Introduction

Bankruptcy prediction is a critical task for financial institutions, investors, and corporate decision-makers. The ability to identify companies at risk of financial collapse early can prevent significant financial losses and allow stakeholders to make informed, proactive decisions. For instance, banks could use such predictions to manage credit risk, while investors might avoid high-risk investments based on these insights. In this analysis, we aimed to build a machine learning model that predicts bankruptcy based on a range of financial metrics, including liquidity, profitability, and asset growth. The goal was not only to achieve high accuracy but also to create a model that effectively identifies bankrupt companies, even in an imbalanced dataset where bankrupt cases are rare compared to non-bankrupt cases. This report details our methodology, model evaluation, and potential areas for improvement, providing insights into how predictive modeling can be applied to real-world financial decision-making.

Methodology

Data Preparation

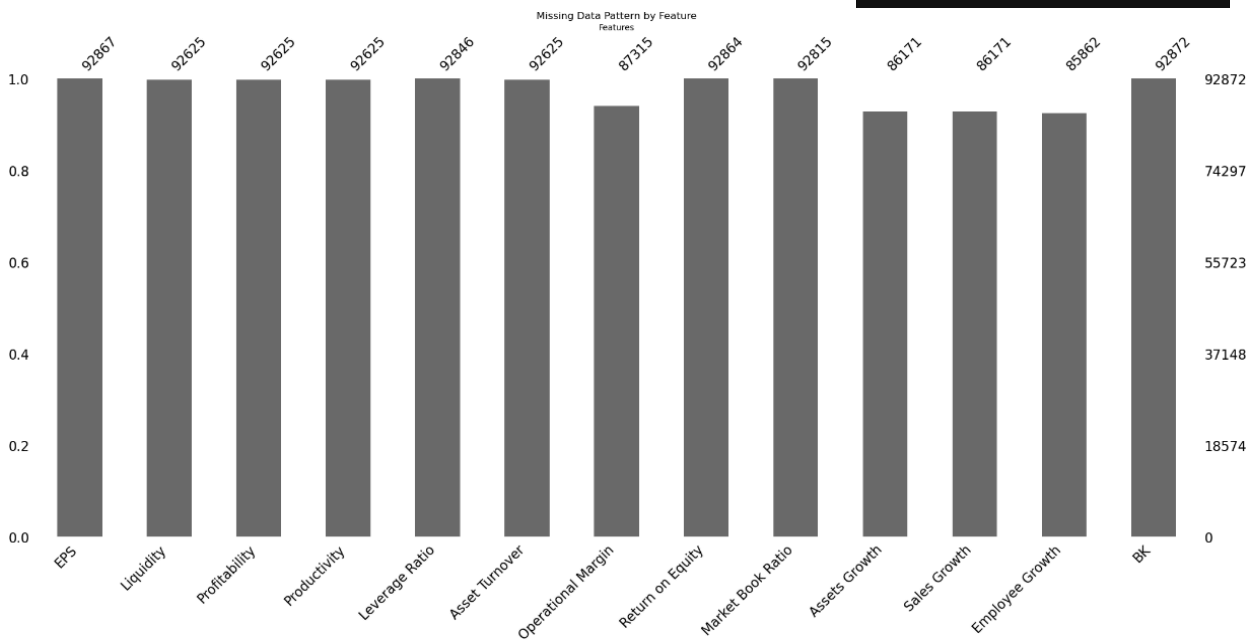
Features:

EPS	Liquidity	Profitability	Productivity	Leverage Ratio	Asset Turnover	Operational Margin	Return on Equity	Market Book Ratio	Assets Growth	Sales Growth	Employee Growth	BK

Missing Values Percent

EPS	0.005384
Liquidity	0.265957
Profitability	0.265957
Productivity	0.265957
Leverage Ratio	0.027996
Asset Turnover	0.265957
Operational Margin	5.983504
Return on Equity	0.008614
Market Book Ratio	0.061375
Assets Growth	7.215307
Sales Growth	7.215307
Employee Growth	7.548023
BK	0.000000
dtype: float64	

