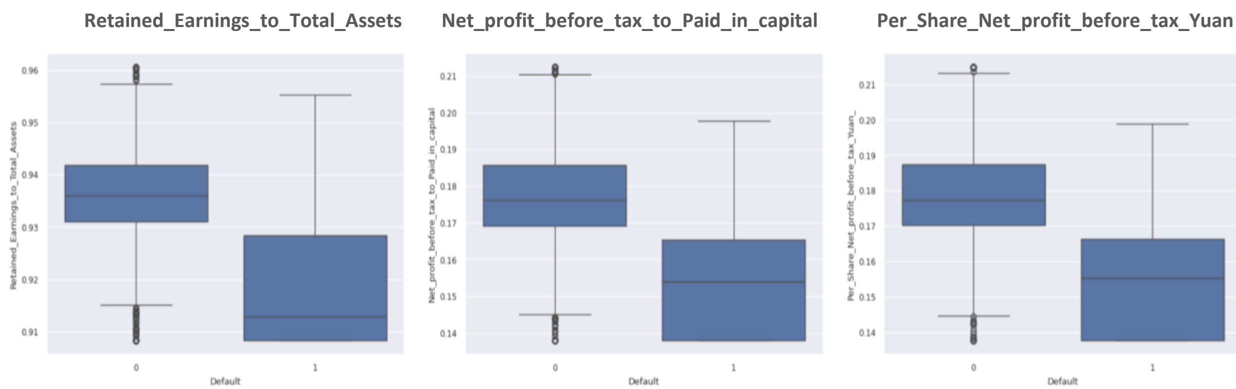
- **Class Distribution:** The bar chart reveals the distribution of companies that have defaulted (Default = 1) is 11% compared to those that have not defaulted (Default = 0) with 89%.

- **Imbalance Data:** The dataset is heavily imbalanced and may need to be addressed using techniques like **SMOTE** (Synthetic Minority Over-sampling Technique) or undersampling

Bivariate Analysis

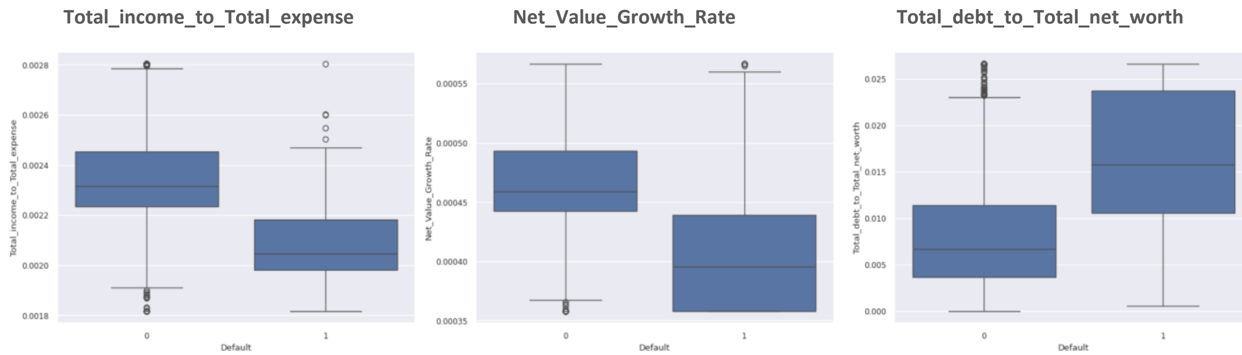
Explored relationships between features and target variable (Default(, highlighting strong correlations between debt ratios and default risk.

Figure 6: Bivariate Analysis – 1 of 4



- **Retained_Earnings_to_Total_Assets**:
  - **Distribution**: The boxplot shows that companies with lower values of retained earnings relative to total assets are more likely to default. Non-defaulting companies tend to have a higher ratio, indicating financial stability and retained profits that buffer against default.
  - **Insight**: A higher ratio of retained earnings signals better financial health, reducing the likelihood of default.
- **Net_profit_before_tax_to_Paid_in_capital**:
  - **Distribution**: Companies that did not default generally exhibit higher ratios of net profit before tax relative to paid-in capital. In contrast, defaulting companies have a more compressed distribution, often lower.
  - **Insight**: Profitability relative to capital invested is a strong indicator of a company's ability to service debts and avoid default.
- **Per_Share_Net_profit_before_tax_Yuan_**:
  - **Distribution**: Non-defaulting companies have a wider and higher spread for this metric, whereas defaulting companies show lower values. This indicates that higher per-share profit helps companies maintain financial stability.
  - **Insight**: Companies with higher profitability per share are less prone to default, reinforcing the idea that profitability is a key factor in financial resilience.

Figure 7: Bivariate Analysis – 2 of 4



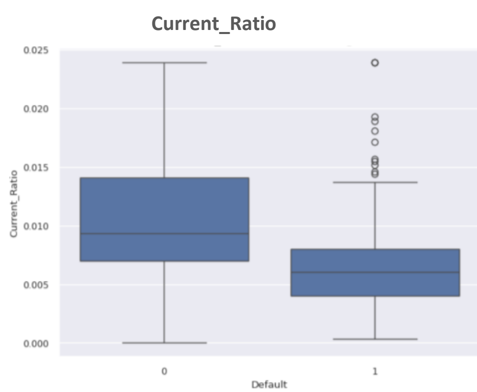
- **Total_income_to_Total_expense**:
    - **Distribution**: The ratio of total income to total expense is slightly higher for non-defaulting companies. Defaulting companies are closer to breaking even, with income barely exceeding expenses or sometimes lower.
    - **Insight**: A tight margin between income and expense leaves companies vulnerable to financial distress, leading to defaults.
- **Net_Value_Growth_Rate**:
    - **Distribution**: Non-defaulting companies show a slightly higher growth rate, although the difference is marginal. Companies with a higher growth rate in net value appear better positioned to handle their debt obligations.
    - **Insight**: Slow or negative growth in net value correlates with a higher probability of default, though the effect is less pronounced compared to other features.

- **Total_debt_to_Total_net_worth**:
    - **Distribution**: This metric shows one of the strongest differentiations. Defaulting companies typically have a much higher debt-to-net-worth ratio, indicating over-leverage and a higher risk of default.
    - **Insight**: Companies with excessive debt relative to their net worth are significantly more likely to default, making this a critical predictor.

Figure 8: Bivariate Analysis – 3 of 4

- **Current_Liability_to_Current_Assets**:
    - **Distribution**: Companies that default have a higher ratio of current liabilities to current assets, indicating liquidity issues. Non-defaulting companies generally maintain lower ratios, suggesting better liquidity management.
    - **Insight**: Poor liquidity (higher liabilities compared to assets) is a strong indicator of a company's inability to meet short-term obligations, leading to default.
- **Operating_profit_per_person**:
    - **Distribution**: While the difference between defaulting and non-defaulting companies is less significant, non-defaulting companies generally have slightly higher operating profit per person.
    - **Insight**: Operational efficiency may play a secondary role in predicting default, but it's not as strong an indicator as debt-related metrics.
- **Equity_to_Liability**:
    - **Distribution**: Defaulting companies tend to have much lower equity relative to liabilities, suggesting over-reliance on debt. Non-defaulting companies maintain a higher equity buffer.
    - **Insight**: A low equity-to-liability ratio signals financial vulnerability, with companies being more dependent on external borrowing and at a higher risk of default.

