Predicting Default Risk

FRA - Part A

Amneh Ghanem

Content/Agenda

- **Business Problem Overview**
- **Data Overview**
- **Data Cleaning**
- **Exploratory Data Analysis**
- **Inference Summary from EDA**
- **Logistic Regression Model (Building and Validation)**
- **Random Forest Model(Building and Validation)**
- **LDA Model (Building and Validation)**
- **Models Comparison**
- **Model Improvement**
- **Models Comparison After Improvement**
- **Feature Importance Analysis**
- **Conclusions and Recommendations**

Business Problem Overview

companies face financial risks that may lead to default if they fail to meet their debt obligations. A company's default can result in a lower credit rating, reduced access to future credit, and increased borrowing costs. From an investor's perspective, assessing a company's financial stability, ability to manage obligations, and potential for growth is crucial before making investment decisions.

The objective of this project is to leverage historical financial data to develop a predictive model for identifying companies at risk of default. This analysis will focus on financial indicators derived from balance sheets and income statements to determine the key factors influencing corporate default. The dependent variable ("Default") is predefined in the dataset, eliminating the need for additional feature engineering.

Data Overview

•Co Code → Unique numerical code assigned to each company.
• Co Name → Name of the company (e.g., Hind. Cables, Tata Tele. Mah.).
• Operating Expense Rate → Ratio of operating expenses to revenue.
• Research and development expense rate → Percentage of revenue spent on R&D._
• Cash flow rate → Measures liquidity and ability to generate cash.
• Interest bearing debt interest rate → Interest rate paid on loans/debt.
• Tax rate A → Company's applicable tax rate.
• Cash Flow Per Share → Cash flow available per share.
• Per Share Net profit before tax Yuan_ → Net profit per share before tax.
• Realized Sales Gross Profit Growth Rate → Growth rate of sales gross profit.
• Cash Flow to Equity → Cash flow available for shareholders.
• Current Liability to Current Assets → Ratio of short-term liabilities to short-term assets.
• Liability Assets Flag → Flag indicating high liabilities compared to assets.

Data Overview

•Total assets to GNP price → Total assets relative to Gross National Product.
• No credit Interval → Duration when the company had no credit access.
• Degree of Financial Leverage DFL → Measure of how financial leverage affects earnings.
• Interest Coverage Ratio Interest expense to EBIT → Ability to cover interest using EBIT (Earnings Before Interest & Taxes).
•Net Income Flag → Binary flag (1 = positive net income, 0 = negative net income).
• Equity to Liability → Ratio of shareholders' equity to liabilities.

- Target Variable (Prediction):Default → 0 (No Default) or 1 (Default)0 → The company did not default.1 → The company defaulted on its financial obligations.

Data Overview

- The dataset contains 2058 entries (rows) and 58 columns (features), including financial indicators, ratios, and a target variable
- 53 columns (float64) → Numerical values (financial metrics, ratios).
- 4 columns (int64) → Discrete values (e.g., flags, binary indicators).
- 1 column (object) → Company Name (Co_Name) (Categorical data).

Data Statistics

Now, let us check the basic measures of descriptive statistics for the continuous variables.

In [10]:
```
Default.describe()
```

Out[10]:

	Co_Code	_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_ra
count	2058.000000	2.058000e+03	2.058000e+03	2058.000000	2.058000e+03	2058.00
mean	17572.113217	2.052389e+09	1.208634e+09	0.465243	1.113022e+07	0.11
std	21892.886518	3.252624e+09	2.144568e+09	0.022663	9.042595e+07	0.15
min	4.000000	1.000260e-04	0.000000e+00	0.000000	0.000000e+00	0.00
25%	3674.000000	1.578727e-04	0.000000e+00	0.460099	2.760280e-04	0.00
50%	6240.000000	3.330330e-04	1.994130e-04	0.463445	4.540450e-04	0.03
75%	24280.750000	4.110000e+09	1.550000e+09	0.468069	6.630660e-04	0.21
max	72493.000000	9.980000e+09	9.980000e+09	1.000000	9.900000e+08	0.99

8 rows × 57 columns

Data Statistics

- Around 10.7% of companies defaulted most companies did not default.
- Operating & Expense Metrics Mean: 2.05 Billion, Std Dev: 3.25 Billion → High variance. Range: Min = 0.0001 | Max = 9.98 Billion that Suggests some companies have extremely high operating expenses.
- Research and development expense rate: Mean: 1.2 Billion, Std Dev: 2.14 Billion25% of companies had no R&D expenses, indicating many companies do not invest in research.
- Average tax rate is **11.47% and** Some companies pay no taxes.
- Realized Sales Gross Profit Growth Rate: Mean: 0.0228 Min: 0.0042, Max: 1 and Some companies show very high sales growth.
- Total Asset Growth Rate: Mean: 5.28 Billion, Max: ~9.98 Billion, High asset growth rate variation.
- Current Ratio (Liquidity Measure):Mean: 1.33 Million (Likely due to outliers).Std Dev: 60.6 Million indicates extreme variation; some companies might struggle with liquidity.
- Total Asset Turnover: Mean: 0.129, Min: 0; Some companies fail to generate revenue from assets.

Data Cleaning

› The dataset has missing values in the following columns:

- Cash Flow Per Share (167 missing values)
- Cash to Total Assets (96 missing values)
- Total debt to Total net worth" (21 missing values)
- Current Liability to Current Assets (14 missing values)

Handling Missing Values:

- These are all financial ratio-based features, meaning they are continuous numerical variables.
- Imputation with median is typically the best approach, as financial data often contains outliers, making the mean less reliable.
- If a column had more than 40% missing values, we would consider dropping it. However, in this case, all missing values are within a manageable range.

Data Cleaning

- **Outlier Treatment**

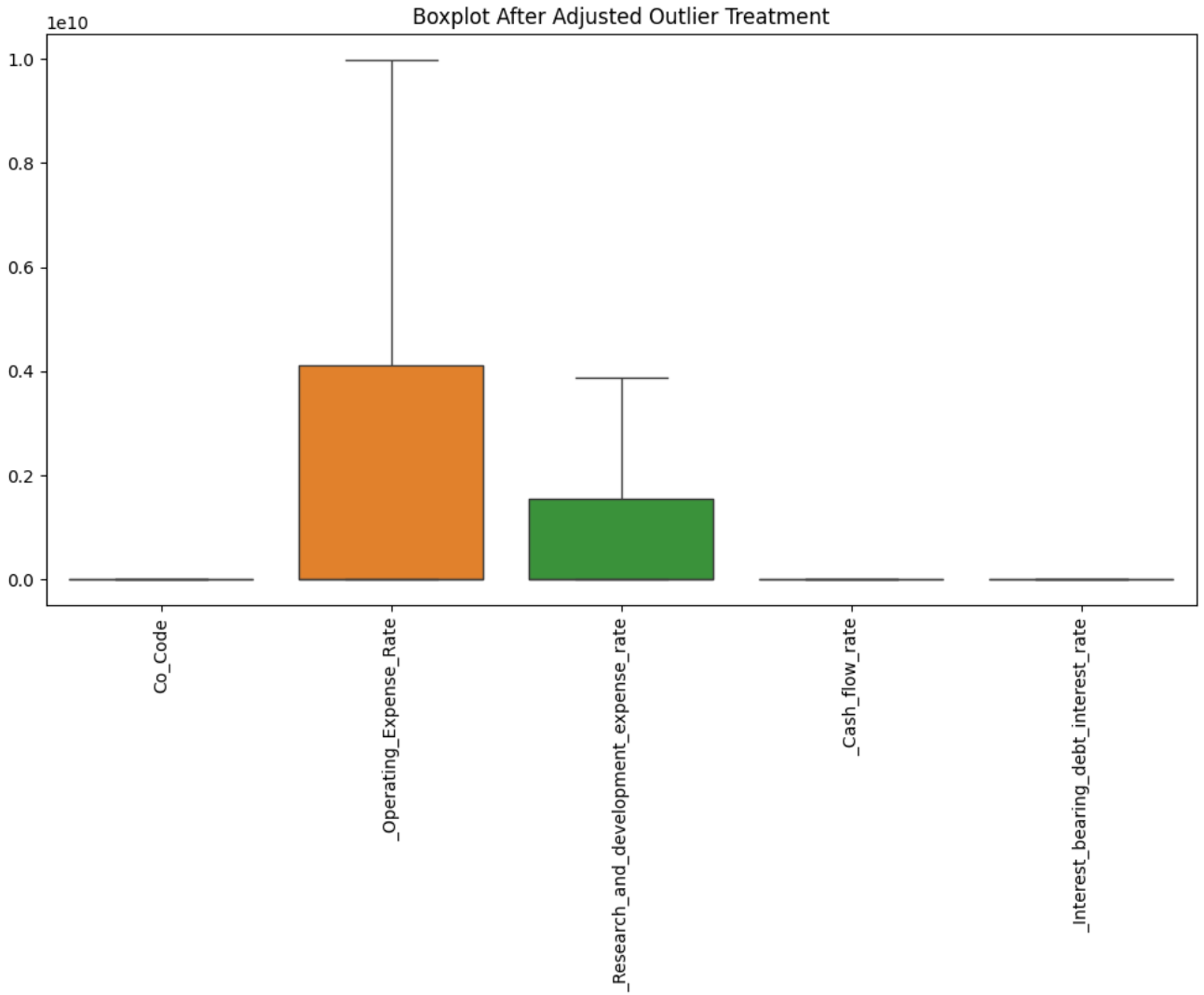
```
_Fixed_Assets_Turnover_Frequency                    501
_Current_Asset_Turnover_Rate                        464
_Degree_of_Financial_Leverage_DFL                   438
_Cash_Flow_to_Liability                             407
_No_credit_Interval                                 396
_Interest_Coverage_Ratio_Interest_expense_to_EBIT   376
_Operating_profit_per_person                        357
_Continuous_Net_Profit_Growth_Rate                  340
_Interest_Expense_Ratio                             328
_Operating_Profit_Growth_Rate                       317
dtype: int64
```

Data Cleaning

Outliers can significantly affect machine learning models and statistical analyses, so handling them properly is crucial

Steps for Outlier Treatment in Financial Data

1. Identify Outliers
2. Choose an Outlier Handling Strategy: Capping (Winsorization)
3. Check Before vs. After treatment using



DATA CLEANING

- The boxplot no longer shows extreme outliers outside the whiskers.
- The distribution looks more balanced, improving model performance.
- The data retains its integrity while removing unwanted extreme deviations.

