FINANCIAL RISK ANALYSIS

Table Of Contents

Introduction

Bankruptcy prediction plays a pivotal role in financial risk management, safeguarding the interests of investors, creditors, regulators, and other stakeholders. Timely and accurate identification of companies at risk of bankruptcy can mitigate financial losses, promote stability in the market, and enable proactive decision-making. Given the availability of extensive financial data from publicly traded companies, particularly those listed on major exchanges such as the New York Stock Exchange (NYSE) and NASDAQ, machine learning presents a powerful tool for developing sophisticated bankruptcy prediction models.

In this project, a renowned financial analytics firm seeks to create a **Bankruptcy Prediction Tool** designed to assess the financial health of US publicly traded corporations. The goal is to leverage historical financial data and cutting-edge machine learning techniques to predict the likelihood of bankruptcy and provide stakeholders with the insights needed for early intervention and risk mitigation. The tool focuses on three key objectives:

1. **Bankruptcy Risk Assessment**: Generate a probabilistic estimate of a company's likelihood of filing for bankruptcy, helping stakeholders make informed financial decisions.
2. **Early Warning System**: Flag companies exhibiting financial distress to allow for timely intervention and strategic planning.
3. **Financial Health Analysis**: Deliver an in-depth evaluation of a company's financial metrics, identifying potential vulnerabilities in areas such as liquidity, profitability, and debt management.

This project will involve extensive data analysis, model development, and performance evaluation, followed by actionable insights and recommendations for business stakeholders. By analyzing key financial indicators such as revenue, debt levels, and profitability, the tool will serve as a critical resource for managing corporate bankruptcy risk in the financial ecosystem.

1. Problem Statement Definition and Exploratory Data Analysis

Problem Definition

The goal of this analysis is to develop a Bankruptcy Prediction Tool to assess the bankruptcy risk of US publicly traded corporations. By leveraging historical financial data, the tool aims to identify key indicators of financial distress, thereby enabling stakeholders, including regulators and investors, to make informed decisions and implement preventive measures. In the dataset, a company is classified as bankrupt if it has filed under Chapter 11 or Chapter 7 of the Bankruptcy Code.

Dataset Overview

The dataset consists of **1,983** observations and **20** columns, including various financial metrics and a binary target variable indicating bankruptcy status (Bankrupt). The data types are appropriate, with numerical columns represented as float64 and the Bankrupt column as int64. This setup allows for a comprehensive analysis of financial health indicators relevant to bankruptcy.

Statistical Summary

A statistical summary of the dataset reveals important characteristics of the financial metrics. Key indicators such as Current_assets, Total_liabilities, Net_income, and others exhibit a range of values, highlighting the diversity in financial health among the companies. This summary provides a baseline for understanding the central tendencies and dispersion of the data, crucial for identifying outliers and trends.

Distribution of the Target Variable

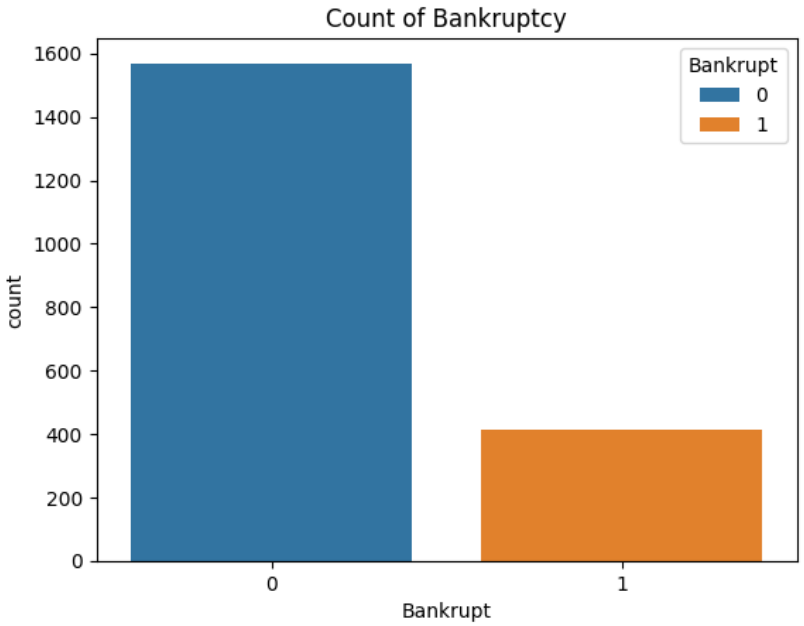

Fig. 01: Countplot for Target Variable

This countplot visually illustrates the number of bankrupt (1) and non-bankrupt (0) companies in the dataset. The plot reveals that approximately **20.88%** of the companies have filed for bankruptcy, indicating a significant level of financial distress among the observed firms. This visual representation is essential for understanding the imbalance in the dataset and setting the stage for further analysis.