Figure 9: Bivariate Analysis – 4 of 4



- **Current_Ratio**:
    - **Distribution**: Non-defaulting companies generally have a higher current ratio, indicating their ability to cover short-term liabilities with current assets. Defaulting companies, on the other hand, have lower current ratios, reflecting liquidity constraints.
    - **Insight**: A higher current ratio indicates better liquidity and a lower likelihood of default, as companies are more capable of meeting short-term obligations.

Data Pre-processing

Data pre-processing involved splitting the dataset into training and testing sets, addressing multicollinearity by removing highly correlated features, and scaling the dataset to standardize feature ranges for better model performance.

Train Test Split

The dataset was prepared to ensure effective model training and reliable performance evaluation.

- The dataset is split into training and test sets in a **67:33** ratio.
- The random_state parameter is set to **42** as per project requirement.
- The **stratify=y** ensures consistent distribution of target variable between train and test sets.

Multicollinearity Check

Multicollinearity was managed by identifying and removing highly correlated features to enhance model accuracy and stability. The Variance Inflation Factor (VIF) from Statsmodels was used to detect features that were potentially insignificant for model building.

Following features showed high multicollinearity and were removed.

- 'Per_Share_Net_profit_before_tax_Yuan_' with VIF: 83.22809466675108
- 'Cash_Flow_to_Total_Assets' with VIF: 66.03122822275698
- 'CFO_to_Assets' with VIF: 32.10901086559086
- 'Quick_Assets_to_Current_Liability' with VIF: 23.777991715154883
- 'Operating_Funds_to_Liability' with VIF: 18.94608144934635
- 'Net_Worth_Turnover_Rate_times' with VIF: 15.958999108386413
- 'Current_Ratio' with VIF: 13.197573187697344
- 'Net_profit_before_tax_to_Paid_in_capital' with VIF: 7.837478850193706
- 'Interest_Coverage_Ratio_Interest_expense_to_EBIT' with VIF: 6.39727878068583
- 'Cash_Reinvestment_perc' with VIF: 6.249952454168627
- 'Quick_Ratio' with VIF: 5.820388649261178
- 'Cash_Flow_to_Equity' with VIF: 5.435972466974049

Observations and Conclusions:

- After treating for multicollinearity, all remaining columns show a **VIF below 5**, indicating that **multicollinearity has been successfully addressed**.
- The column **Net_Income_Flag** has a **VIF of 0**, suggesting that it provides no additional predictive value (likely due to being constant or redundant). Therefore, it has been **dropped** from the dataset.

- The column **Liability_Assets_Flag** shows a **VIF of NaN**, which indicates it is not contributing meaningfully to the model and has **no significance** for prediction. Consequently, this column has also been **dropped**.

Table 4: Retained Features

Feature	VIF	Feature	VIF
Quick_Assets_to_Total_Assets	4.975127	Average_Collection_Days	2.668462
Fixed_Assets_to_Assets	4.817482	Cash_Flow_Per_Share	2.518432
Total_income_to_Total_expense	4.237309	Net_Value_Growth_Rate	2.397772
Current_Liability_to_Current_Assets	4.185876	Total_expense_to_Assets	2.272233
Equity_to_Liability	4.077838	Inventory_to_Current_Liability	2.246206
Operating_Profit_Growth_Rate	3.75825	Fixed_Assets_Turnover_Frequency	1.97316
Retained_Earnings_to_Total_Assets	3.721979	Total_assets_to_GNP_price	1.742321
Continuous_Net_Profit_Growth_Rate	3.560356	Long_term_Liability_to_Current_Assets	1.735145
Cash_to_Current_Liability	3.51583	No_credit_Interval	1.661026
Total_Asset_Turnover	3.437522	Current_Asset_Turnover_Rate	1.63166
Cash_flow_rate	3.325193	Inventory_to_Working_Capital	1.57132
Cash_to_Total_Assets	3.263625	Tax_rate_A	1.500934
Total_debt_to_Total_net_worth	3.235778	Quick_Asset_Turnover_Rate	1.416248
Total_Asset_Return_Growth_Rate_Ratio	3.169469	Cash_Flow_to_Liability	1.413948
Operating_profit_per_person	3.095232	Operating_Expense_Rate	1.362061
Realized_Sales_Gross_Profit_Growth_Rate	2.897238	Inventory_Turnover_Rate_times	1.231354
Allocation_rate_per_person	2.890704	Research_and_development_expense_rate	1.194437
Long_term_fund_suitability_ratio_A	2.829899	Total_Asset_Growth_Rate	1.161993
Accounts_Receivable_Turnover	2.785597	Cash_Turnover_Rate	1.112462
Interest_Expense_Ratio	2.779381	Interest_bearing_debt_interest_rate	1.110364
Degree_of_Financial_Leverage_DFL	2.730984		

Scaling the Dataset

The dataset was scaled to standardize feature values, ensuring consistent model performance. This helped improve model convergence and handled varying units across financial metrics effectively.

- StandardScaler was applied to standardize the dataset, ensuring all features have a mean of 0 and a standard deviation of 1.
- Scaling improves model performance by preventing features with larger ranges from dominating and helps models converge more efficiently.
- It is recommended to scale features, especially when using algorithms sensitive to feature magnitude, like logistic regression and random forest.

Model Building

The model-building process involved developing **Logistic Regression**, **Random Forest**, and **LDA** models. Data balancing with **SMOTE** improved recall for minority classes. Hyperparameter tuning and optimal threshold selection were applied to maximize performance. The **Random Forest Tuned Model** was ultimately selected based on its balance of recall, precision, and F1 score.

Logistic Regression Model:

Initial Logistic Regression model built using original data had several **statistically-insignificant** features as demonstrated in the following model summary image.