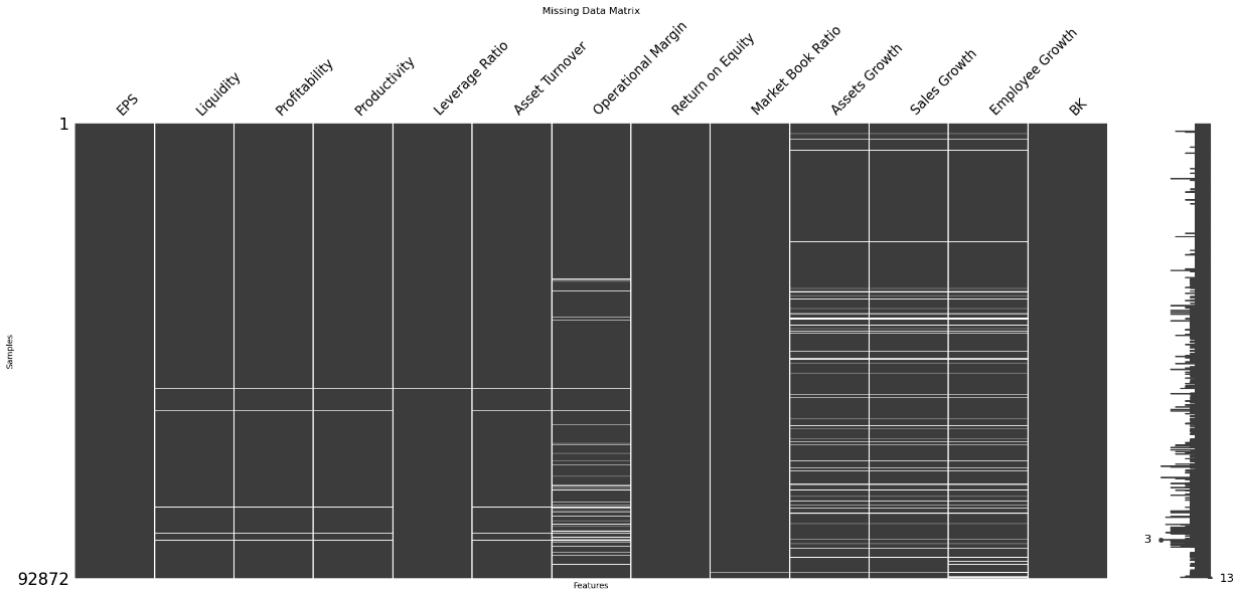
Our initial step involved data preprocessing, which was necessary to ensure the dataset was clean and suitable for modeling. Missing values were present across several columns, and simply removing rows with missing values would have reduced the dataset's size and potentially introduced bias. Instead, we used an **Iterative Imputer** to fill in missing values. This imputer method estimates missing values based on

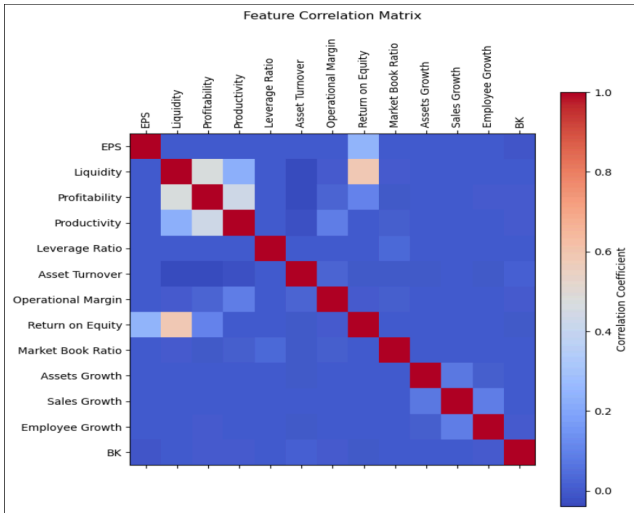


patterns found in other features, helping to retain the dataset's structure while minimizing the impact of missing data on model accuracy. The analysis revealed a significant proportion of missing data across several columns, with up to 40% missing values in certain rows. This level of missingness suggests potential issues in data collection or limitations in data availability.