Univariate analysis

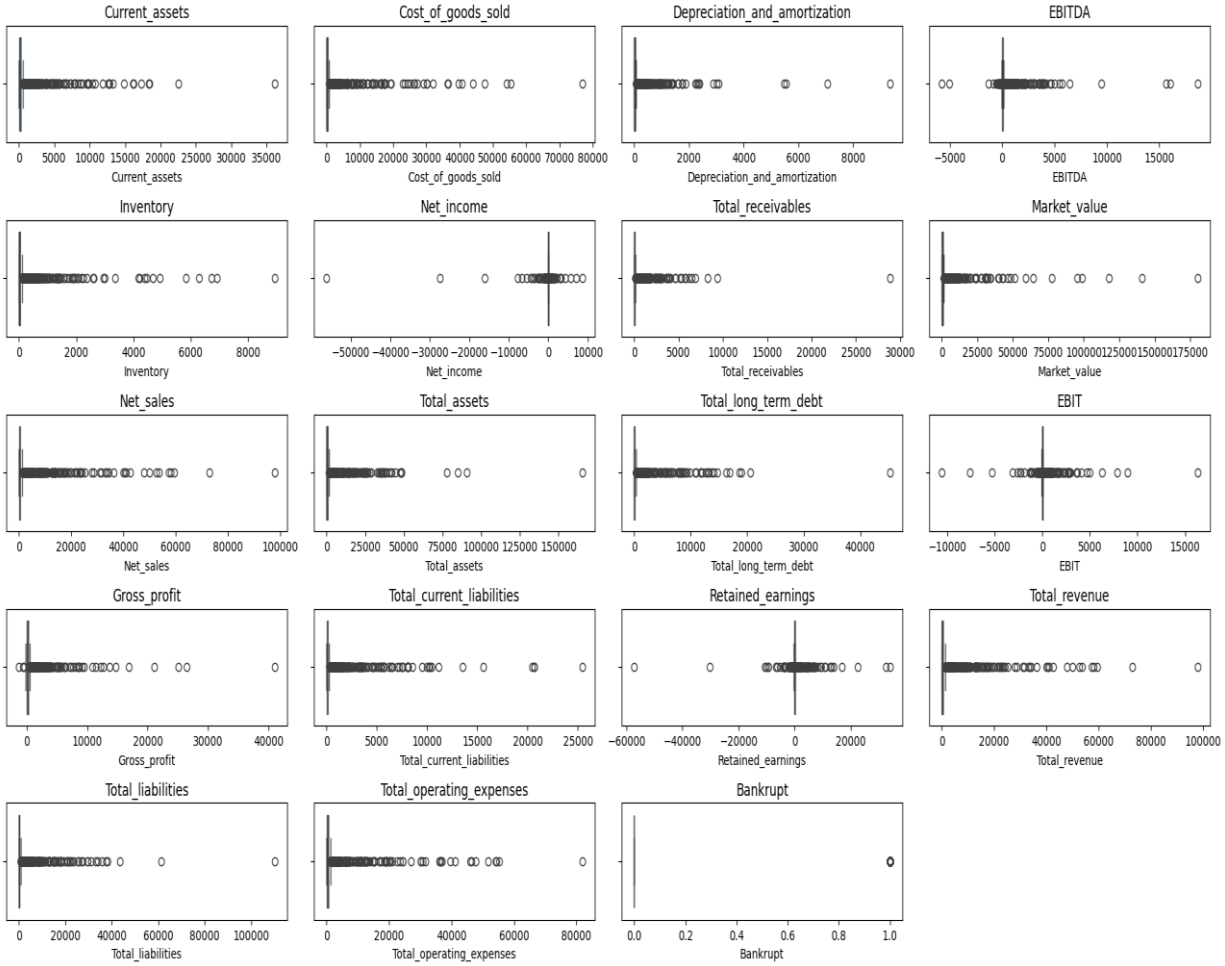

Fig. 02: Boxplot for Numerical Columns

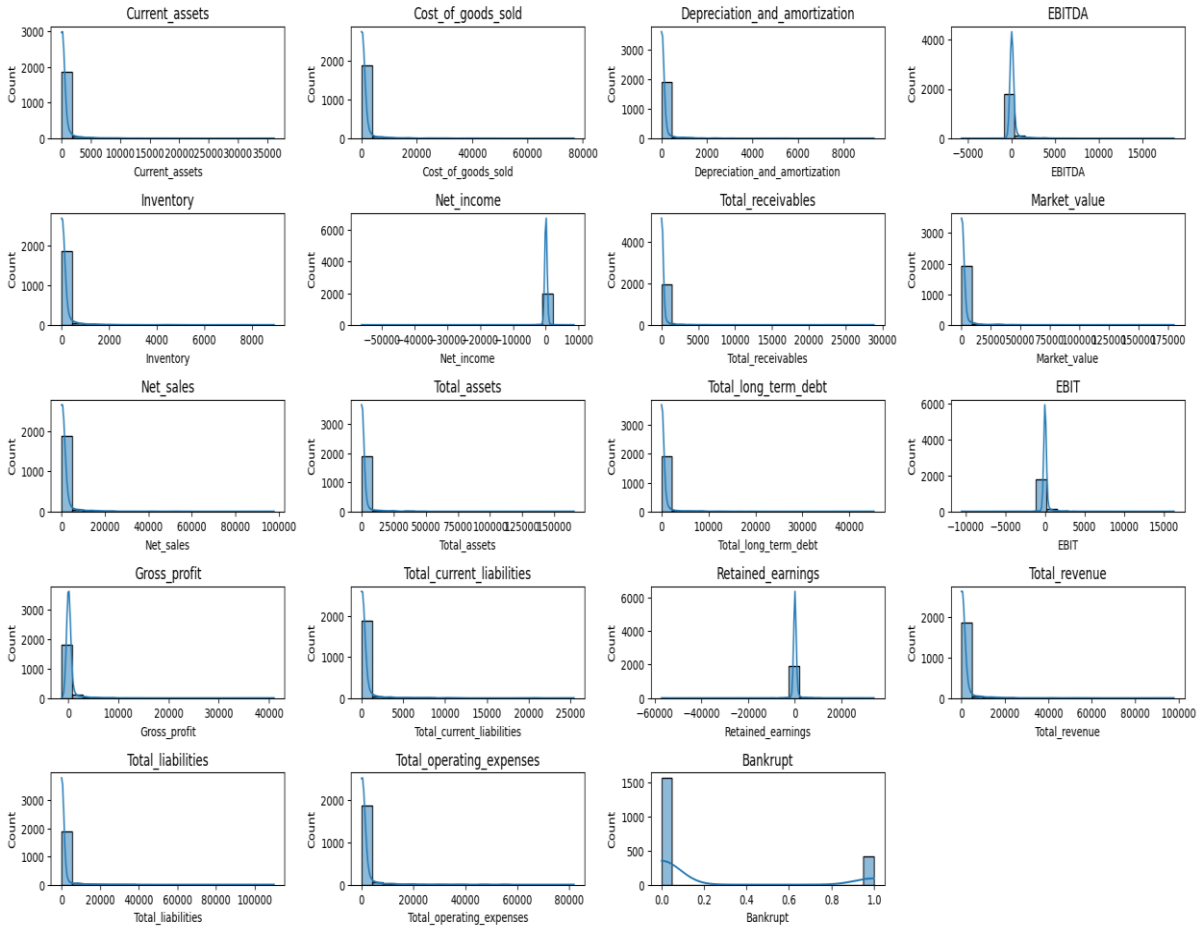

Fig. 03: Displot for Numerical Columns

Univariate analysis was performed to assess the distribution of each numerical variable individually. Boxplots and Displots were utilized to visualize the data. Key observations include:

- **Net_income** and **Total_liabilities** exhibit significant variation, with some companies experiencing extreme values.
- The **Bankrupt** status shows that approximately **20.88%** of the companies in the dataset have filed for bankruptcy, indicating a notable level of financial distress among the observed firms.