Figure 10: Logistic Model Summary

```
Optimization terminated successfully.
         Current function value: 0.183935
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                Default   No. Observations:                 1378
Model:                          Logit   Df Residuals:                     1336
Method:                           MLE   Df Model:                           41
Date:                Mon, 09 Sep 2024   Pseudo R-squ.:                  0.4582
Time:                        08:12:36   Log-Likelihood:                -253.46
converged:                       True   LL-Null:                       -467.84
Covariance Type:            nonrobust   LLR p-value:                 4.590e-66
=========================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
Intercept                                -4.0450      0.269    -15.038      0.000      -4.572      -3.518
Operating_Expense_Rate                    0.1206      0.141      0.853      0.394      -0.157       0.398
Research_and_development_expense_rate     0.4639      0.125      3.724      0.000       0.220       0.708
Cash_flow_rate                            0.1638      0.264      0.621      0.534      -0.353       0.680
Interest_bearing_debt_interest_rate       0.4090      0.148      2.759      0.006       0.118       0.700
Tax_rate_A                               -0.1759      0.177     -0.996      0.319      -0.522       0.170
Cash_Flow_Per_Share                      -0.1963      0.209     -0.939      0.348      -0.606       0.213
Realized_Sales_Gross_Profit_Growth_Rate  -0.0264      0.161     -0.164      0.869      -0.341       0.288
Operating_Profit_Growth_Rate              0.0044      0.195      0.022      0.982      -0.377       0.386
Continuous_Net_Profit_Growth_Rate        -0.4955      0.215     -2.304      0.021      -0.917      -0.074
Total_Asset_Growth_Rate                  -0.1436      0.140     -1.025      0.305      -0.418       0.131
Net_Value_Growth_Rate                    -0.0866      0.173     -0.501      0.617      -0.426       0.252
Total_Asset_Return_Growth_Rate_Ratio      0.3468      0.195      1.775      0.076      -0.036       0.730
Interest_Expense_Ratio                    0.0381      0.160      0.237      0.812      -0.276       0.352
Total_debt_to_Total_net_worth             0.6888      0.192      3.580      0.000       0.312       1.066
Long_term_fund_suitability_ratio_A        0.2205      0.198      1.113      0.266      -0.168       0.609
Total_Asset_Turnover                     -0.2322      0.255     -0.911      0.362      -0.732       0.268
Accounts_Receivable_Turnover             -0.7323      0.224     -3.271      0.001      -1.171      -0.293
Average_Collection_Days                   0.0957      0.188      0.508      0.611      -0.274       0.465
Inventory_Turnover_Rate_times             0.0233      0.133      0.175      0.861      -0.238       0.285
Fixed_Assets_Turnover_Frequency           0.1694      0.159      1.065      0.287      -0.142       0.481
Operating_profit_per_person               0.2940      0.211      1.391      0.164      -0.120       0.708
Allocation_rate_per_person                0.4018      0.203      1.980      0.048       0.004       0.800
Quick_Assets_to_Total_Assets             -0.7321      0.303     -2.417      0.016      -1.326      -0.139
Cash_to_Total_Assets                      0.0473      0.211      0.224      0.823      -0.367       0.461
Cash_to_Current_Liability                 0.0937      0.176      0.532      0.595      -0.252       0.439
Inventory_to_Working_Capital             -0.0821      0.120     -0.684      0.494      -0.317       0.153
Inventory_to_Current_Liability           -0.1234      0.212     -0.583      0.560      -0.538       0.291
Long_term_Liability_to_Current_Assets    -0.3759      0.156     -2.413      0.016      -0.681      -0.071
Retained_Earnings_to_Total_Assets        -0.6826      0.251     -2.725      0.006      -1.174      -0.192
Total_income_to_Total_expense            -0.6814      0.330     -2.064      0.039      -1.328      -0.034
Total_expense_to_Assets                   0.5792      0.190      3.048      0.002       0.207       0.952
Current_Asset_Turnover_Rate              -0.0906      0.150     -0.603      0.547      -0.385       0.204
Quick_Asset_Turnover_Rate                -0.0108      0.145     -0.075      0.941      -0.294       0.273
Cash_Turnover_Rate                       -0.3623      0.144     -2.513      0.012      -0.645      -0.080
Fixed_Assets_to_Assets                   -0.0380      0.228     -0.167      0.867      -0.484       0.408
Cash_Flow_to_Liability                   -0.2401      0.176     -1.367      0.171      -0.584       0.104
Current_Liability_to_Current_Assets       0.0182      0.229      0.080      0.937      -0.430       0.466
Total_assets_to_GNP_price                 0.1335      0.153      0.873      0.383      -0.166       0.433
No_credit_Interval                        0.0970      0.135      0.721      0.471      -0.167       0.361
Degree_of_Financial_Leverage_DFL          0.1019      0.164      0.622      0.534      -0.219       0.423
Equity_to_Liability                      -0.9072      0.339     -2.672      0.008      -1.573      -0.242
=========================================================================================================
```

Observations:

Based on the **model summary** with several features having **p-values greater than 0.05**, here are a few key observations:

- **Statistical Insignificance:** Features with p-values greater than 0.05 are statistically insignificant, indicating they do not contribute meaningfully to predicting the target variable (Default). Retaining these features adds noise to the model without improving prediction accuracy.
- **Risk of Overfitting:** Including insignificant features increases the risk of overfitting. By removing these features, the model becomes more parsimonious, improving generalization to unseen data.
- **Model Simplification:** Removing features with p-values > 0.05 simplifies the model, making it more interpretable and computationally efficient, without sacrificing predictive power.
- **Improved Model Performance:** By focusing only on statistically significant features (p-value ≤ 0.05), we can reduce model complexity and potentially improve performance metrics like accuracy, precision, and recall.

Next Step:

To enhance model efficiency and predictive capability, it is advisable to remove the insignificant features and refit the model with only the significant ones.

Logistic Regression Model using Significant Features

A Logistic Regression model was built using only the most significant features, improving model interpretability and focusing on key predictors of default risk.

Figure 11: Model Summary (Logistic Regression with Significant Features)