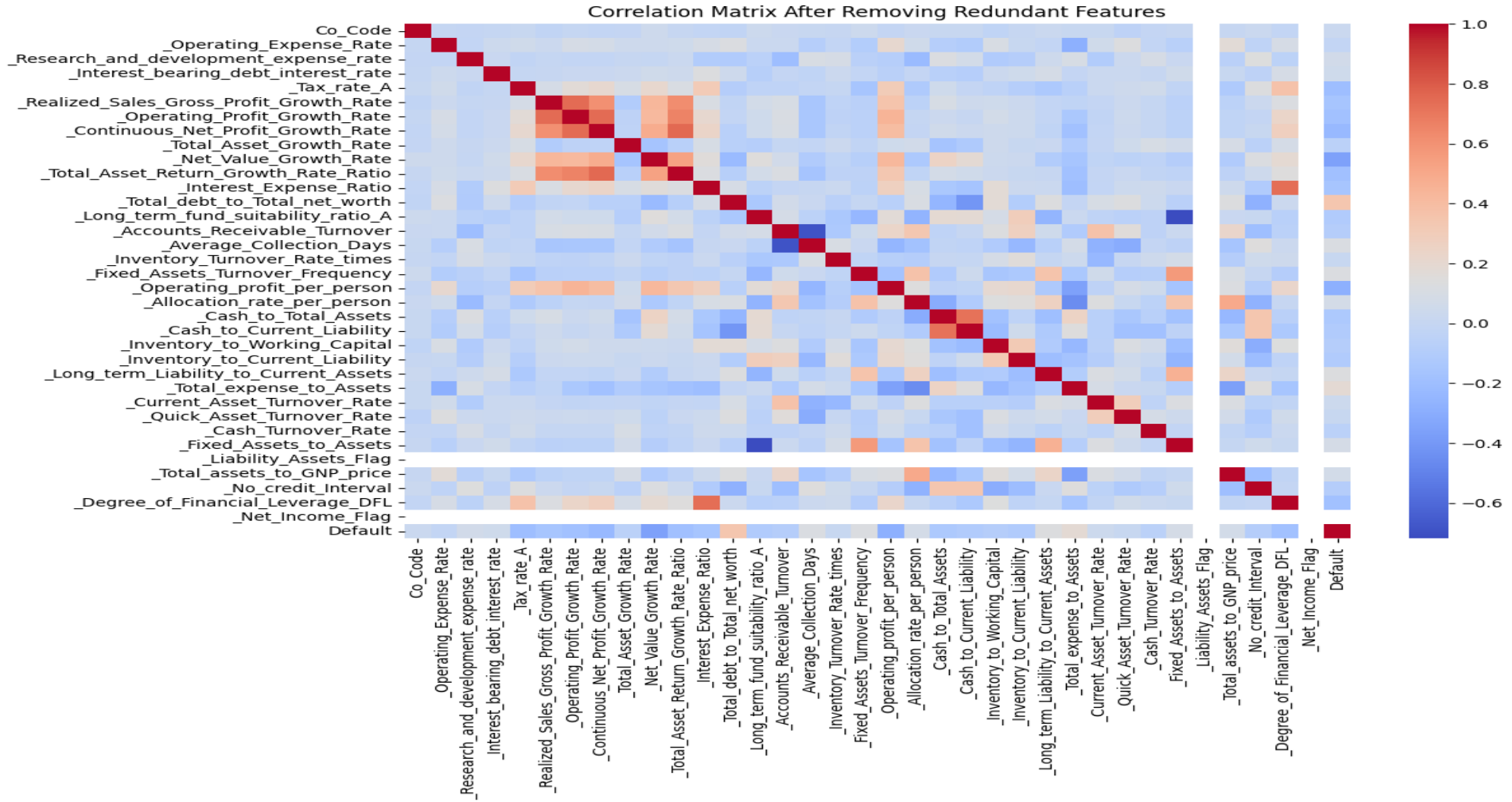
Data Cleaning

Eliminating redundant variables:

To systematically remove redundant variables, we will follow these steps:

- Check Multicollinearity Using a Correlation Matrix
- Identify highly correlated variables (threshold > 0.85).
- Retain only one variable from each correlated pair.
- Use Variance Inflation Factor (VIF) to Confirm Redundancy
- VIF detects multicollinearity among independent variables.
- VIF > 5 indicates a high correlation with other predictors.
- Drop Redundant Features Based on Correlation and VIF

Data Cleaning



Data Cleaning

Observations from the Heatmap:

- Diagonal is always 1 (self-correlation).
- Some features have high correlations, meaning one can be removed.
- redundant features:
- Total Asset Growth Rate and Net Value Growth Rate are Both measure asset growth.
- Quick Asset Turnover Rate and Current Asset Turnover Rate are Similar liquidity measures.
- Interest Coverage Ratio Interest expense to EBIT and Interest Expense Ratio are Related to debt obligations.

Data Cleaning

List of redundant features identified from high correlation analysis

redundant_features = [

'Co_Code','_Net_profit_before_tax_to_Paid_in_capital', '_Quick_Assets_to_Total_Assets', '_Cash_flow_rate',

'_Per_Share_Net_profit_before_tax_Yuan_', '_Current_Liability_to_Current_Assets', '_Cash_Flow_to_Equity',

'_Cash_Reinvestment_perc', '_Quick_Assets_to_Current_Liability', '_Operating_Funds_to_Liability',

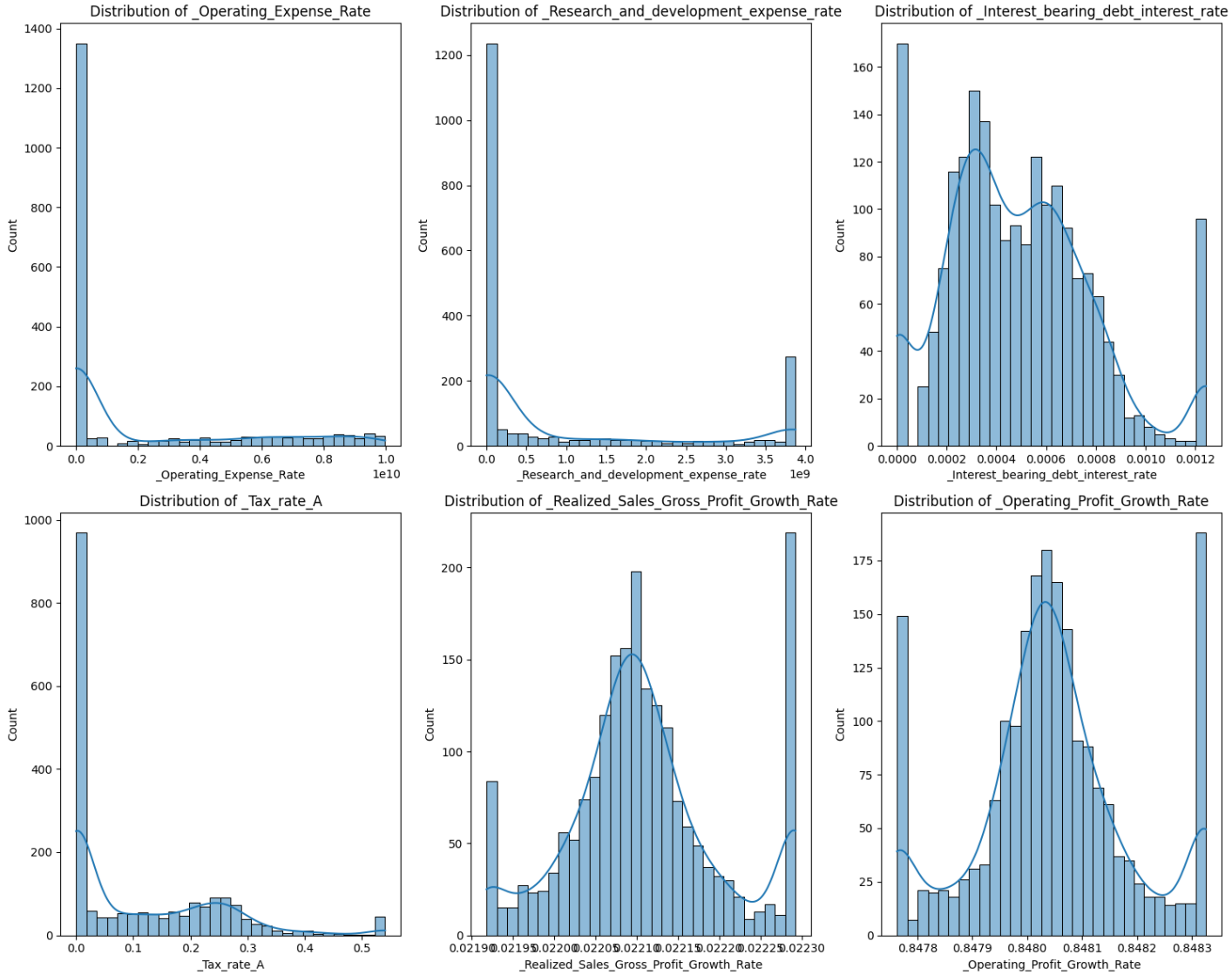
'_Total_Asset_Turnover', '_Retained_Earnings_to_Total_Assets', '_Interest_Coverage_Ratio_Interest_expense_to_EBIT',

'_Net_Worth_Turnover_Rate_times', '_Total_income_to_Total_expense', '_Cash_Flow_Per_Share', '_Quick_Ratio',

'_Cash_Flow_to_Total_Assets', '_CFO_to_Assets', '_Equity_to_Liability', '_Cash_Flow_to_Liability',

Now we have 36 columns

UNIVARIATE ANALYSIS:

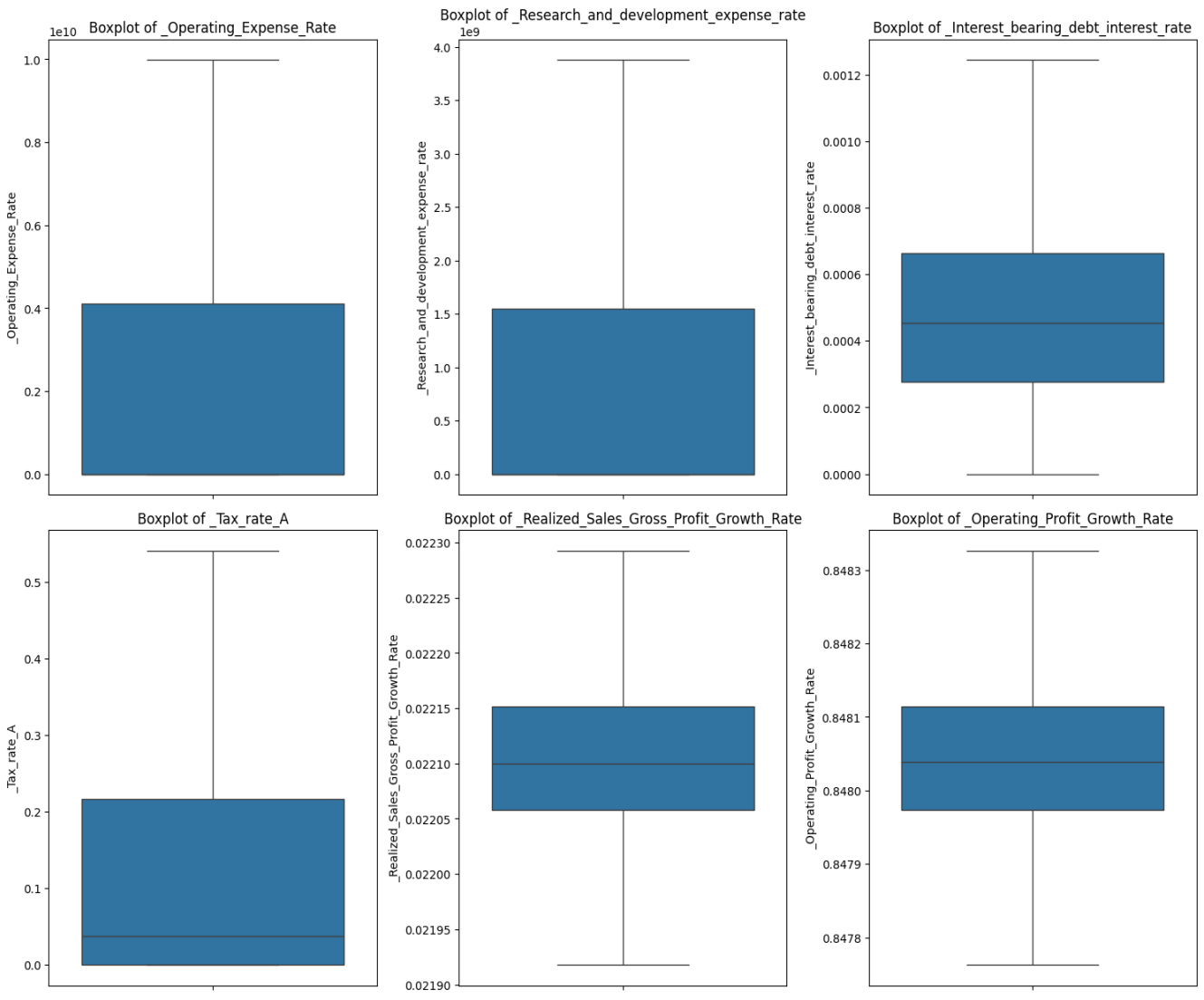


EXPLORATORY DATA ANALYSIS

The histograms show the distribution of six key financial variables:

- Operating Expense Rate right skewed, Most companies have very low operating expenses, but a few have extremely high values.
- Interest bearing debt interest rate : Moderately skewed with multiple peaks bimodal distribution. This suggests two groups of companies (low and high debt interest).
- Operating Profit Growth Rate Slightly bimodal distribution, meaning companies may fall into two different performance groups.

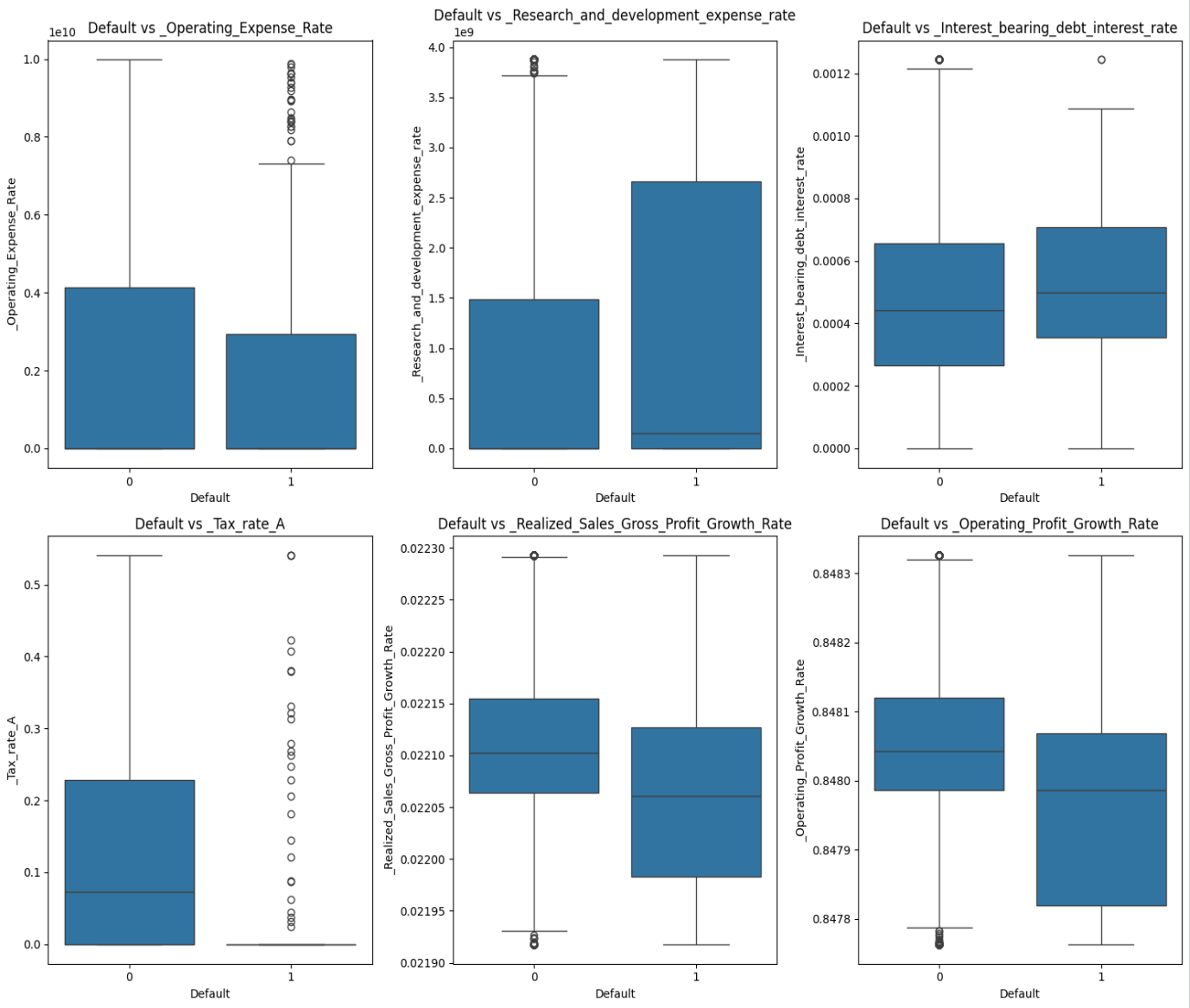
UNIVARIATE ANALYSIS:



EXPLORATORY DATA ANALYSIS

Boxplot after outlier treatment with more balanced distribution:

- Interest bearing debt interest rate compared to other financial indicators. The data is more evenly distributed, but some companies have very high debt
- Operating Expense Rate the upper whisker is shorter than before, indicating outliers were capped and Less extreme values
- Research and development expense rate Similar pattern to Operating Expense Rate right skewed, Large variation in R&D spending across companies.
- Tax rate A Many companies have very low tax rates (left-skewed).A small number of companies have significantly higher tax rates .

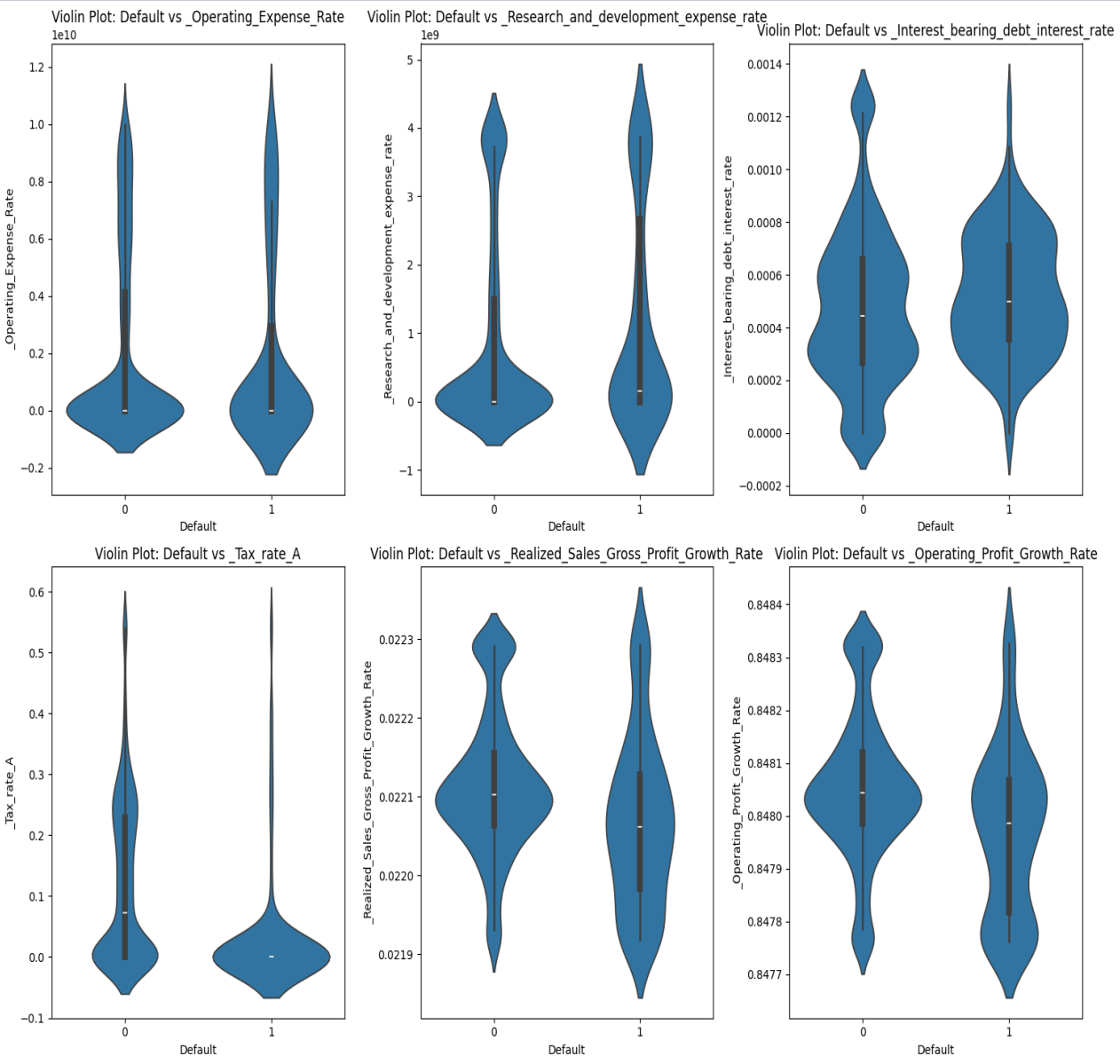


EXPLORATORY DATA ANALYSIS: BIVARIATE ANALYSIS

Feature Distributions for Default Categories:

- Interest bearing debt interest rate Defaulted companies tend to have slightly higher debt interest rates. Higher debt costs might contribute to financial distress, but the difference is not extreme.
- Non defaulted companies have higher median operating expenses and Defaulted companies have more extreme values Some defaulted companies had very high expenses, potentially leading to financial distress.
- Research and development expense rate Non-defaulted companies have higher median R&D spending. Defaulted companies show lower R&D investment overall. Companies that invest more in R&D may be financially stronger and less likely

Bivariate Analysis

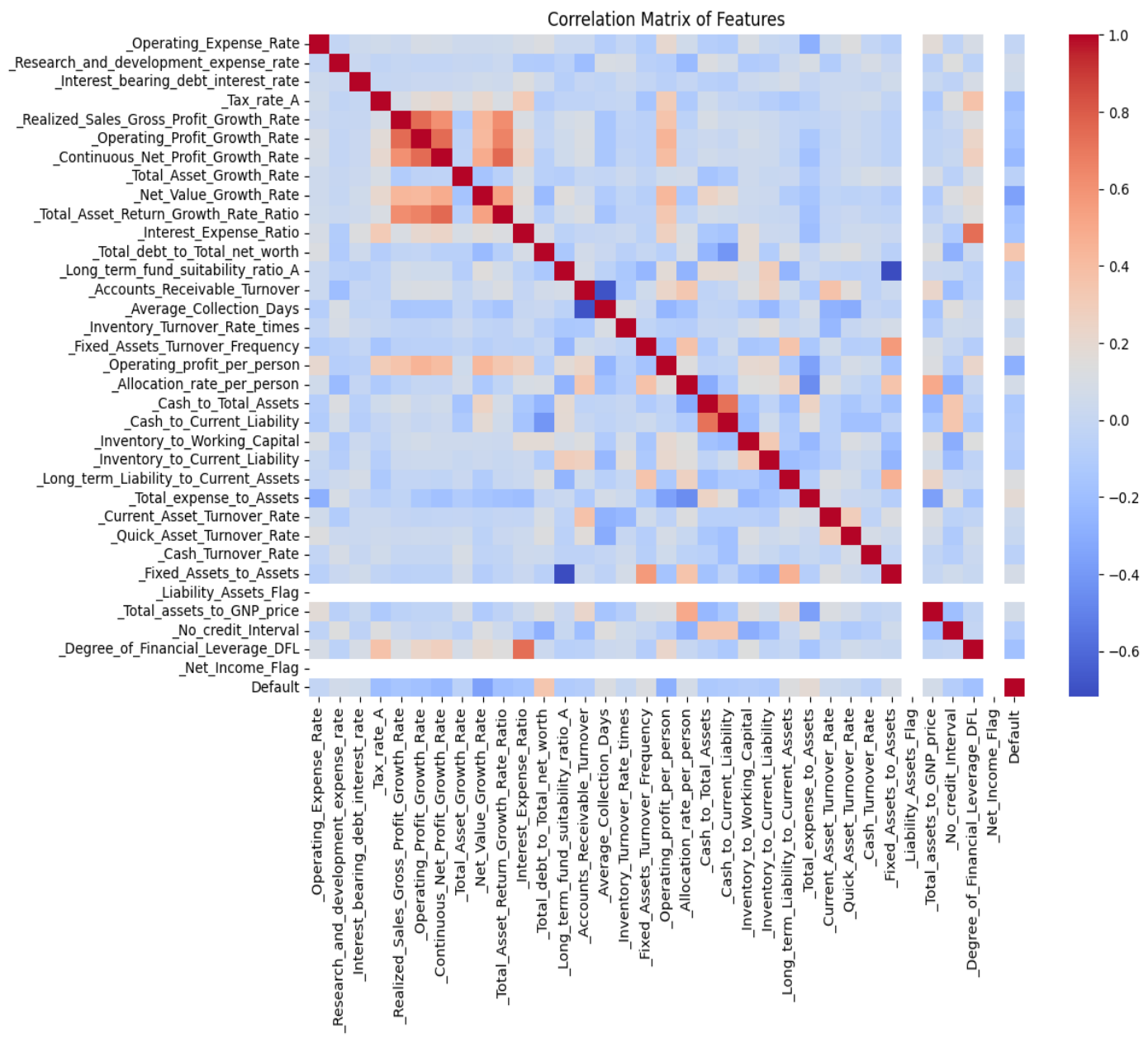


EXPLORATORY DATA ANALYSIS:

Violin plots provide insight into data distribution, density, and spread for defaulted and non-defaulted companies:

- Operating Profit Growth Rate Both groups have similar distributions, but non-defaulted companies have slightly higher profit growth. Higher profit growth slightly reduces default risk.
- Tax rate A Defaulted companies have more density at low tax rates. Non-defaulted companies show a slightly higher tax burden. Companies that pay little or no tax may be in financial distress.
- Operating Expense Rate: Wide distribution for both defaulted and non-defaulted companies. Higher density near lower values for both categories. Defaulted companies show more extreme high values. Most companies have low operating expenses, but some high-expense firms' default.
- Research and development expense rate: Non-defaulted companies have higher median R&D spending. Defaulted companies show lower density in high R&D spending. Companies with higher R&D spending are less likely to default.

Multivariate Analysis



EXPLORATORY DATA ANALYSIS: MULTIVARIATE ANALYSIS

The heatmap provides insights into how different financial variables correlate with each other and with the Default variable.

- Total expense to Assets, Higher expense ratios indicate financial, Companies with unsustainable expenses may struggle with profitability and loan repayments.
- Interest bearing debt interest rate ,Higher debt interest rates are linked to higher default risk. companies paying higher interest likely have weaker credit ratings or increased financial risk.
- Research and development expense rate, Higher R&D spending correlates with a lower chance of default. Companies that invest in innovation and future growth are more financially stable.
- Operating Profit Growth Rate, Higher profit growth reduces default risk. Strong profit growth suggests a sustainable business model.

Inference Summary from EDA

- Higher Interest-Bearing Debt Interest Rate Increases Default Risk Companies with high debt interest rates are more likely to default.
- Higher R&D Spending Reduces Default Risk Companies investing more in research & development (R&D) are less likely to default.
- High Operating Expenses (for Some) Default Risk Increases Most companies have low expenses, but some defaulted firms show extreme high operating expenses.
- Low Tax Rate Sign of Financial Distress Many defaulted companies pay little or no tax, indicating potential financial instability.
- Higher Operating Profit Growth Rate Reduces Default Risk Companies with higher profit growth tend to survive financial difficulties.

Logistic Regression Model (Building and Validation)

Feature Selection:

- Identifying Key Predictors
- Ensuring the target variable 'Default' exists
- Defining features (X) and target (y)
- Selecting important features
- Feature Importance using Random Forest
- Intersection of RFE & Random Forest

Logistic Regression Model: Feature Selecting

Top Features Selected by RFE:
['_Operating_Expense_Rate', '_Research_and_development_expense_rate', '_Operating_Profit_Growth_Rate', '_Total_Asset_Growth_Rate', '_Interest_Expense_Ratio', '_Inventory_Turnover_Rate_times', '_Quick_Asset_Turnover_Rate', '_Cash_Turnover_Rate', '_No_credit_Interval', '_Net_Income_Flag']

Top Features Selected by Random Forest:
['_Net_Value_Growth_Rate', '_Total_debt_to_Total_net_worth', '_Interest_Expense_Ratio', '_Inventory_to_Working_Capital', '_Degree_of_Financial_Leverage_DFL', '_Operating_profit_per_person', '_Long_term_fund_suitability_ratio_A', '_Cash_to_Current_Liability', '_Cash_to_Total_Assets', '_Total_Asset_Return_Growth_Rate_Ratio']

Final Selected Features for Modeling:
['_Net_Income_Flag', '_Net_Value_Growth_Rate', '_Total_Asset_Growth_Rate', '_Cash_to_Total_Assets', '_Inventory_Turnover_Rate_times', '_Operating_profit_per_person', '_Degree_of_Financial_Leverage_DFL', '_Interest_Expense_Ratio', '_Cash_Turnover_Rate', '_Quick_Asset_Turnover_Rate', '_Total_Asset_Return_Growth_Rate_Ratio', '_Operating_Profit_Growth_Rate', '_Operating_Expense_Rate', '_Inventory_to_Working_Capital', '_Cash_to_Current_Liability', '_Long_term_fund_suitability_ratio_A', '_Research_and_development_expense_rate', '_Total_debt_to_Total_net_worth', '_No_credit_Interval']
Final Dataset Shape: (2058, 19)

Logistic Regression Model (Building and Validation)

- Train-Test Split: We will split the dataset into 67% training and 33% testing using train_test_split to ensure that we train the models on one part of the data and test them on unseen data.
- Build Logistic Regression Model: We will use stats models to build a Logistic Regression model, the optimal cut-off will be determined using the ROC Curve.
- Dropping Highly Correlated Features
- Removing Features with zero variance
- Add Constant for Intercept
- Train Logistic Regression

Logistic Regression Model (Building and Validation)

```
Optimization terminated successfully.
         Current function value: 0.332871
         Iterations 6
```

✅ Updated Logistic Regression Model Summary:

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                Default   No. Observations:                 1378
Model:                          Logit   Df Residuals:                     1371
Method:                           MLE   Df Model:                            6
Date:                Thu, 13 Feb 2025   Pseudo R-squ.:                 0.01955
Time:                        19:08:37   Log-Likelihood:                -458.70
converged:                       True   LL-Null:                       -467.84
Covariance Type:            nonrobust   LLR p-value:                  0.005531
================================================================================================
                                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                             -1.8620      0.201     -9.281      0.000      -2.255      -1.469
_Total_Asset_Growth_Rate       -4.546e-11   2.89e-11     -1.574      0.116   -1.02e-10    1.12e-11
_Inventory_Turnover_Rate_times  4.672e-12   2.83e-11      0.165      0.869   -5.07e-11    6.01e-11
_Cash_Turnover_Rate            -7.985e-11   3.43e-11     -2.330      0.020   -1.47e-10   -1.27e-11
_Quick_Asset_Turnover_Rate      3.23e-12    2.61e-11      0.124      0.901   -4.79e-11    5.43e-11
_Operating_Expense_Rate        -2.729e-11   2.81e-11     -0.970      0.332   -8.24e-11    2.79e-11
_Research_and_development_expense_rate  1.763e-10   5.54e-11      3.184      0.001    6.78e-11    2.85e-10
================================================================================================
```

Logistic Regression Model (Building and Validation)

- Removing High p-Value Features
- High p-values (> 0.05) mean the feature has little to no effect on the target variable (Default).
- Keeping irrelevant features can reduce model interpretability and performance.
- Removing them helps improve the model's statistical reliability and predictive power.
- Steps to Remove High p-Value Features and Re-run the Model
- Identifying Features with p > 0.05

From model summary, these features have high p-values:

- _Total_Asset_Growth_Rate (p = 0.116)
- _Inventory_Turnover_Rate_times (p = 0.869)
- _Quick_Asset_Turnover_Rate (p = 0.901)
- _Operating_Expense_Rate (p = 0.332)

Logistic Regression Model (Building and Validation)

```
Optimization terminated successfully.
         Current function value: 0.334140
         Iterations 6
```

✅ Updated Logistic Regression Model Summary:

Logit Regression Results

Dep. Variable:	Default	No. Observations:	1378
Model:	Logit	Df Residuals:	1375
Method:	MLE	Df Model:	2
Date:	Thu, 13 Feb 2025	Pseudo R-squ.:	0.01582
Time:	19:09:56	Log-Likelihood:	-460.44
converged:	True	LL-Null:	-467.84
Covariance Type:	nonrobust	LLR p-value:	0.0006117

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1185	0.131	-16.196	0.000	-2.375	-1.862
_Cash_Turnover_Rate	-8.306e-11	3.42e-11	-2.432	0.015	-1.5e-10	-1.61e-11
_Research_and_development_expense_rate	1.776e-10	5.5e-11	3.230	0.001	6.99e-11	2.85e-10

Logistic Regression Model (Building and Validation)

evaluating a trained machine learning model to ensure it performs well on unseen data. It helps assess:

- Accuracy: How often the model makes correct predictions.
- Precision & Recall: How well the model identifies actual defaults.
- ROC-AUC Score: The model’s ability to distinguish between defaulters and non-defaulters.
- Confusion Matrix: The breakdown of correct and incorrect classifications.

Logistic Regression Model (Building and Validation)

› Validate the Logistic Regression Model on Test Data

✅ Logistic Regression Model Performance on Test Data:

	precision	recall	f1-score	support
0	0.89	1.00	0.94	607
1	0.00	0.00	0.00	73
accuracy			0.89	680
macro avg	0.45	0.50	0.47	680
weighted avg	0.80	0.89	0.84	680

ROC-AUC Score: 0.5576267743901063

✅ Confusion Matrix:

```
[[607   0]
 [ 73   0]]
```

Logistic Regression Model (Building and Validation)

- Accuracy is 89% but only because it predicts non-default cases correctly.
- ROC-AUC Score: 0.557 Poor discrimination ability between default and non-default.
- The model completely fails to predict default cases (1), as seen from the 0% recall and precision for class 1.
- The model predicts all companies as No Default (0).
- 73 actual default cases were completely ignored.
- Not a useful model for risk prediction, as it fails to identify defaulting companies.

Random Forest Model(Building and Validation)

- Model Initialization by n_estimators=100 ,The model will train 100 decision trees and use their aggregated output.
- random_state=42 to Ensure reproducibility, so results remain consistent across runs.
- Using multiple trees reduces variance to avoid overfitting.
- Random Forest captures non-linear relationships in the data. Handles class imbalance better compared to traditional models.
- Train the Model: (X_train) Training dataset containing financial features,(y_train) Corresponding labels (0 = No Default, 1 = Default).
- Make Predictions: Predicts class labels (0 or 1) for the test dataset (X_test).
- Predict Probabilities: (predict_proba X_test) returns probabilities instead of class labels.

Random Forest Model(Building and Validation)

Model Validation

- classification_report(y_test, y_pred_rf) generates key evaluation metrics:
- Precision :How many predicted defaults (1) were actual defaults?
- Recall :How many actual defaults (1) were correctly predicted?
- F1-Score :Harmonic mean of precision and recall (balances both metrics).
- Support : Number of instances per class (0 = No Default, 1 = Default).

Random Forest Model(Building and Validation)

✅ Random Forest Model Performance:

	precision	recall	f1-score	support
0	0.89	0.94	0.91	607
1	0.13	0.08	0.10	73
accuracy			0.84	680
macro avg	0.51	0.51	0.51	680
weighted avg	0.81	0.84	0.83	680

ROC-AUC Score: 0.5792241204215658

Random Forest Model(Building and Validation)

Key Observations:

- Model Accuracy is 84% Seems high, but accuracy is misleading due to class imbalance.

Class 0 (No Default)

- Precision is 89% of predicted non-default cases were correct.
- Recall is 94% of actual non-defaults were correctly classified.
- F1-Score (0.91): Good balance of precision & recall for non-defaults.

Class 1 (Default) Performance:

- Precision Only 13% of predicted default cases were actually defaults.
- Recall Model identified only 8% of actual defaults.
- F1-Score (0.10): Very poor, indicating the model fails to capture default cases.
- The model still fails to predict most default cases (1), as seen from the low recall (0.08).

LDA Model (Building and Validation)

- Model Initialization: Initializes an LDA classifier with default parameters, It learns how to separate default (1) and non-default (0) cases by maximizing the separation between their distributions.
- Train the LDA Model: Training dataset containing financial features (X_train) , Corresponding labels 0 = No Default, 1 = Default (y_train).
- Make Predictions on Test Data: Predicts class labels (0 or 1) for the test dataset (X_test).
- y_pred_lda[i] = 0: Company is predicted to NOT default.
- y_pred_lda[i] = 1: Company is predicted to default.

LDA Model (Building and Validation)

- **Model Validation:**

Generates classification metrics:

- Precision: How many predicted defaults (1) were defaults.
- Recall :How many actual defaults (1) were correctly identified.
- F1-Score: Harmonic mean of precision and recall.
- Support : Number of instances per class (0 = No Default, 1 = Default).
- Compute ROC-AUC Score: Measures the model’s ability to differentiate Default (1) vs. No Default (0).

LDA Model (Building and Validation)

✅ LDA Model Performance:

	precision	recall	f1-score	support
0	0.89	1.00	0.94	607
1	0.00	0.00	0.00	73
accuracy			0.89	680
macro avg	0.45	0.50	0.47	680
weighted avg	0.80	0.89	0.84	680

ROC-AUC Score: 0.5616099839768907

LDA Model (Building and Validation)

Key Observations:

› Accuracy is 89% : Misleading metric due to class imbalance.

Class 0 (No Default):

› Precision is 89% of predicted non-default cases were correct.

› Recall (1.00): 100% of actual non-defaults were correctly identified.

› F1-Score (0.94) → Strong performance in predicting No Default cases.

Class 1 (Default) :

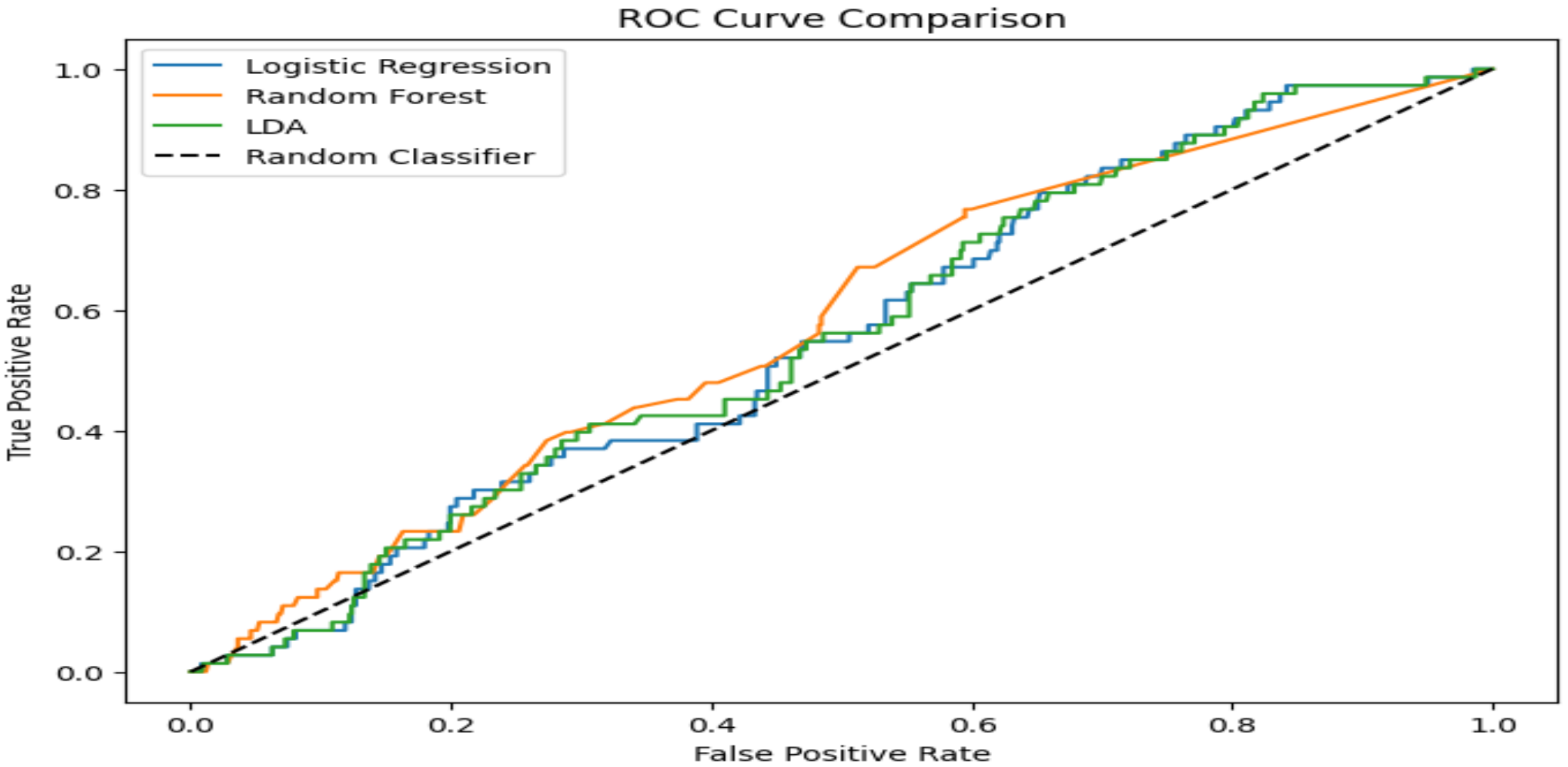
› Precision (0.00) : The model never predicts defaults correctly.

› Recall (0.00) :The model fails to identify any actual default cases.

› F1-Score (0.00): Indicates the model completely ignores the default class.

› The model predicts all cases as No Default (0), ignoring actual defaults.

Models Comparison



Models Comparison

- The Random Forest model (Orange Line) is slightly above the other models.
- Better at detecting defaults than Logistic Regression & LDA.
- Captures more complex patterns than linear models.
- Still not a strong classifier (close to random guessing at some points).
- Logistic Regression and LDA Have Almost Identical Performance
- Both models (Blue & Green Lines) are very close.
- Neither model effectively separates default from non-default cases. Their curves stay near the 45-degree line (random guessing).
- Poor ability to distinguish between risky and non-risky firms.

Model Improvement