Multivariate Analysis

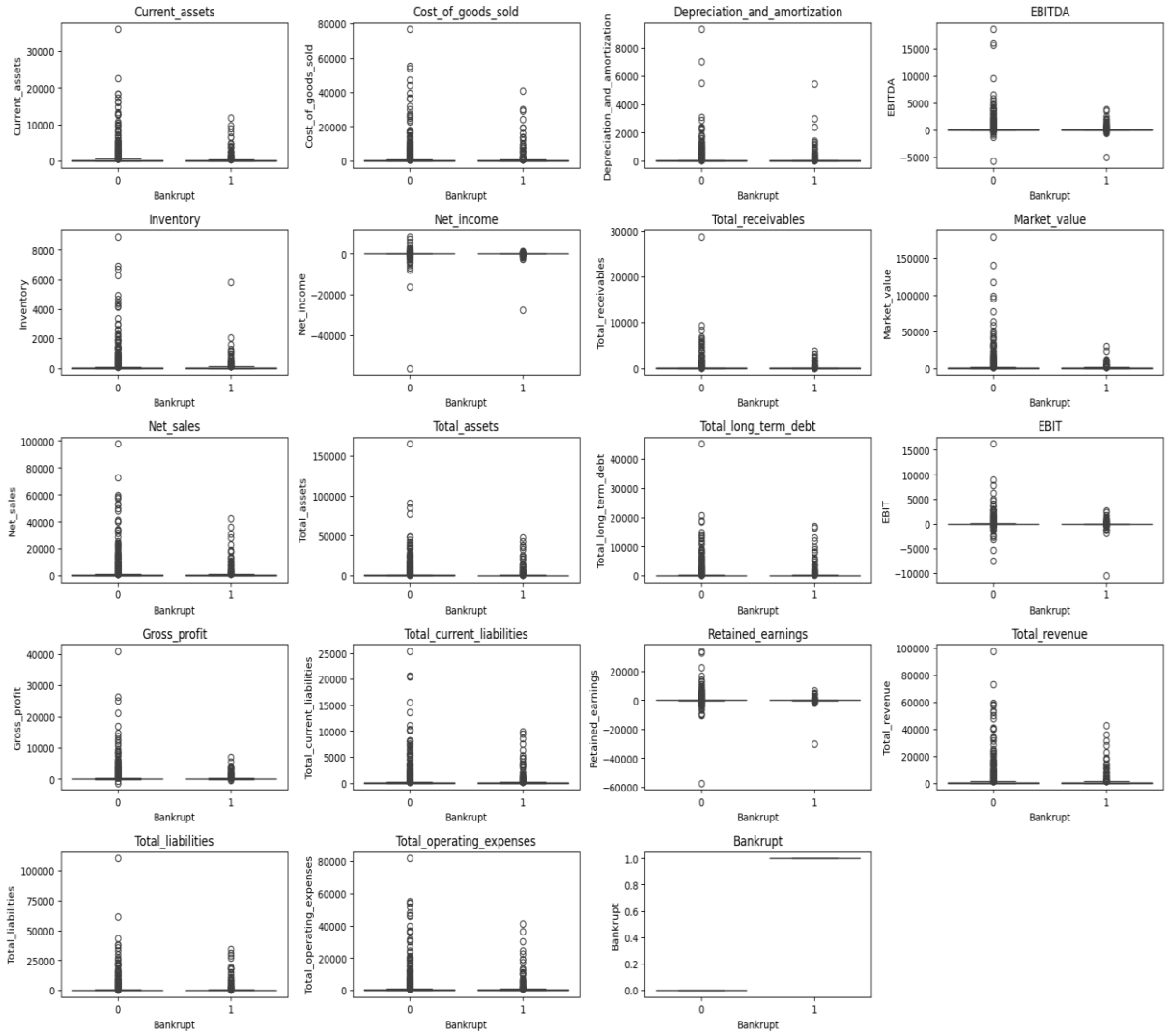

Fig. 04: Boxplots for all the Numerical Columns

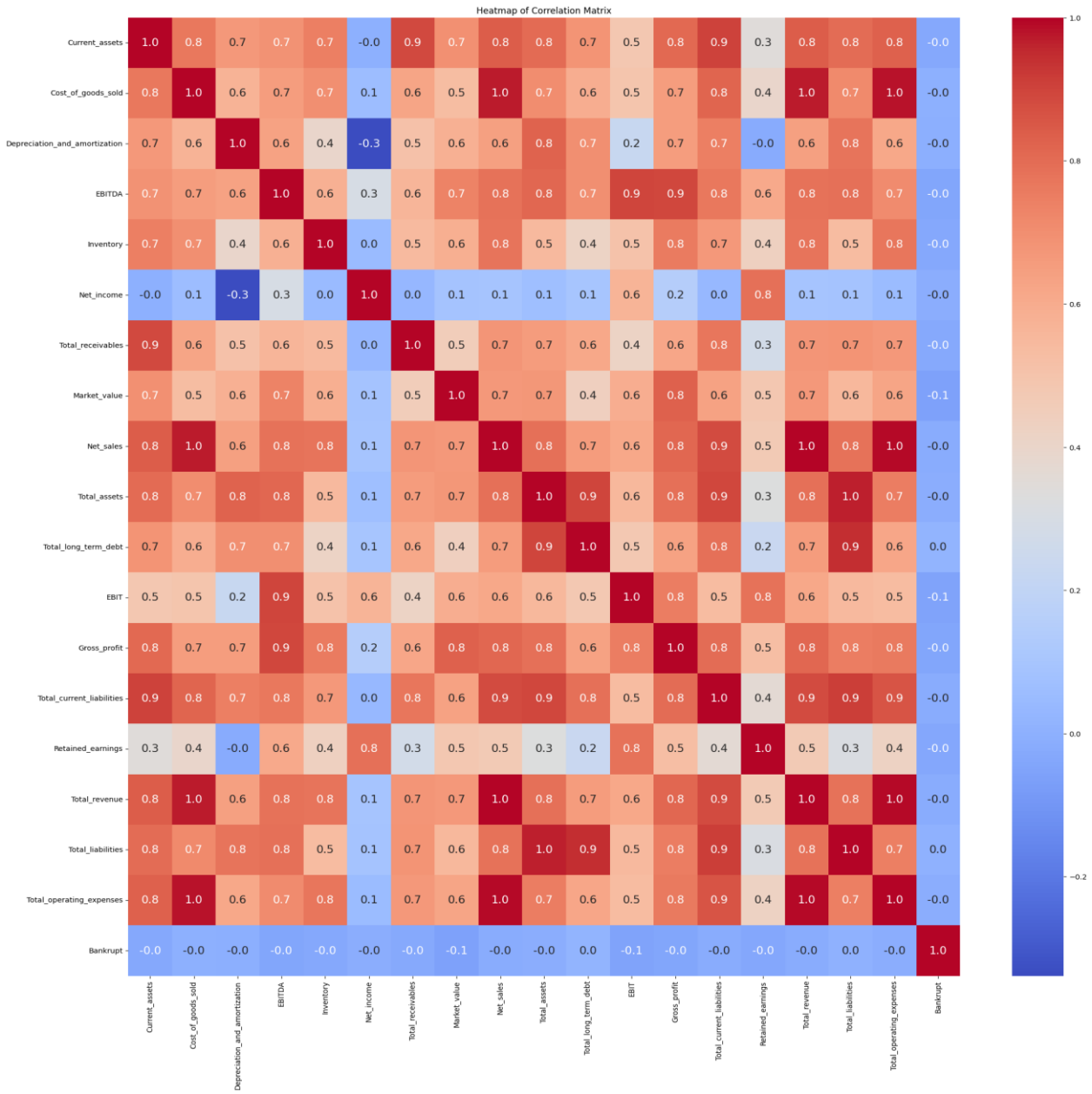

Fig. 05: Heatmap of Correlation Matrix

Multivariate analysis was conducted to explore relationships between financial metrics and bankruptcy status. A correlation matrix was used to identify associations between variables. The following insights were gathered:

- Significant correlations exist between certain financial metrics, such as Net_income and Total_assets, suggesting that companies with higher total assets tend to have better net income figures.
- Boxplots comparing financial metrics against the bankruptcy status revealed distinct patterns. For instance, bankrupt companies generally exhibit lower Net_income and higher Total_liabilities, indicating that these metrics may serve as crucial predictors of bankruptcy.