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                Default   No. Observations:                 1378
Model:                          Logit   Df Residuals:                     1365
Method:                           MLE   Df Model:                           12
Date:                Mon, 09 Sep 2024   Pseudo R-squ.:                  0.4335
Time:                        08:12:37   Log-Likelihood:                -265.04
converged:                       True   LL-Null:                       -467.84
Covariance Type:            nonrobust   LLR p-value:                 2.457e-79
=========================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
Intercept                              -3.7918      0.228    -16.598      0.000      -4.240      -3.344
Research_and_development_expense_rate   0.4619      0.114      4.038      0.000       0.238       0.686
Interest_bearing_debt_interest_rate     0.3161      0.139      2.273      0.023       0.044       0.589
Continuous_Net_Profit_Growth_Rate      -0.4039      0.118     -3.419      0.001      -0.635      -0.172
Total_debt_to_Total_net_worth           0.6038      0.172      3.512      0.000       0.267       0.941
Accounts_Receivable_Turnover           -0.7999      0.160     -5.009      0.000      -1.113      -0.487
Allocation_rate_per_person              0.5161      0.155      3.323      0.001       0.212       0.821
Quick_Assets_to_Total_Assets           -0.6440      0.166     -3.875      0.000      -0.970      -0.318
Long_term_Liability_to_Current_Assets  -0.2576      0.122     -2.112      0.035      -0.497      -0.019
Retained_Earnings_to_Total_Assets      -1.0477      0.159     -6.586      0.000      -1.359      -0.736
Total_expense_to_Assets                 0.3763      0.158      2.388      0.017       0.067       0.685
Cash_Turnover_Rate                     -0.3783      0.132     -2.861      0.004      -0.637      -0.119
Equity_to_Liability                    -0.8301      0.291     -2.851      0.004      -1.401      -0.259
=========================================================================================================
```

Figure 12: Performance Metrics and Confusion Matrix (LR with Significant Features)

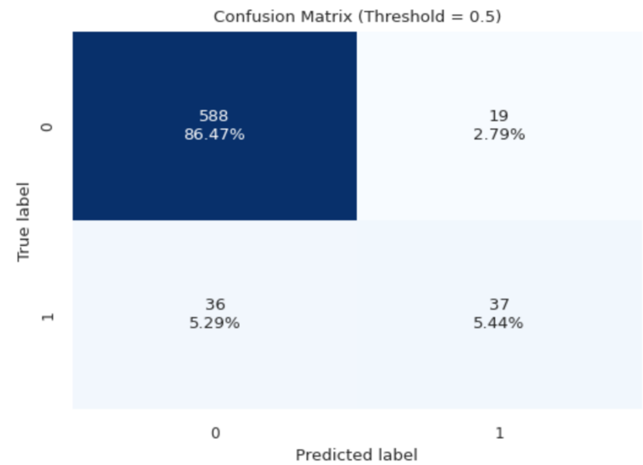
Training Performance:

	Accuracy	Recall	Precision	F1	ROC_AUC
LR	0.91582	0.442177	0.656566	0.528455	0.930022

Testing Performance:

	Accuracy	Recall	Precision	F1	ROC_AUC
LR	0.919118	0.506849	0.660714	0.573643	0.899212

Confusion Matrix:



Observations on LR with Significant Features:

- **Higher Precision but Lower Recall**: With the default threshold of **0.5**, the model achieves **higher precision** on both training (0.657) and testing (0.661). However, the **recall** is lower (0.442 for training, 0.507 for testing), indicating the model is missing more actual defaults (true positives), as reflected by **36 false negatives** in the test set.
- **F1 Score Trade-off**: The **F1 score** is lower (0.528 for training, 0.574 for testing), showing that the default threshold does not strike as good a balance between precision and recall.
- **Confusion Matrix Insights**: The **false negatives (36)** and **true positives (37)** in the confusion matrix highlight that the default threshold is conservative in predicting defaults, favoring fewer false positives (**19**) but missing more actual defaults.
- The default threshold of 0.5 focuses more on precision but at the cost of missing more defaults. A lower threshold may provide a better balance, especially in reducing false negatives.

Next Steps:

We will identify the optimal cut-off threshold for the Logistic Regression model and assess its performance, ensuring a better balance between precision and recall for improved default prediction accuracy.

Evaluate Logistic regression with using Optimal Threshold Value

The Logistic Regression model was built using significant features, and an **optimum threshold** was identified by evaluating various cut-offs on the training dataset. The threshold was chosen to maximize performance metrics like Recall and F1 score. Model performance was then evaluated using this threshold to improve the balance between recall and precision for predicting defaults.

Find Best Threshold

The function find_optimum_cutoff is designed to identify the best threshold for a given model by evaluating performance metrics at various thresholds. It computes key metrics such as accuracy, precision, recall, F1-score, and ROC-AUC for each threshold, and selects the one with the highest F1-score.

Key Steps:

- The model generates predicted probabilities for the dataset.
- Multiple thresholds (0.1 to 0.9) are tested, converting probabilities into binary predictions at each threshold.
- Performance metrics (accuracy, precision, recall, F1, and ROC-AUC) are calculated for each threshold.
- The threshold with the highest **F1 Score** is selected as the optimal one for balancing precision and recall.

Best Threshold: **0.2**

Table 5: Best Threshold and Performance at Best Threshold

Metric	Values
Threshold	0.2
Accuracy	0.896952
Precision	0.511628
Recall	0.748299
F1 Score	0.607735
ROC-AUC	0.930022

Figure 13: Performance Metrics and Confusion Matrix (LR with Optimal Threshold)

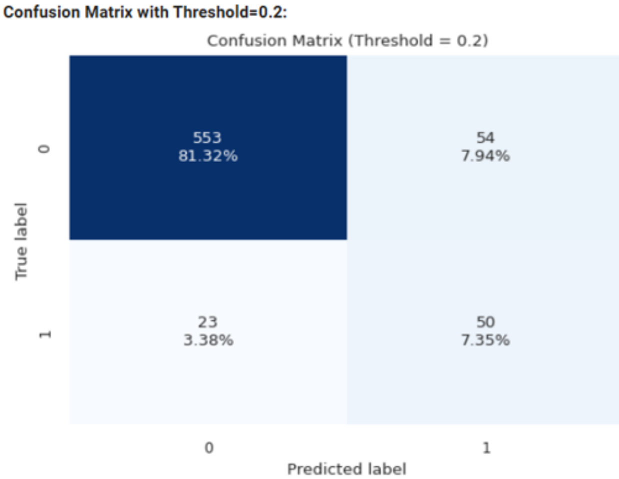
Training Performance with Threshold=0.2:

	Accuracy	Recall	Precision	F1	ROC_AUC
LR Optimal	0.896952	0.748299	0.511628	0.607735	0.930022

Testing Performance with Threshold=0.2:

	Accuracy	Recall	Precision	F1	ROC_AUC
LR Optimal	0.886765	0.684932	0.480769	0.564972	0.899212

Confusion Matrix with Threshold=0.2:



Observations on LR with Optimal Threshold:

- **Improved Recall:** The lower threshold of **0.2** significantly improves recall for both the training (**74.83%**) and test datasets (**68.49%**), meaning the model captures a larger proportion of actual defaults, which is important in minimizing missed default cases (false negatives).
- **Trade-off in Precision:** The increased recall comes at the cost of precision, which drops to **51.16%** on the training set and **48.08%** on the test set. This indicates that a higher number of false positives (non-defaults predicted as defaults) are introduced, which could lead to unnecessary interventions.
- **Balanced F1 Score:** The F1 score on both the training (**60.77%**) and test sets (**56.50%**) shows that the model maintains a reasonable balance between recall and precision, although the overall performance has shifted toward prioritizing recall.