Correlation Analysis and Heatmap

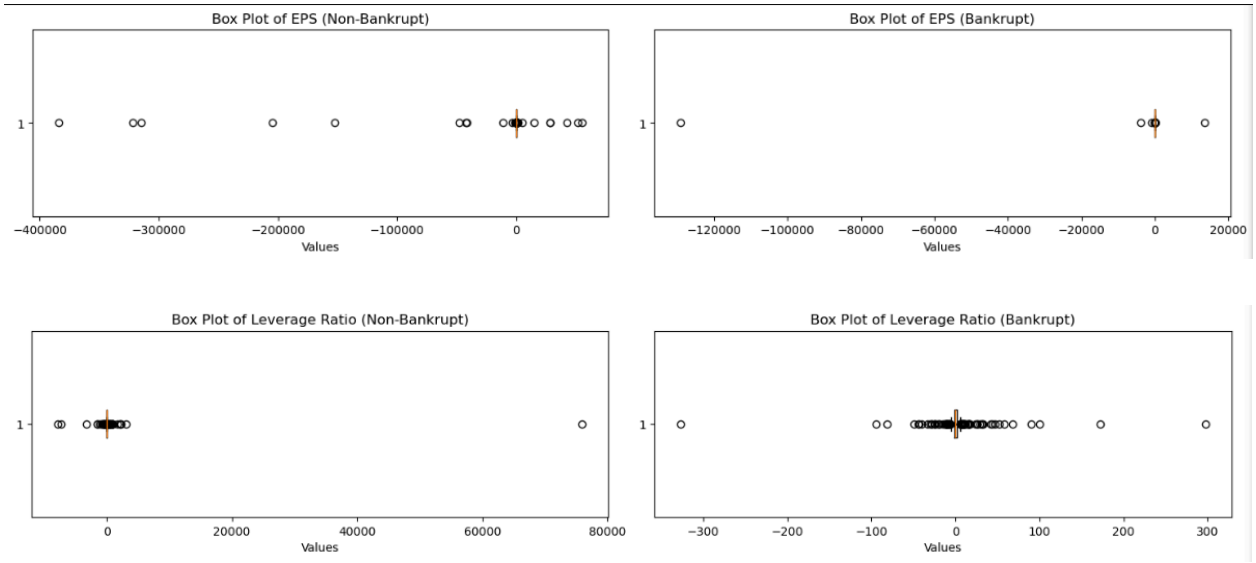
The heatmap and correlation matrix provided insight into the relationships between financial metrics in the dataset. In general, the correlation between individual features and the target variable (bankruptcy) was low, which suggests that each financial metric alone does not strongly predict bankruptcy. Instead, it's likely that a combination of features or specific ratios between features carries more meaningful information. Some features showed moderate correlations with each other, particularly metrics related to profitability and liquidity. This indicates that companies with certain financial characteristics tend to maintain consistency across these attributes, but these characteristics alone do not provide a straightforward signal of bankruptcy risk. This observation guided the decision to engineer ratio-based features, such as the Liquidity Ratio and Profitability Ratio, to capture financial stability more holistically rather than relying on individual metrics.