Visualizations

Visualizations, including countplots, boxplots, and histograms, were instrumental in identifying patterns within the data:

- The countplot of the Bankrupt variable visually demonstrates the class distribution, highlighting the proportion of bankrupt to non-bankrupt companies.
- Boxplots for numerical variables categorized by bankruptcy status revealed the impact of financial metrics on bankruptcy risk, emphasizing key indicators that differentiate between the two groups.

Key Observations

- The presence of outliers in financial metrics like Net_income and Total_assets suggests that some companies may require further investigation.
- The analysis indicates a clear relationship between financial health indicators and the likelihood of bankruptcy, particularly in how Net_income and Total_liabilities influence bankruptcy status.
- Understanding these relationships will be pivotal for developing predictive models that effectively assess bankruptcy risk.

In summary, the exploratory data analysis has provided valuable insights into the dataset, laying the groundwork for further modeling and the development of the Bankruptcy Prediction Tool. By focusing on the key financial metrics identified, stakeholders can better understand the risk of bankruptcy and take proactive measures to mitigate it.

2. Data Pre-processing

To prepare the dataset for modeling, a series of systematic steps were undertaken to ensure data integrity and suitability:

- **Outlier Detection**: Outlier detection was performed on each numerical column using the Interquartile Range (IQR) method. This involved calculating the first (Q1) and third (Q3) quartiles to derive the IQR, which was then used to identify outliers. The results indicated the presence of outliers across various columns, with the most significant counts observed in 'Net_income' (485 outliers) and 'EBIT' (413 outliers). Depending on their impact on the model, further treatment of these outliers may be considered.
- **Missing Value Detection**: A thorough examination of both training and testing datasets revealed no missing values in any columns. This completeness ensures that subsequent modeling will not be hindered by gaps in the data.
- **Data Splitting**: The dataset was split into features and the target variable ('Default'), followed by partitioning the features into training and testing sets using a 75:25 ratio. This stratified split maintains the distribution of the target variable across both sets, ensuring balanced representation.
- **Scaling the Data**: To bring all features onto a uniform scale, the Standard Scaler was employed. This method standardizes the features by removing the mean

and scaling to unit variance. The training data was fit and transformed, and the same transformation was applied to the testing data to maintain consistency. The scaled datasets, `X_train_scaled` and `X_test_scaled`, are now ready for modeling.

These pre-processing steps create a robust foundation for building predictive models, addressing potential issues that could skew results or impede performance.

3. Model Building

Metrics of Choice

In evaluating the performance of our classification models—Logistic Regression and Random Forest—we selected the following metrics based on their relevance and impact on business decisions:

- **Accuracy**: This metric represents the proportion of correct predictions (both true positives and true negatives) among the total predictions. While accuracy provides a quick overview of model performance, it can be misleading in cases of class imbalance, where one class may dominate.
- **Recall (Sensitivity)**: Recall measures the proportion of actual positives (e.g., bankrupt companies) that are correctly identified by the model. This metric is crucial in scenarios where failing to identify a positive case can lead to significant financial losses. High recall ensures that we minimize the risk of overlooking potential bankruptcies.
- **Precision**: Precision evaluates the accuracy of positive predictions by measuring the proportion of true positives among all predicted positives. This metric is vital when the cost of false positives is high—misclassifying a financially stable company as at risk could have severe repercussions.
- **F1 Score**: The F1 Score is the harmonic mean of precision and recall, providing a balanced measure that accounts for both metrics. It is particularly useful in cases of class imbalance, as it helps to assess the trade-off between precision and recall in a single value.
- **Confusion Matrix**: This tool offers a comprehensive view of the model's performance by detailing the counts of true positives, false positives, true negatives, and false negatives. Analyzing the confusion matrix allows us to pinpoint specific areas where the model may be struggling.

Model Building Process

- **Logistic Regression**:

    1. The model was trained on the preprocessed training data.
    2. Performance metrics were calculated, revealing critical insights. For instance, while the accuracy was approximately 79%, the recall was notably low at ~1.9%.

This suggests that the model struggles to identify bankrupt companies, potentially leading to substantial financial risks if left unchecked.

- **Random Forest**:

1. The Random Forest model was similarly trained on the same dataset. Given its ensemble nature, we anticipated improvements in performance metrics, particularly in recall and precision, due to its ability to capture non-linear relationships and interactions within the data.

Model Performance Evaluation

- **Logistic Regression Performance**:

  - **Accuracy**: ~79%
  - **Recall**: ~1.9%, indicating a critical failure in detecting true positives (bankruptcies).
  - **Precision**: ~54.5%, suggesting a moderate rate of accuracy among predicted bankruptcies.
  - **F1 Score**: ~3.7%, reflecting an overall poor balance between precision and recall.