```
Optimization terminated successfully.
         Current function value: 0.620223
         Iterations 5

 Updated Logistic Regression Model Summary (After SMOTE):

                           Logit Regression Results
==============================================================================
Dep. Variable:                Default   No. Observations:                 1846
Model:                          Logit   Df Residuals:                     1839
Method:                           MLE   Df Model:                            6
Date:                Fri, 14 Feb 2025   Pseudo R-squ.:                 0.02540
Time:                        10:32:42   Log-Likelihood:                -1144.9
converged:                       True   LL-Null:                       -1174.8
Covariance Type:            nonrobust   LLR p-value:                 5.214e-11
================================================================================================
                                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                                -0.3888      0.124     -3.146      0.002      -0.631      -0.147
_Total_Asset_Growth_Rate          -2.903e-11   1.82e-11     -1.599      0.110   -6.46e-11    6.55e-12
_Inventory_Turnover_Rate_times    -2.343e-11   1.72e-11     -1.359      0.174   -5.72e-11    1.04e-11
_Cash_Turnover_Rate               -9.861e-11      2e-11     -4.920      0.000   -1.38e-10   -5.93e-11
_Quick_Asset_Turnover_Rate          1.11e-12    1.5e-11      0.074      0.941   -2.84e-11    3.06e-11
_Operating_Expense_Rate           -4.094e-11   1.61e-11     -2.541      0.011   -7.25e-11   -9.36e-12
_Research_and_development_expense_rate  1.867e-10   3.39e-11      5.506      0.000     1.2e-10    2.53e-10
================================================================================================
```

Model Improvement

Observations from the Updated Logistic Regression Model

- The model is statistically significant (LLR p-value < 0.05).
- Some features have significant coefficients (p < 0.05), meaning they contribute to predicting default.

Issues:

- Some features still have high p-values (> 0.05) and should be removed for better performance.
- Pseudo R^2 is low, indicating the model still needs better predictors.

Removing Features with High p-values

› We will drop the following features and re-train the model:

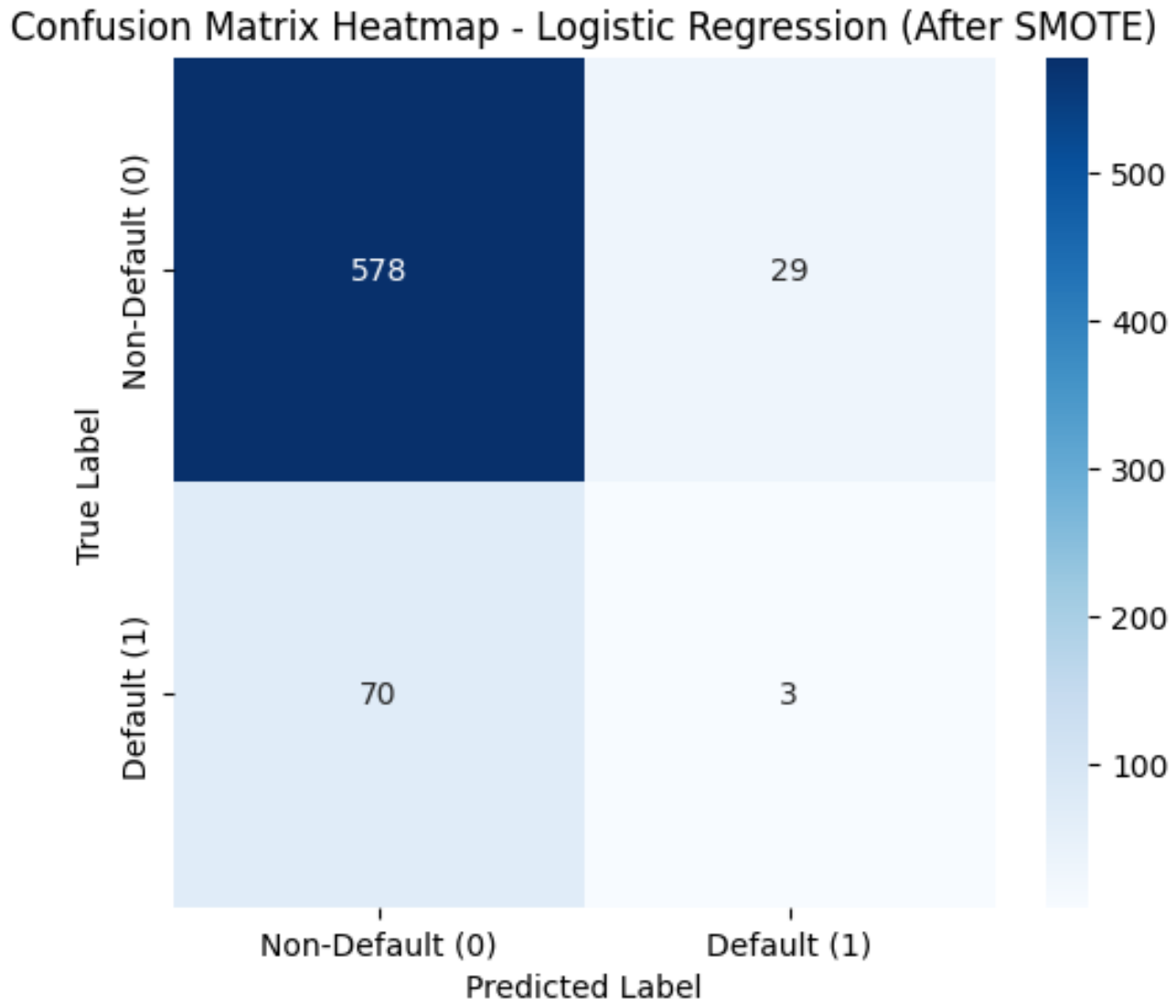
- _Total_Asset_Growth_Rate (p = 0.110)
- _Inventory_Turnover_Rate_times (p = 0.174)
- _Quick_Asset_Turnover_Rate (p = 0.941)

Model Improvement

```
Optimization terminated successfully.
         Current function value: 0.621452
         Iterations 5
```

✅ Updated Logistic Regression Model Summary (After Feature Refinement):

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                Default   No. Observations:                 1846
Model:                          Logit   Df Residuals:                     1842
Method:                           MLE   Df Model:                            3
Date:                Fri, 14 Feb 2025   Pseudo R-squ.:                 0.02347
Time:                        10:43:44   Log-Likelihood:                -1147.2
converged:                       True   LL-Null:                       -1174.8
Covariance Type:            nonrobust   LLR p-value:                 6.390e-12
================================================================================================
                                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                                -0.5674      0.080     -7.111      0.000      -0.724      -0.411
_Cash_Turnover_Rate                -1.008e-10   1.99e-11     -5.069      0.000    -1.4e-10   -6.18e-11
_Operating_Expense_Rate            -4.091e-11   1.59e-11     -2.577      0.010    -7.2e-11   -9.79e-12
_Research_and_development_expense_rate  1.804e-10   3.35e-11      5.385      0.000    1.15e-10    2.46e-10
================================================================================================
```



MODEL IMPROVEMENT

Analysis of Logistic Regression Model Performance (After Feature Refinement):

Logistic Regression model has been re-evaluated after refining features, but the results still indicate poor detection of defaulters.

Logistic Regression is Failing to Detect Defaulters (Low Recall: 4%):Only 3 out of 73 actual defaulters were correctly classified,This means 70 defaulters were misclassified as non-defaulters, This is a critical issue because the goal is to detect risky customers.

High Precision for Non-Defaulters, But Useless for Defaulters, The model is good at identifying non-defaulters (95% recall). But it sacrifices defaulter detection (only 4% recall), making it highly biased towards predicting "No Default" (0).This is likely due to class imbalance, even after applying SMOTE.

Confusion Matrix Confirms the Problem

578 non-defaulters correctly identified (TN) → Good

Only 3 defaulters correctly identified (TP) → Very bad recall (4%)

70 defaulters misclassified as non-defaulters (FN) → Huge issue!

29 non-defaulters incorrectly flagged as defaulters (FP) → Some false alarms, but not the main issue.

Model Improvement

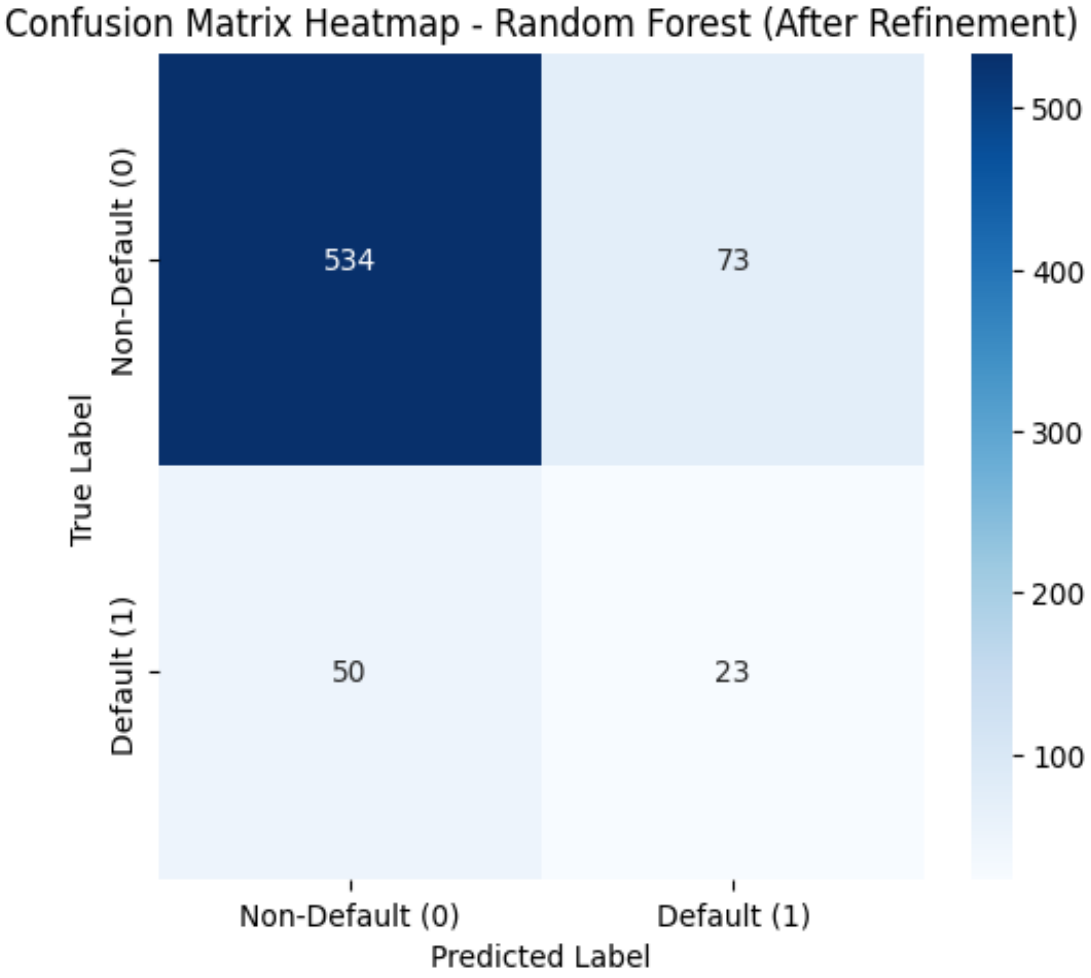
- **Random Forest using SMOTE-balanced training data:**