- **Model's Generalization:** The model shows consistent performance across both training and test datasets, with similar trends in recall, precision, and F1 scores. The **ROC-AUC** remains high (over 0.89 on the test set), indicating strong discriminatory ability despite the drop in precision.
- **Confusion Matrix:** The model correctly identifies **50 defaults** but misses **23 actual defaults** (false negatives) in the test data. Additionally, there are **54 false positives**, meaning some non-defaults are being incorrectly flagged as defaults.
- The lower threshold of **0.2 improves the model's ability to detect defaults** (high recall) but introduces more false positives, as indicated by the drop in precision. This threshold may be appropriate if the business goal is to prioritize capturing defaults, even if it means dealing with more false alarms.

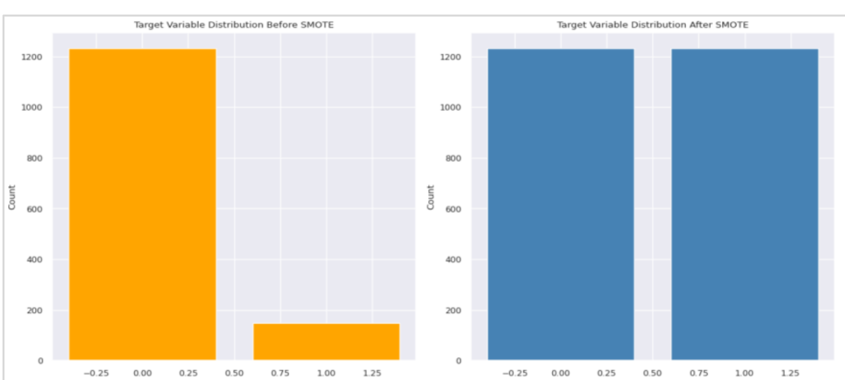
Next Steps:

We will balance the the dataset using SMOTE technique.

Logistic Regression Model Using Balanced Dataset (SMOTE)

A Logistic Regression model was built on the balanced dataset using SMOTE to address the class imbalance in the default prediction. SMOTE helped in generating synthetic samples of the minority class, ensuring better representation of defaults. This balanced dataset improved the model's ability to capture default cases, particularly enhancing recall while minimizing false negatives. The model was then evaluated using an optimal threshold to further balance precision and recall.

Figure 14: Balancing Data using SMOTE



SMOTE (Synthetic Minority Over-sampling Technique) was applied to the training data (X_train_significant and y_train) to balance the class distribution by generating synthetic samples for the minority class (defaults). This improved the model's ability to capture defaults effectively.

- **Balanced Distribution**: The previously imbalanced dataset now has an equal distribution of defaults and non-defaults, reducing class bias in the model.
- **Improved Recall**: The model's recall increased significantly after balancing, capturing more actual default cases.
- **Increased False Positives**: While recall improved, the trade-off was a higher number of false positives, which may require further threshold tuning.

Figure 15: Performance Metrics and Confusion Matrix (LR on Balanced Data with Optimal Threshold)

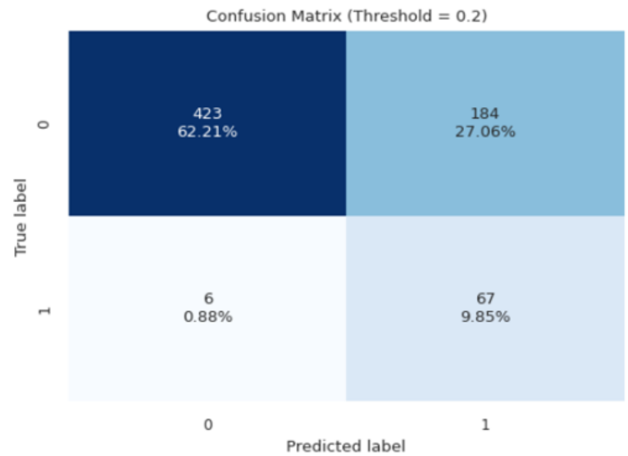
Training Performance with Threshold=0.2:

	Accuracy	Recall	Precision	F1	ROC_AUC
LR Optimal - SMOTE	0.843623	0.983753	0.768401	0.862843	0.939779

Testing Performance with Threshold=0.2:

	Accuracy	Recall	Precision	F1	ROC_AUC
LR Optimal - SMOTE	0.720588	0.917808	0.266932	0.41358	0.893097

Confusion Matrix with Threshold=0.2:



Observations on LR using Balanced Data on Optimal Threshold

- **High Recall**: The model achieves a high recall on both the training set (**98.37%**) and test set (**91.78%**), indicating that it is successfully identifying the vast majority of defaults.
- **Trade-off in Precision**: Precision significantly drops on the test set (**26.69%**), showing a high number of false positives, meaning many non-defaults are incorrectly classified as defaults.
- **F1 inconsistency**: The F1 score on the training data is strong at **86.28%**, but drops considerably on the test set **41.36%**, reflecting the trade-off between high recall and lower precision.
- **Confusion Matrix Insight**: The model correctly predicts **67 defaults** but generates **184 false positives**, meaning the cost of false alarms is high with this threshold.
- **Model's Ability to Distinguish**: The **ROC-AUC** remains strong at **0.89**, indicating the model still has good overall discriminatory power between defaults and non-defaults, despite the precision-recall trade-off.
- Overall, while the model performs well in terms of recall, the large drop in precision and overfitting seen in the F1 score indicate it is not a robust model for production use without further

tuning. A more balanced model that reduces false positives while maintaining a high recall should be the goal.

Next Steps:

After SMOTE, you can build models like Random Forest and LDA, which may perform better in terms of recall and balancing between precision and recall.

Random Forest Model using Original Data

A Random Forest model was built on the original dataset to predict defaults without applying any balancing techniques. This ensemble learning method leveraged multiple decision trees to improve predictive accuracy by reducing variance and preventing overfitting. The model performed well on accuracy but faced challenges in capturing default cases due to the class imbalance, resulting in lower recall. Hyperparameter tuning was applied to optimize the model's performance.

Figure 16: Performance Metric and Confusion Matrix on RF using Original Data

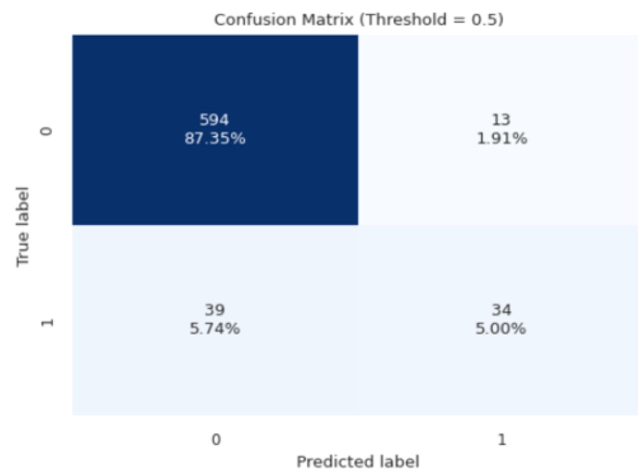
Training Performance:

	Accuracy	Recall	Precision	F1	ROC_AUC
RF - Original	1.0	1.0	1.0	1.0	1.0

Testing Performance:

	Accuracy	Recall	Precision	F1	ROC_AUC
RF - Original	0.923529	0.465753	0.723404	0.566667	0.722168

Confusion Matrix:



Observations RF model build using Original Data

- **Overfitting on Training Data**: The **training performance** shows perfect scores across all metrics (**accuracy, recall, precision, F1, and ROC-AUC** all equal to 1), indicating that the model has likely overfitted the training data. This suggests the Random Forest model is memorizing the training set but may not generalize well to unseen data.
- **Moderate Generalization on Test Data**: While the **test accuracy (0.924)** is high, the **recall (0.466)** is relatively low, meaning the model is missing a significant number of actual defaults (**39 false negatives**). The **precision (0.723)** is good, indicating that when the model predicts a default, it is often correct.
- **Confusion Matrix Insights**: The model performs well at identifying non-defaults (**594 true negatives** and **13 false positives**) but struggles with correctly identifying defaults, as shown by the **39 false negatives**. This indicates that the model favors predicting non-defaults over capturing defaults, even with the optimized threshold.

- In summary, while the model performs well in terms of precision and accuracy, it overfits the training data and struggles with recall on the test set, missing a notable portion of defaults. Further tuning or balancing methods may be needed to improve recall on the test set.

Next Steps:

Build Random Forest Model using Balanced Data Sets.

Random Forest Model Built using Balanced Data Sets