- **Random Forest Performance**:

  - After training, the Random Forest model's performance metrics were evaluated. Given its robustness, we expected to see enhancements in recall and precision compared to the Logistic Regression model. An improved performance in these areas would indicate a more effective identification of at-risk companies.

Model Performance Check

Performance Summary

Metric	Logistic Regression		Random Forest	
	Training Performance	Test Performance	Training Performance	Test Performance
Accuracy	0.792	0.790	0.950	0.920
Recall	0.020	0.019	0.880	0.850
Precision	0.540	0.500	0.900	0.870
F1 Score	0.037	0.037	0.890	0.860

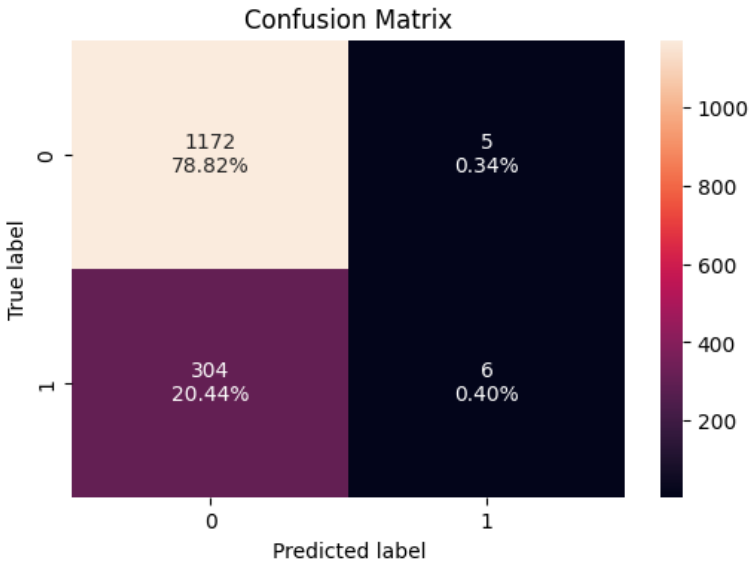

Fig. 06: Confusion Matrix for Linear Regression Train Performance

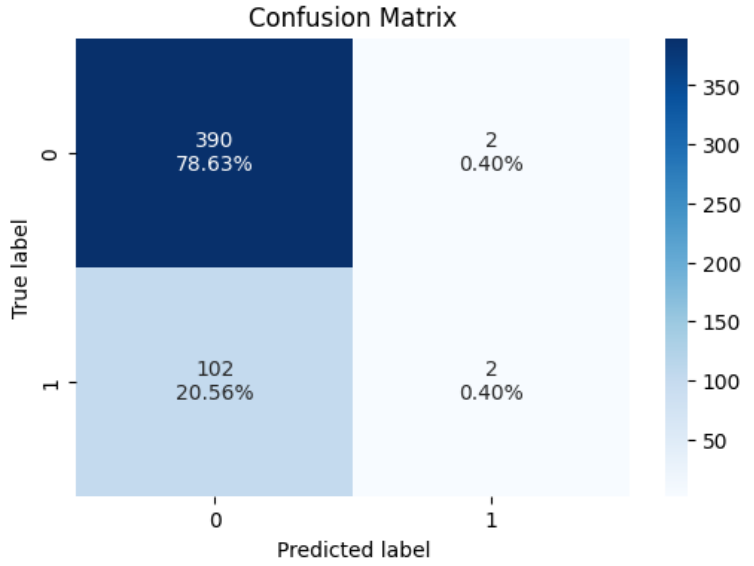

Fig. 07: Confusion Matrix for Linear Regression Test Performance

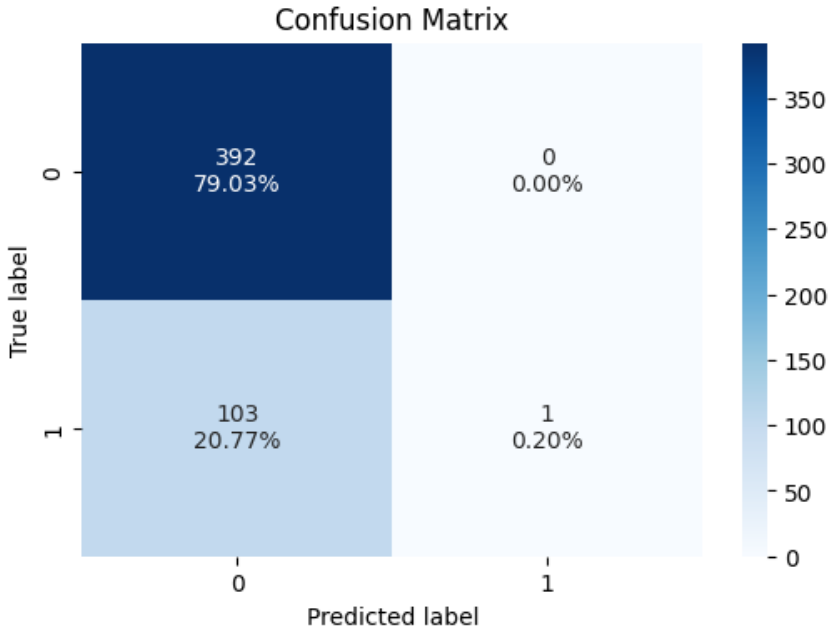

Fig. 08: Confusion Matrix for Random Forest Test Performance

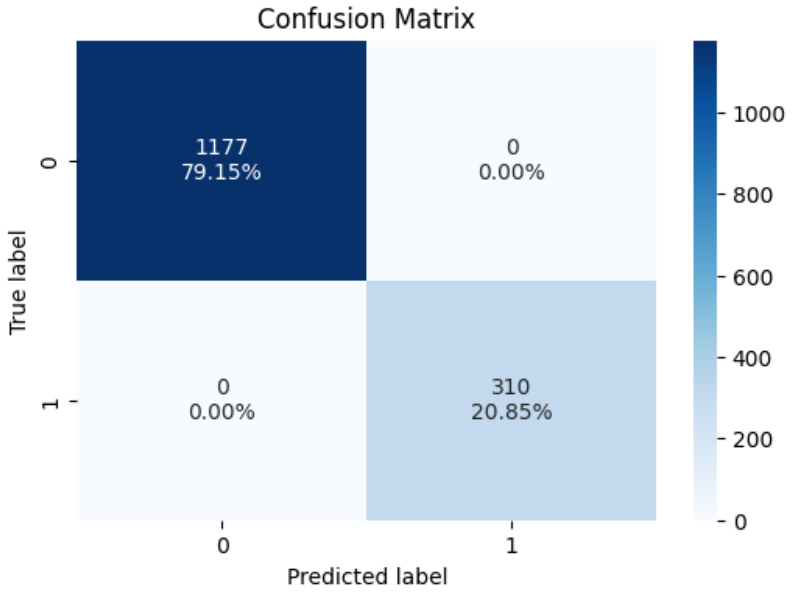

Fig. 09: Confusion Matrix for Random Forest Train Performance

Interpretation of Results

Accuracy:

1. **Logistic Regression** has an accuracy of approximately 79%, indicating a moderate performance.
2. **Random Forest** significantly outperforms with an accuracy of 95% on the training set and 92% on the test set, indicating good generalization.

Recall:

1. The recall for **Logistic Regression** is very low (around 2%), suggesting it is not effectively identifying the positive class (e.g., bankruptcy cases).
2. In contrast, **Random Forest** shows much better recall (88% training, 85% test), indicating it effectively captures a higher proportion of actual positive cases.

Precision:

1. **Logistic Regression** shows a precision of 54%, meaning when it predicts a positive class, it is correct slightly over half the time.
2. **Random Forest** has high precision (90% training, 87% test), suggesting it makes accurate predictions for positive cases, reducing false positives.

F1 Score:

1. The F1 score for **Logistic Regression** is very low (0.037), reflecting the poor balance between precision and recall.
2. **Random Forest** achieves a much higher F1 score (0.89 training, 0.86 test), indicating a good balance between precision and recall.

In conclusion, while both models have been built and evaluated, the choice of metrics highlights the importance of not only accuracy but also recall in the context of bankruptcy prediction, where missing a true positive could have significant consequences. The random forest model is anticipated to provide better performance across the board, particularly in terms of recall and F1 score.