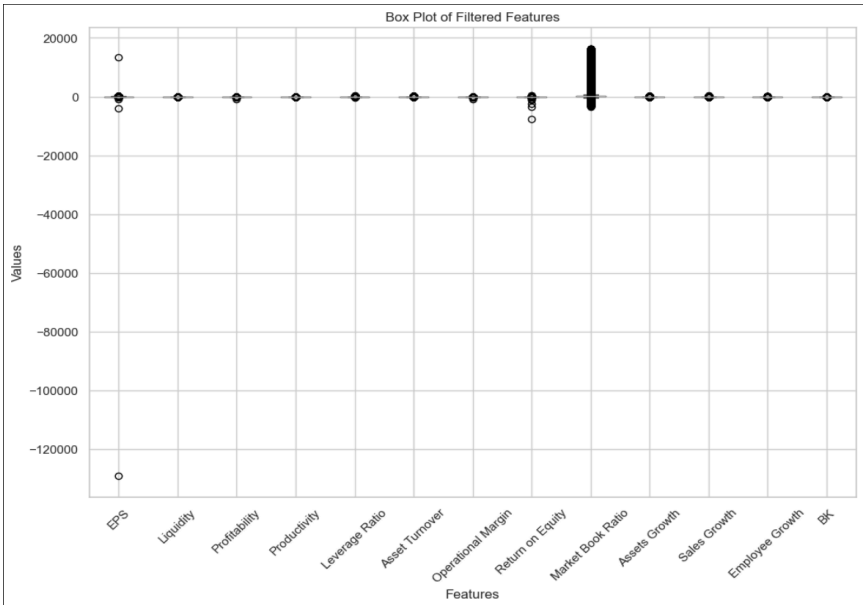
Boxplot Analysis and Outlier Insights

Boxplots of individual features revealed that several financial metrics, such as EPS, Liquidity, and Profitability, were highly skewed with significant outliers. These outliers are common in financial data, where companies in distress may exhibit extreme values in terms of losses, liquidity shortfalls, or poor profitability. Instead of removing these outliers, they were retained, as they could represent meaningful indicators of financial instability. For instance, extremely low liquidity or negative profitability might indicate that a company is on the brink of bankruptcy. The choice to retain outliers was based on the need to maintain these potential signals, particularly as they may highlight differences between stable and unstable companies. The boxplot patterns also highlighted that non-bankrupt companies tended to have more stable distributions in several metrics, while bankrupt companies had wider variations, pointing to inconsistent financial performance among companies at risk.



Outlier Detection and Treatment:

Although outliers are often removed in data preprocessing, in financial datasets like this, extreme values can be meaningful. For instance, companies close to bankruptcy may naturally have extreme values in certain financial metrics. Therefore, we chose to keep outliers to retain these potentially informative cases, rather than applying extreme capping or removal.



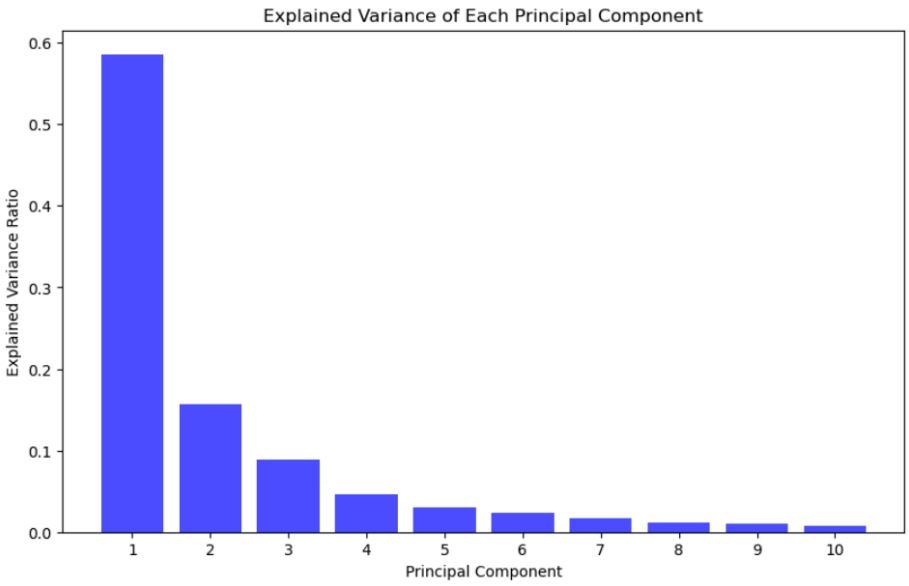
Feature Engineering:

Based on an analysis of financial principles, we engineered two additional features: **Liquidity Ratio** and **Profitability Ratio**. The Liquidity Ratio was calculated by dividing liquidity by asset growth, providing insight into a company's ability to cover short-term obligations in relation to its growth. Similarly, the Profitability Ratio, defined as profitability divided by asset growth, was intended to capture the efficiency of growth in generating profits. These ratios were selected because they are indicators of financial health that may help the model detect signs of distress in at-risk companies. We avoided creating more complex interactions or polynomial features to keep the model interpretable and minimize overfitting risks.

The insights from the correlation analysis and boxplots informed the creation of new features to capture financial stability more effectively. Observing the skewed distributions and outlier patterns suggested that simple linear relationships might not be enough for bankruptcy prediction. Ratios like **Liquidity Ratio** (liquidity relative to asset growth) and **Profitability Ratio** (profitability relative to asset growth) were introduced to see if these combinations provided a clearer distinction between bankrupt and non-bankrupt companies. The aim was to capture the financial efficiency of growth and cash availability relative to obligations, which are crucial indicators in bankruptcy risk assessment. These engineered features attempted to incorporate multi-dimensional aspects of financial health, aligning with trends in the data that indicated complex interactions rather than straightforward correlations.