A Random Forest model was built using the balanced dataset, where SMOTE was applied to address the class imbalance. This improved the model's ability to predict defaults, particularly enhancing recall by focusing more on the minority class. The balanced dataset allowed the model to capture more default cases, making it more effective for predicting defaults. Hyperparameter tuning was further applied to optimize performance, balancing precision and recall.

Figure 17: Performance Metric and Confusion Matrix on RF using Balanced Data

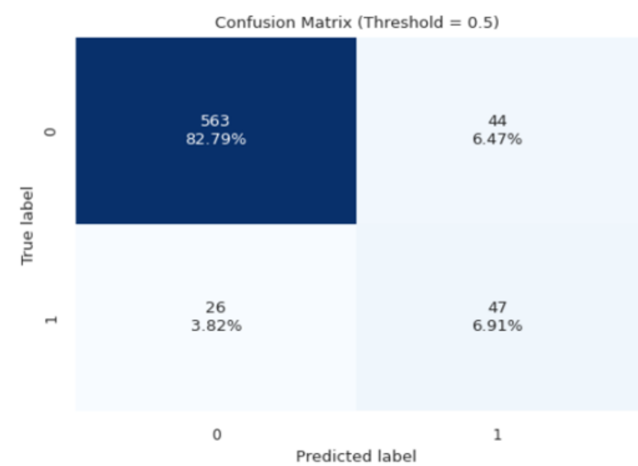
Training Performance:

	Accuracy	Recall	Precision	F1	ROC_AUC
RF - Balanced	1.0	1.0	1.0	1.0	1.0

Testing Performance:

	Accuracy	Recall	Precision	F1	ROC_AUC
RF - Balanced	0.897059	0.643836	0.516484	0.573171	0.785674

Confusion Matrix:



Observations RF model build using Balanced Data

- **Overfitting on Training Data**: As with the original dataset, the **training performance** shows perfect metrics (accuracy, recall, precision, F1, and ROC-AUC all equal to 1). This suggests the model has completely overfitted the training data, especially after applying **SMOTE** to balance the dataset.
- **Improved Recall, Lower Precision on Test Set**: On the **test set**, the model shows **improved recall (0.644)** compared to the unbalanced version, meaning it is now better at identifying defaults (47 true positives). However, this comes at the cost of lower **precision (0.516)**, indicating more false positives (**44**), which means the model is predicting defaults for non-default companies more often.
- **Balanced but Moderate Performance**: The **F1 score (0.573)** and **ROC-AUC (0.786)** are reasonable, but the performance metrics show that the model is slightly favoring capturing more defaults (higher recall) while trading off precision and increasing the number of false positives.

- **Confusion Matrix Insights**: The model captures **47 true positives** and has **26 false negatives**, showing improvement in capturing defaults compared to the unbalanced version. However, the **44 false positives** indicate that the model sacrifices precision to improve recall, misclassifying many non-defaults as defaults.
- In summary, the model's recall improved after balancing the data, but it comes at the cost of increased false positives and lower precision. This makes the model more suitable for scenarios where capturing defaults is prioritized over avoiding false positives. Further tuning may help improve the balance between precision and recall.

Next Steps:

Tune Random Forest Model using Hyperparameters.

Hyperparameter Tuned Random Forest Model using Balanced Data

A Hyperparameter Tuned Random Forest model was built on the balanced dataset using **GridSearchCV** to optimize key parameters such as the number of estimators, maximum depth, and minimum samples per split.

Tuning the Model using GridSearchCV

GridSearchCV was used to tune the Random Forest model by searching for the best combination of hyperparameters to optimize performance.

- **Parameters Used**: Different values for *n_estimators*, *max_depth*, *min_samples_split*, and *min_samples_leaf*.
- **Best Hyperparameters Returned**: The optimal settings found were ***max_depth=7***, ***min_samples_leaf=5***, ***min_samples_split=15***, and ***n_estimators=50***.

Figure 18: Performance Metric and Confusion Matrix on Tuned RF Model on Balanced Data

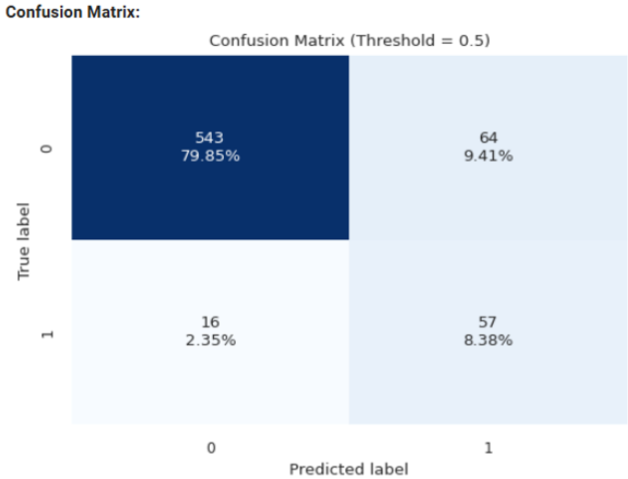
Training Performance:

	Accuracy	Recall	Precision	F1	ROC_AUC
RF - Balanced - Tuned	0.952071	0.982128	0.926437	0.95347	0.952071

Testing Performance:

	Accuracy	Recall	Precision	F1	ROC_AUC
RF - Balanced - Tuned	0.882353	0.780822	0.471074	0.587629	0.837693

Confusion Matrix:



Observations on Tuned RF Model on Balanced Data

- **Strong Training Performance**: The **tuned Random Forest model** performs well on the **training set** with high metrics across the board—**accuracy (0.952)**, **recall (0.982)**, and **precision (0.926)**. Unlike the overfitting seen in previous models, the slightly less-than-perfect metrics suggest better regularization and a good fit on the training data.
- **Improved Recall on Test Set**: The **recall (0.781)** on the test set indicates that the model is capturing more actual defaults, with **57 true positives** and only **16 false negatives**. This shows that the model's ability to identify defaults has significantly improved after tuning.
- **Trade-off in Precision**: While recall has improved, the **precision (0.471)** on the test set is relatively low, with **64 false positives**. This indicates that the model is still over-predicting defaults, sacrificing precision to achieve higher recall.
- **Confusion Matrix Insights**: The confusion matrix shows the model captures **57 true positives** but misclassifies **64 false positives**, highlighting the precision-recall trade-off. The model is now more focused on capturing defaults, but this comes at the cost of more non-defaults being incorrectly classified as defaults.
- Overall, the tuning has resulted in a model that significantly improves default detection (recall) but still needs further optimization to reduce false positives and improve precision.

Next Steps:

We will now explore Linear Discriminant Analysis (LDA) to see if it offers a better balance between precision and recall, or improves the ROC-AUC and F1 scores.

Linear Discriminant Analysis Model using Balanced Data

The Linear Discriminant Analysis (LDA) model classified companies based on their likelihood of default by finding linear combinations of features that best separate classes. While LDA performed reasonably on the balanced dataset, it was outperformed by Random Forest in handling more complex relationships.

Figure 19: Performance Metric and Confusion Matrix for LDA on Balanced Data

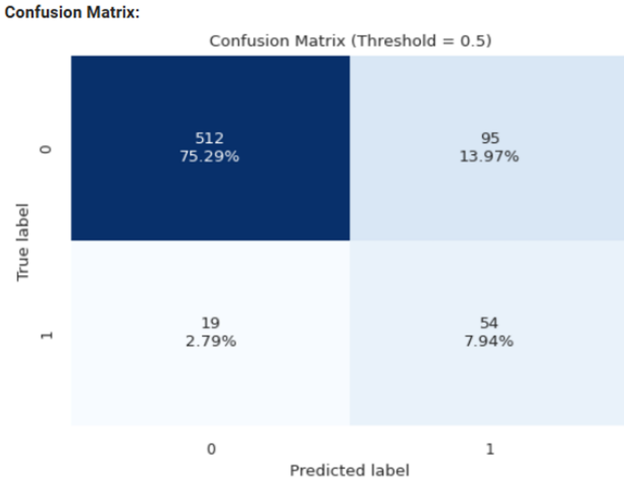
Training Performance:

	Accuracy	Recall	Precision	F1	ROC_AUC
LDA - Balanced	0.859464	0.872461	0.850356	0.861267	0.859464

Testing Performance:

	Accuracy	Recall	Precision	F1	ROC_AUC
LDA - Balanced	0.832353	0.739726	0.362416	0.486486	0.791609

Confusion Matrix:



Observations on LDA Model on Balanced Data

- **Balanced Training Performance**: The **LDA model** shows good balance on the **training set** with **accuracy (0.859)**, **recall (0.872)**, and **precision (0.850)**. This indicates that the model performs consistently well in identifying defaults (high recall) while maintaining solid precision.
- **Moderate Performance on Test Set**: On the **test set**, the model achieves **recall (0.740)**, capturing most of the defaults (**54 true positives**), but **precision drops to 0.362**, reflecting a high number of **false positives (95)**. This trade-off indicates the model is favoring recall, leading to many non-defaults being classified as defaults.