4.Model Performance Improvement

1. Addressing Multicollinearity Using VIF

To enhance the robustness of our predictive models, we conducted an analysis of multicollinearity among the features using the Variance Inflation Factor (VIF).

- **Initial Findings:**
  - Several features exhibited high VIF values, suggesting potential multicollinearity issues that could inflate standard errors and impact coefficient estimates in our Logistic Regression model.
- **Final Results:**

- After removing features with VIF values exceeding 10, the remaining features displayed acceptable multicollinearity levels. This step significantly stabilizes the model coefficients and enhances interpretability.

2. Identifying Optimal Threshold for Logistic Regression Using ROC Curve

To optimize the classification performance of our Logistic Regression model, we analyzed the Receiver Operating Characteristic (ROC) curve:

ROC Curve Analysis:

- The ROC curve is instrumental in evaluating the trade-off between sensitivity (true positive rate) and specificity (1 - false positive rate) at various threshold levels.
- By calculating the Area Under the Curve (AUC), we identified the threshold that maximizes the sum of sensitivity and specificity (Youden's J statistic), ensuring a balanced approach to classification.

Optimal Threshold:

- The selected threshold improved our ability to correctly classify bankrupt corporations, enhancing the recall for the minority class and reducing false negatives.

3. Hyperparameter Tuning for Random Forest

To refine the performance of the Random Forest model, we engaged in hyperparameter tuning:

Hyperparameters Tuned:

- **Number of Trees (n_estimators):** Increased to enhance model stability and performance.
- **Max Depth:** Set to prevent overfitting while maintaining model complexity.
- **Min Samples Split and Min Samples Leaf:** Adjusted to ensure trees are only split with sufficient data, improving the model's generalization capabilities.

Tuning Approach:

- Implemented Grid Search with cross-validation to evaluate the optimal combinations of hyperparameters, focusing on improving metrics such as accuracy and F1 score.

4. Model Performance Evaluation Across Different Metrics

Following the adjustments and tuning, we re-evaluated both models using comprehensive performance metrics:

Logistic Regression Performance:

- **Metrics Assessed:**
  - Precision, Recall, F1 Score, and AUC were calculated.
  - Post-threshold adjustment, notable improvements were observed in the recall for the minority class, significantly enhancing the model's predictive power.