✅ Random Forest Model Performance (After SMOTE):

	precision	recall	f1-score	support
0	0.91	0.96	0.94	607
1	0.41	0.22	0.29	73
accuracy			0.88	680
macro avg	0.66	0.59	0.61	680
weighted avg	0.86	0.88	0.87	680

ROC-AUC Score: 0.734738552503893

Model Improvement



- Interpretation of Random Forest Model Results (After SMOTE)
- Random Forest model has improved significantly compared to Logistic Regression. Let's analyze the results:
- Overall Accuracy is Strong (88%):The model correctly classifies most cases, but accuracy alone is misleading because of class imbalance.
- Improved ROC-AUC Score (0.734):A good improvement over Logistic Regression (0.567), indicating better separation between defaulters and non-defaulters.
- Recall for Defaulters (22%) is Low :The model misses a lot of actual defaulters (false negatives).This means banks or financial institutions using this model might still issue risky loans to defaulters.
- Precision for Defaulters (41%) is Low: When the model predicts a default, it is correct only 41% of the time.This means there are many false positives, potentially rejecting customers who are not actually risky.

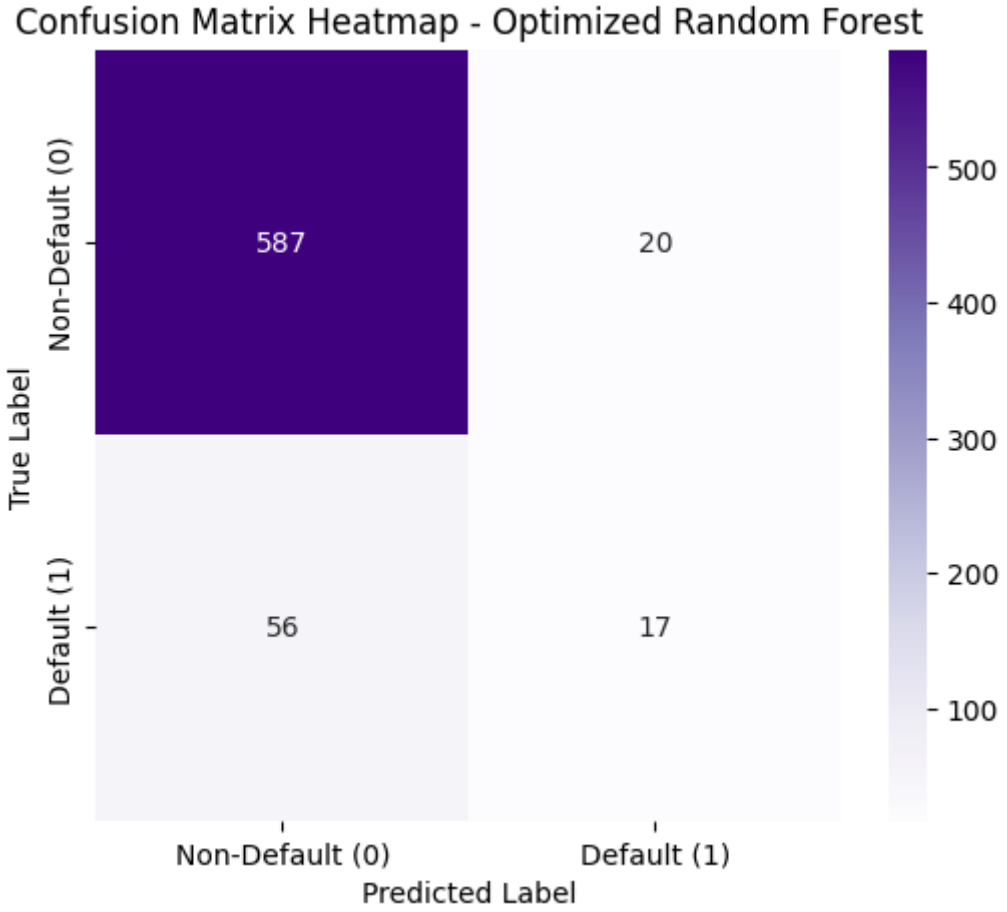
Model Improvement

Further Random Forest Improvement:

- Hyperparameter Tuning for Random Forest: Optimize n_estimators, max_depth, min_samples_split to reduce false positives.Using GridSearchCV or RandomizedSearchCV.

```
Fitting 5 folds for each of 27 candidates, totalling 135 fits

✅ Best Random Forest Parameters Found: {'max_depth': 30, 'min_samples_split': 2, 'n_estimators': 300}

✅ Optimized Random Forest Model Performance:
              precision    recall  f1-score   support

           0       0.91      0.97      0.94       607
           1       0.46      0.23      0.31        73

    accuracy                           0.89       680
   macro avg       0.69      0.60      0.62       680
weighted avg       0.86      0.89      0.87       680

ROC-AUC Score: 0.7437092369840447
```

Model Improvement



- Accuracy Improved to 89%:The model correctly classifies 89% of all cases. Slight improvement from 88% before tuning.
- Better Precision for Default Cases (41% to 46%):Now, when the model predicts "Default," it is correct 46% of the time (compared to 41% before).Fewer false alarms (false positives reduced from 57 to 20).
- Recall for Default Cases (22% to 23%) Remains Low: Still misses 77% of actual defaulters. 56 defaulters are still misclassified as non-defaulters (false negatives).

Model Improvement

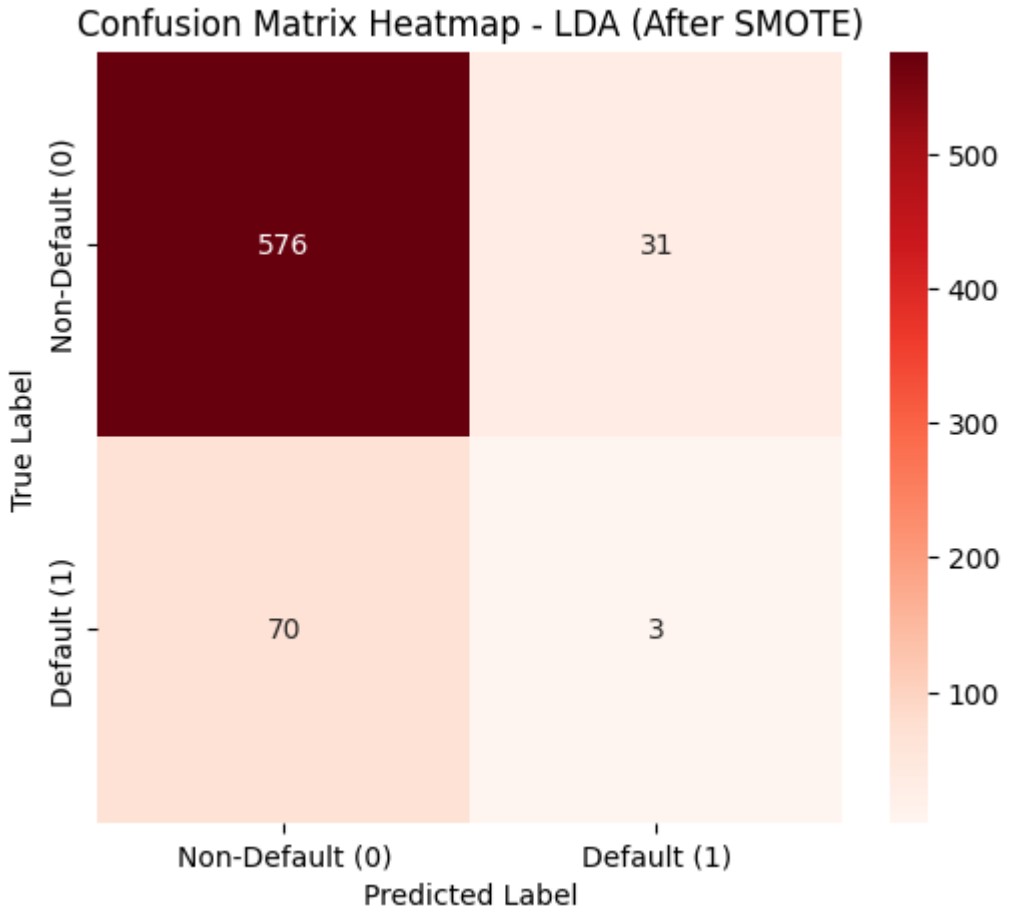
- **Linear Discriminant Analysis Model Improvement Using SMOTE**

✅ LDA Model Performance (After SMOTE):

	precision	recall	f1-score	support
0	0.89	0.95	0.92	607
1	0.09	0.04	0.06	73
accuracy			0.85	680
macro avg	0.49	0.50	0.49	680
weighted avg	0.81	0.85	0.83	680