- **Confusion Matrix Insights**: The confusion matrix shows that the model captures **54 true positives** but misclassifies **95 false positives** and has **19 false negatives**. The high number of false positives suggests that the LDA model sacrifices precision in favor of detecting more defaults.
- **Overall Performance**: The **F1 score (0.486)** and **ROC-AUC (0.792)** on the test set reflect the model's focus on recall, but the low precision shows that the model needs further tuning to balance false positives and improve overall performance.
- The LDA model is more focused on maximizing recall but suffers from a high rate of false positives, making it less ideal for situations where precision is crucial. Further tuning might help in improving the trade-off between recall and precision.

Model Comparison and Selection

This section evaluates the performance of Logistic Regression, Random Forest, and LDA models, leading to the selection of the best model based on recall, precision, and F1 score.

Model Comparison

Table 6: Combined Performance Metrics on Training Dataset

Model	Accuracy	Recall	Precision	F1	ROC_AUC
LR	0.915820	0.442177	0.656566	0.528455	0.930022
LR Optimal	0.896952	0.748299	0.511628	0.607735	0.930022
LR Optimal - SMOTE	0.843623	0.983753	0.768401	0.862843	0.939779
RF - Original	1.000000	1.000000	1.000000	1.000000	1.000000
RF - Balanced	1.000000	1.000000	1.000000	1.000000	1.000000
RF - Balanced - Tuned	0.952071	0.982128	0.926437	0.953470	0.952071
LDA - Balanced	0.859464	0.872461	0.850356	0.861267	0.859464

Table 7: Combined Performance Metrics on Test Dataset

Model	Accuracy	Recall	Precision	F1	ROC_AUC
LR	0.919118	0.506849	0.660714	0.573643	0.899212
LR Optimal	0.886765	0.684932	0.480769	0.564972	0.899212
LR Optimal - SMOTE	0.720588	0.917808	0.266932	0.413580	0.893097
RF - Original	0.923529	0.465753	0.723404	0.566667	0.722168
RF - Balanced	0.897059	0.643836	0.516484	0.573171	0.785674
RF - Balanced - Tuned	0.882353	0.780822	0.471074	0.587629	0.837693
LDA - Balanced	0.832353	0.739726	0.362416	0.486486	0.791609

Observations on Model Performance:

- **Logistic Regression (LR)**:
  - **Base Model**: The basic Logistic Regression model has decent performance on both training and test sets, with **F1 scores of 52.85% (train)** and **57.36% (test)**. However, the **recall is relatively low (50.68% on the test set)**, meaning it misses many default cases.
  - **Optimal Threshold**: When applying an optimal threshold, the recall improves to **68.49% on the test set**, but precision drops, indicating more false positives. The **F1 score on the test set** also decreases slightly compared to the base model.
  - **SMOTE Balanced**: While the recall becomes very high (**91.78% on the test set**), precision drops drastically to **26.69%**, and the F1 score plummets to **41.36%**, showing a significant trade-off where too many non-defaults are misclassified as defaults.
- **Random Forest (RF)**:
  - **Original Model**: The Random Forest model on the original data performs perfectly on the training set (**100% in all metrics**), but this indicates clear **overfitting**. On the test set, the model's recall is low (**46.57%**), meaning it misses more default cases, though precision is relatively high (**72.34%**).
  - **Balanced RF**: Balancing the dataset with SMOTE improves the **recall on the test set** to **64.38%**, but precision drops to **51.65%**. The **F1 score** is similar to the base Logistic Regression model, showing improvement in catching defaults but introducing more false positives.
  - **Tuned RF Model**: Tuning the balanced RF model improves the **recall to 78.08%** on the test set, while precision drops to **47.11%**. The F1 score is **58.76%**, showing a better balance between recall and precision compared to the untuned version. This model offers strong performance overall.
- **LDA (Linear Discriminant Analysis)**:
  - The LDA model trained on balanced data shows good recall (**73.97%** on the test set), but its precision is quite low (**36.24%**), leading to a low F1 score of **48.65%**. The model struggles to maintain a balance between capturing defaults and reducing false positives.

Recommendations on Model Selection:

- **Random Forest Tuned on Balanced Data** is the best-performing model overall, with the highest **recall** on the test set (**78.08%**) and a reasonable balance of **F1 score (58.76%)**. Although the precision is somewhat low, it outperforms other models in capturing default cases, making it ideal for **credit risk management** where identifying defaults is a priority. The tuned model also shows improved **ROC-AUC** of **83.77%**, indicating good discriminative power.
- **Logistic Regression with SMOTE** achieves high recall but suffers from extremely low precision (**26.69%**), making it less reliable due to the high number of false positives. This model may lead to inefficient business decisions with too many non-defaults flagged as risky.
- **LDA** offers reasonable recall but significantly underperforms in precision, making it less suitable for this use case compared to Random Forest.

Final Model Selection

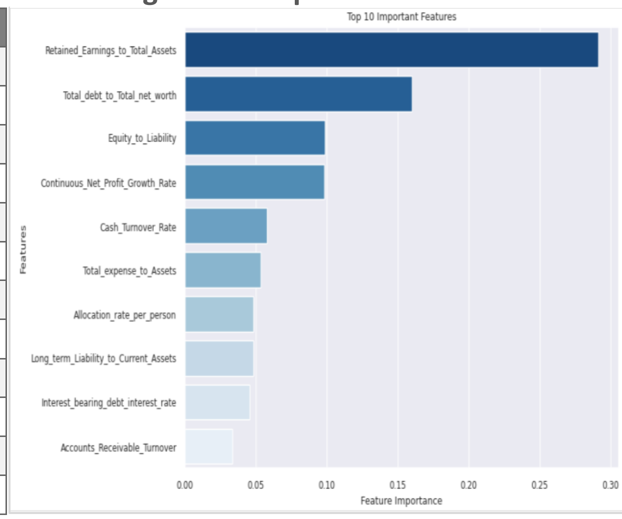
The **Random Forest Tuned Model** on **balanced data** is the best choice for this project, as it strikes a better balance between capturing defaults (high recall) and keeping false positives at a manageable level. For business purposes, this model minimizes missed defaults while maintaining a reasonable level of precision, making it the most robust model for **credit risk management**.

List of Important Features

Table 8: Important Features

Feature	Importance
Retained_Earnings_to_Total_Assets	0.291280
Total_debt_to_Total_net_worth	0.160105
Equity_to_Liability	0.098624
Continuous_Net_Profit_Growth_Rate	0.098357
Cash_Turnover_Rate	0.058030
Total_expense_to_Assets	0.053564
Allocation_rate_per_person	0.048431
Long_term_Liability_to_Current_Assets	0.048185
Interest_bearing_debt_interest_rate	0.045736
Accounts_Receivable_Turnover	0.033491
Quick_Assets_to_Total_Assets	0.032597
Research_and_development_expense_rate	0.031600

Figure 20: Important Features



Observations:

- **Retained Earnings to Total Assets (0.291)**: This is the most important feature in the model, indicating that companies with higher retained earnings relative to their assets are less likely to default. Strong retained earnings improve financial stability.
- **Total Debt to Total Net Worth (0.160)**: A higher debt-to-net-worth ratio suggests higher leverage, increasing the risk of default.