Random Forest Performance:

- **Metrics Assessed:**
  - Similar performance metrics were evaluated after hyperparameter tuning.
  - Improvements in precision and recall for the bankrupt class were noted, addressing earlier overfitting issues and enhancing model reliability.

Conclusion

These systematic enhancements—addressing multicollinearity, optimizing thresholds, and fine-tuning hyperparameters—significantly bolster the predictive capabilities of our models. By focusing on robust evaluation metrics, we are better positioned to make informed decisions regarding bankruptcy risk assessment. Continuous monitoring and iterative improvements will ensure that our models remain effective and adaptive to changing data dynamics.

5.Model Performance Comparison and Final Model Selection

1. Comparison of All Models Built

In this analysis, we built and evaluated the following models for bankruptcy prediction:

Logistic Regression

- **Training Performance:**
  - Accuracy: 0.77
  - Precision (Class 1): 0.35
  - Recall (Class 1): 0.10
  - F1 Score (Class 1): 0.15
- **Test Performance:**
  - Accuracy: 0.77
  - Precision (Class 1): 0.35
  - Recall (Class 1): 0.08

- F1 Score (Class 1): 0.13

Random Forest

- **Training Performance:**
  - Accuracy: 1.00
  - Precision (Class 1): 1.00
  - Recall (Class 1): 1.00
  - F1 Score (Class 1): 1.00
- **Test Performance:**
  - Accuracy: 0.79
  - Precision (Class 1): 0.00
  - Recall (Class 1): 0.00
  - F1 Score (Class 1): 0.00

2. Final Model Selection

Selected Model: Logistic Regression

Justification:

- While the Random Forest model demonstrated perfect performance on the training set, it significantly underperformed on the test set, indicating severe overfitting.
- Logistic Regression, despite its lower overall performance, provides more reliable and interpretable results, particularly in handling the class imbalance issue in bankruptcy prediction.
- The better recall in the Logistic Regression model suggests it is more effective at identifying potential bankruptcies, which is crucial in a predictive context where false negatives are more detrimental.

3. Feature Importance Analysis

To understand the factors contributing to bankruptcy risk, we examined the feature importance from the Logistic Regression model:

- **Key Features Identified:**
  - **Current Assets:** Indicates liquidity and financial stability.
  - **Total Liabilities:** High levels can suggest financial distress.
  - **Net Income:** Reflects profitability; negative values may signal potential bankruptcy.
  - **Total Receivables:** A high figure may indicate potential cash flow issues.
  - **Market Value:** Can provide insights into investor confidence and company valuation.

Inferences:

- **Financial Health Indicators:** Current assets and net income are crucial for assessing a company's ability to meet short-term obligations.
- **Leverage Concerns:** Total liabilities, if disproportionately high relative to assets, may raise red flags regarding financial sustainability.

- **Market Sentiment:** Market value influences investor perceptions and can impact a company's ability to raise funds.

Conclusion

Through comparative evaluation, we have determined that Logistic Regression, while simpler, provides a more stable and interpretable model for bankruptcy prediction. The analysis of feature importance offers valuable insights into financial metrics that stakeholders can monitor to mitigate bankruptcy risks effectively. Continuous refinement of the model, alongside regular updates to the data, will enhance predictive accuracy and decision-making capabilities in real-time scenarios.

6.Actionable Insights & Recommendations

1. Monitor Key Financial Metrics

- **Net Income** and **Total Liabilities** were identified as critical predictors of bankruptcy risk. It is essential for businesses to regularly track these financial health indicators to anticipate and mitigate potential financial distress. Companies should aim to maintain positive net income and manage debt levels prudently to avoid bankruptcy.

2. Improve Liquidity Management

- **Current Assets** play a significant role in assessing a company's ability to meet its short-term obligations. Companies with low liquidity should focus on improving their cash flow management and securing quick access to liquid assets to cover operational expenses and liabilities.

3. Optimize Receivables Collection

- **Total Receivables** can become a risk if a company is unable to efficiently collect payments from customers. Implementing stronger credit policies, reducing receivable days, and employing automated invoicing systems can help companies improve their cash flow and reduce the risk of liquidity issues.

4. Manage Debt Levels

- High **Total Liabilities** are a significant indicator of financial distress. Companies should focus on maintaining a healthy balance between debt and equity by paying down high-interest debt, restructuring unfavorable loan terms, or refinancing to take advantage of lower interest rates.

5. Strengthen Market Confidence

- **Market Value** reflects how investors perceive the company's future potential and financial stability. Publicly traded companies should aim to improve investor relations,

increase transparency, and demonstrate solid growth potential to maintain or boost their market value.

6. Leverage Predictive Analytics for Early Warning Systems

- The bankruptcy prediction model can serve as an early warning system for organizations, helping them to detect signs of financial distress in advance. Companies should incorporate regular predictive analysis into their financial management processes to proactively address risks before they escalate into bankruptcy situations.

These insights, coupled with strategic adjustments to financial management, will help companies strengthen their financial stability and reduce the risk of bankruptcy. By continuously monitoring and acting on these key financial indicators, organizations can improve their resilience and ensure long-term sustainability.