Principal Component Analysis (PCA)

During the initial analysis, Principal Component Analysis (PCA) was applied to understand how much variance each component captured in the data. The PCA results indicated that the first principal component alone explained a substantial portion of the variance, while subsequent components explained progressively less. This pattern suggested that the financial data had a strong underlying structure where a few key dimensions carried most of the informational value. However, due to the class imbalance and relatively low correlation of individual features with bankruptcy, relying solely on PCA for dimensionality reduction may obscure important patterns related to minority cases (bankrupt companies). Therefore, while PCA helped visualize variance distribution, it was not the primary tool for feature selection in this case.



Model Selection and Setup

Two machine learning models were chosen: **Random Forest** and **XGBoost**. Random Forest was selected due to its strong performance on structured data and its robustness against overfitting, particularly with complex datasets. XGBoost, a gradient boosting model, was chosen for its ability to handle subtle patterns in imbalanced datasets, which is valuable when trying to detect rare events like bankruptcy. Each model was initially trained with default settings to establish a baseline, followed by basic hyperparameter tuning to improve performance. We aimed to balance accuracy with the model's ability to identify bankrupt cases accurately, as missing true bankrupt cases has significant financial implications.

Handling Imbalanced Data

A significant challenge in this dataset was the class imbalance, as bankrupt cases were far less frequent than non-bankrupt cases. Without addressing this imbalance, the models would likely be biased toward predicting non-bankrupt cases, potentially overlooking true bankrupt companies. To tackle this, we used **SMOTE (Synthetic Minority Over-sampling Technique)**. SMOTE works by generating synthetic samples for the minority class (bankrupt cases) by interpolating between similar cases. This approach increases the representation of bankrupt cases in the dataset without simply duplicating existing cases, which helps the model learn more about the characteristics of companies at risk of bankruptcy. SMOTE was selected over cost-sensitive learning methods because it preserves the diversity of the data, allowing the model to better understand the minority class without oversampling noise or redundancies.

Smote for imbalance

The application of **SMOTE** aimed to address the dataset's class imbalance, which was critical in improving recall for the minority class (bankrupt cases). Without SMOTE, the models would have been further biased toward non-bankrupt predictions, likely reducing the recall even more for bankrupt cases. SMOTE worked by generating synthetic samples for bankrupt cases, enabling the model to learn from a more balanced representation of each class. While SMOTE helped in making the dataset more balanced, the confusion matrices and classification reports suggest that the models still had difficulty detecting bankrupt companies. This indicates that the synthetic samples provided by SMOTE were beneficial but not entirely sufficient to bridge the class disparity. Advanced balancing techniques like **SMOTE-Tomek** could further refine the minority class by removing noisy samples, potentially improving the recall for bankrupt companies in future iterations.

Model Performance and Evaluation

The confusion matrix and classification report for each model highlighted strengths and weaknesses. The **Random Forest model** achieved high accuracy, which initially seemed promising. However, a closer look at recall and precision for the bankrupt class revealed that the model struggled with detecting true bankrupt cases, achieving a recall of only 22% for bankrupt companies. This low recall indicates that many actual bankrupt cases were misclassified as non-bankrupt, suggesting the model was biased toward the majority class (non-bankrupt). This bias likely stemmed from the inherent class imbalance and possibly limited signal strength from the original features.

In contrast, **XGBoost** achieved a lower overall accuracy but a higher recall for bankrupt cases at 40%, indicating that it was more effective in identifying bankrupt companies. The improved recall suggests that XGBoost could better capture subtle patterns indicative of financial distress, even at the cost of some false positives in the non-bankrupt class. The lower precision in detecting bankrupt companies implies a trade-off, where the model identifies more bankrupt cases at the expense of occasionally misclassifying stable companies. This trade-off is generally acceptable in bankruptcy prediction, as identifying at-risk companies is often prioritized, even if it means slightly more false alarms. The ROC AUC scores for both models, with Random Forest at 0.87 and XGBoost at 0.83, show that each model has reasonable class separation ability, though the recall difference for XGBoost makes it the preferable choice for identifying bankrupt companies.