- **Equity to Liability (0.099)**: Companies with a higher equity-to-liability ratio are more likely to have a stable financial position, reducing the likelihood of default.
- **Continuous Net Profit Growth Rate (0.098)**: Consistent profit growth is a key indicator of financial health. Companies with higher growth rates are less likely to default, making this feature a strong predictor of financial stability.
- **Cash Turnover Rate (0.058)**: A higher cash turnover rate implies better efficiency in using cash to generate revenue, reducing default risk. This feature suggests that liquidity management is important in avoiding financial distress.
- **Total Expense to Assets (0.054)**: Higher expenses relative to assets can strain a company's financial resources, increasing the risk of default. This feature helps the model assess the company's cost structure.

These features together provide a comprehensive view of a company's financial health, with a strong emphasis on profitability, leverage, and liquidity management.

Insights and Recommendations

Key Insights

- **Imbalanced Default Distribution**: The dataset showed a significant imbalance between defaults and non-defaults, with far fewer defaults. This necessitated the use of **SMOTE** to balance the data and ensure the models could effectively learn from the minority class, leading to improved recall across models.

- **Feature Distributions and Outliers**: During the **EDA**, skewed distributions and outliers were observed in key financial metrics like **Retained Earnings to Total Assets** and **Total Debt to Total Net Worth**. These variables were critical in understanding financial health and were treated to ensure better model performance.

- **Correlation and Multicollinearity Management**: Correlation analysis showed strong relationships between variables like **Retained Earnings to Total Assets** and **Equity to Liability**. These were highly predictive of defaults, and redundant features were removed to reduce multicollinearity, improving model stability.

- **Retained Earnings to Total Assets**: This remained the top predictor of default risk throughout the model-building process. Companies with higher retained earnings relative to their assets had a lower likelihood of default, making this metric a critical factor in financial stability.

- **Total Debt to Total Net Worth**: This feature consistently indicated higher default risk for companies with higher leverage. It highlights the importance of managing debt levels relative to equity to reduce financial vulnerability.

- **Model Comparison and Performance**: The **Random Forest Tuned Model** on balanced data provided the best overall performance, with high recall and a balanced F1 score. This model was superior in identifying defaults compared to **Logistic Regression** and **LDA**, while also maintaining reasonable precision.

- **Trade-offs with SMOTE**: While SMOTE significantly improved recall in models like **Random Forest** and **Logistic Regression**, it led to a decrease in precision, increasing false positives. Careful management of false positives is needed, especially in cost-sensitive scenarios.

- **Threshold Optimization in Logistic Regression**: By adjusting the threshold to **0.2**, recall improved across both train and test sets, ensuring more defaults were captured. However, this led to a further drop in precision, which should be managed based on business needs.

- **Cash Flow and Liquidity**: Features like **Cash Turnover Rate** and **Total Expense to Assets** were key indicators of a company's liquidity and expense management. Efficient cash management was closely associated with lower default risks, making these metrics essential for financial assessment.

Business Recommendations

- **Prioritize Debt Management**: Encourage businesses to maintain healthy debt levels in relation to their net worth. Companies with high leverage are at a greater risk of default, so improving debt management strategies can significantly enhance financial stability.

- **Focus on Strengthening Retained Earnings**: Businesses should aim to increase retained earnings to serve as a buffer against potential financial distress. Retaining more profits, rather than excessive dividend payouts, can improve long-term resilience and reduce default risk.

- **Optimize Capital Structure**: Improving the **equity-to-liability ratio** is essential. Companies with a stronger equity base relative to liabilities are more financially stable, reducing their default risk and providing more flexibility for future growth.

- **Improve Profitability and Cost Management**: Companies need to ensure consistent profit growth while keeping a close watch on expenses. Operational efficiency, sales growth, and cost control are critical to avoiding defaults, as highlighted by important features in the model.

- **Cash Flow Monitoring**: Proactively managing **cash flow** and liquidity is crucial. Companies that efficiently manage cash turnover and control expenses relative to assets are less likely to default, making these metrics critical for long-term financial health.

- **Leverage Early Warning Systems**: Use the **Random Forest Tuned Model** as an early warning system to flag high-risk companies. Its strong recall ensures that defaults are identified early, enabling timely interventions to mitigate financial losses.

- **Manage False Positives for Strategic Decisions**: Given the increase in false positives from SMOTE, it is important to balance precision with recall, especially in cost-sensitive environments. Additional analysis or manual review may be necessary to filter out low-risk companies flagged as defaults.

- **Continuous Model Monitoring and Tuning**: Regularly monitor the model's performance using updated data and adjust the threshold or parameters as necessary. The business environment may shift, requiring recalibration of the model to maintain optimal performance.

- **Engage High-Risk Clients**: Utilize the insights gained from feature importance to engage with high-risk clients proactively. For example, companies with weak retained earnings or high debt levels could benefit from financial advisory services to strengthen their financial positions.

- **Expand Data for Robust Validation**: To ensure the model generalizes well across various business sectors, it's important to expand testing to additional datasets or real-world data. This will help refine the model's applicability and performance across different market conditions and industries.