ROC-AUC Score: 0.5677822662544288

Model Improvement



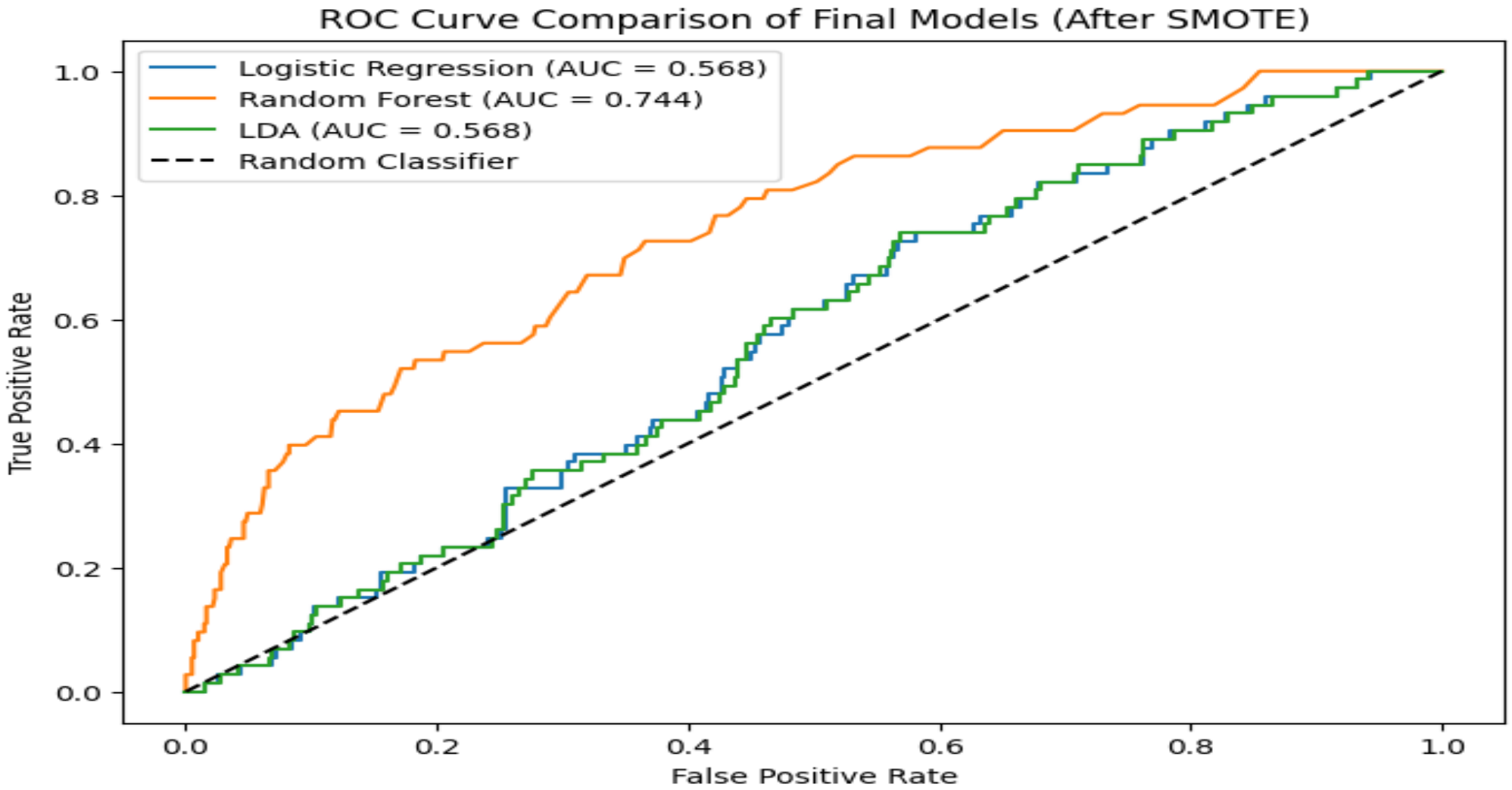
LDA Model Performance After SMOTE:

- LDA is Failing to Detect Defaulters (Low Recall: 4%):The model almost never predicts "Default" correctly. Only 3 out of 73 actual defaulters were correctly classified. This is a major issue because financial risk models must detect defaulters.
- High Precision for Non-Defaulters, But Useless for Defaulters: The model correctly classifies "Non-Defaults" with 95% recall.

Confusion Matrix Shows the Issue:

- 607 non-defaulters correctly identified (True Negatives).
- Only 3 defaulters correctly identified (True Positives).
- 70 defaulters were misclassified as non-defaulters (False Negatives).

Models Comparison After Improvement



Models Comparison After Improvement

Explanation of the ROC Curve Comparison: ROC Curve compares the performance of Logistic Regression, Random Forest, and LDA after SMOTE (Synthetic Minority Over-sampling Technique) was applied to handle class imbalance.

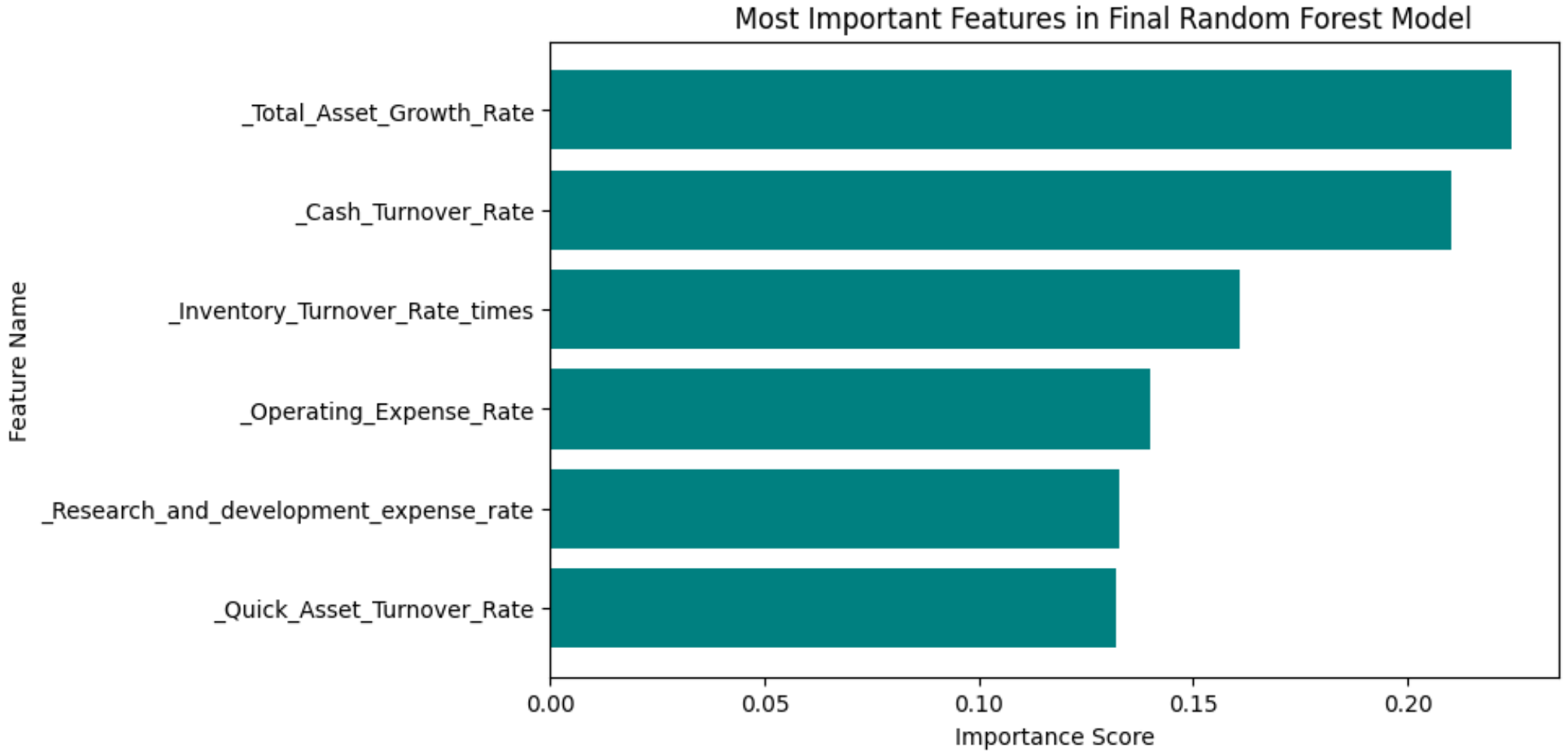
Random Forest is the Best Model:

- AUC Score of 0.744 means 74.4% chance of correctly distinguishing defaulters from non-defaulters.
- It significantly outperforms Logistic Regression and LDA.
- This makes Random Forest the preferred model.
- Logistic Regression & LDA Perform Poorly
- Both have AUC = 0.568, which is barely better than a random guess (AUC = 0.5).
- This indicates that Logistic Regression and LDA are ineffective at distinguishing defaulters.
- These models fail because they assume linear decision boundaries, while financial data is often non-linear.

The ROC Curve Confirms

- Random Forest (Orange Line) stays well above the diagonal (better classification).
- Logistic Regression & LDA (Green & Blue Lines) stay close to the diagonal, meaning they struggle to differentiate classes.

Feature Importance Analysis



Feature Importance Analysis

Feature Importance in Final Random Forest Model

- Total Asset Growth Rate is the Most Important Factor
- Companies with negative or declining asset growth are at higher risk of default.
- Suggests that rapid asset depreciation or low reinvestment in assets signals financial instability.
- Cash Turnover Rate is Critical for Liquidity
- Poor cash flow management limits a company’s ability to meet short-term obligations.
- Liquidity risk is a key indicator of financial distress.
- Inventory Turnover and Operating Expense Ratio Matter
- A low inventory turnover suggests that a company is struggling to sell goods, reducing revenue.
- High operating expenses reduce net profit, increasing the chance of financial difficulty.
- R&D Investment and Quick Assets Play a Role
- Heavy spending on R&D without strong financial backing could strain resources.
- Quick asset ratio helps evaluate liquidity but is less impactful compared to other features.

Conclusions and Recommendations

- Focus on Asset Growth & Cash Flow Stability Monitor firms with extremely low or high asset growth rates. Ensure cash turnover is maintained at healthy levels to prevent liquidity crises.
- Improve Inventory & Expense Management Track firms with low inventory turnover, as they may have unsold stock leading to financial inefficiency. Reduce unnecessary operating expenses to improve profit margins.
- Encourage Balanced R&D Investment Companies investing too little in R&D risk stagnation, while excessive R&D without profitability is risky. Encourage firms to align R&D with revenue growth.
- Use These Features for Credit Risk Scoring Models These factors should be weighted in credit risk models to improve default prediction accuracy. Regularly monitor these indicators to detect early financial distress.
- Declining Quick Asset Turnover Indicates Inefficiency If quick assets (cash, receivables) are not generating revenue efficiently, liquidity stress increases. Companies may struggle to pay short-term obligations.