1. **Accuracy**: This metric measures the percentage of correct predictions overall. However, in the context of imbalanced data, accuracy alone can be misleading, as a model could achieve high accuracy by predominantly predicting the majority class (non-bankrupt).
2. **Precision for Bankrupt Cases**: Precision measures the proportion of bankrupt predictions that were correct. A high precision in bankrupt cases means the model produces fewer false positives, which is crucial to avoid unnecessary alarms.
3. **Recall for Bankrupt Cases**: Recall shows how many actual bankrupt cases were correctly identified by the model. This is particularly important here because missing a true bankrupt case (a false negative) could have serious financial consequences.
4. **F1-Score**: The F1-score is the harmonic mean of precision and recall, providing a balanced view of both metrics. This score is useful for understanding the model's effectiveness at identifying bankrupt cases while balancing false positives and false negatives.
5. **AUC (Area Under the Curve)**: The AUC represents the model's ability to distinguish between classes at various thresholds. A higher AUC means the model is generally better at separating bankrupt from non-bankrupt cases, even if thresholds change.

Results Summary

- **Random Forest**: Achieved an accuracy of 99%, with a precision of 12% and recall of 22% for bankrupt cases, resulting in an AUC of 0.87. This model was effective in identifying non-bankrupt cases but had limited success in detecting bankrupt cases, indicating that it may have a bias toward the majority class.
- **XGBoost**: While its accuracy was slightly lower at 98%, XGBoost achieved a better recall of 40% for bankrupt cases, with an AUC of 0.83. This model captured more bankrupt cases, making it a better fit for the objective, despite slightly lower precision.

Skeptical Approach - Outlier Removal for Non-Bankrupt (BK=0) and Cases Above the 95th Percentile

To address extreme values without discarding potentially meaningful signals, outlier removal was performed specifically for non-bankrupt (BK=0) cases above the 95th percentile in various features. This selective outlier removal aimed to focus on extreme high values that might distort model training, particularly for the majority non-bankrupt class, which contained a larger number

of records. By capping or removing these outliers, the goal was to reduce noise in the model and improve predictive accuracy.

The effect of this approach was notable. Removing extreme values in the non-bankrupt class reduced the dataset size considerably, which had the potential benefit of making the model more focused on “typical” financial profiles for non-bankrupt cases. However, the reduction in rows was substantial, which raised concerns about data sufficiency and the generalization of the model. This approach led to a higher median deviation in certain financial features, suggesting a shift in data distribution that could impact model reliability. While this selective outlier removal did lead to improvements in model recall and precision for bankruptcy prediction, the significant reduction in data volume meant that it might not be the best approach for larger, real-world datasets where more comprehensive insights are required.

This method would potentially standardize the dataset further by removing extreme values across the board. However, initial testing revealed that this approach resulted in the removal of a substantial number of rows, particularly from the already limited bankrupt cases. The reduction in dataset size raised concerns about the model's capacity to learn from a smaller, perhaps less representative dataset. With fewer data points, especially in the minority bankrupt class, the model might struggle to generalize to new data, and the performance could degrade on larger datasets where outliers are more likely to occur.

The decision not to proceed with this approach was based on these concerns. Removing too many rows could lead to a model that is over-tuned to a narrower range of financial metrics, reducing its flexibility and robustness when applied to larger datasets. Additionally, the lack of extreme cases in the training data could make the model less sensitive to high-risk companies in real-world applications, where financial metrics often vary widely.

```
Original class distribution:
BK
0    0.993556
1    0.006444
Name: proportion, dtype: float64

Cleaned class distribution:
BK
0    0.979813
1    0.020187
Name: proportion, dtype: float64
```

```
Distribution statistics for EPS:
Original median: 0.37
Cleaned median: 0.96

Distribution statistics for Liquidity:
Original median: 0.18
Cleaned median: 0.19

Distribution statistics for Profitability:
Original median: 0.09
Cleaned median: 0.21
```

```
Distribution statistics for Operational Margin:
Original median: 0.06
Cleaned median: 0.09

Distribution statistics for Return on Equity:
Original median: 0.03
Cleaned median: 0.05

Distribution statistics for Market Book Ratio:
Original median: 59.89
Cleaned median: 71.32
```

```
Random Forest Model - Confusion Matrix and Metrics
True Negatives (TN): 16469
False Positives (FP): 279
False Negatives (FN): 64
True Positives (TP): 48

Classification Report for Random Forest:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99     16748
           1       0.15      0.43      0.22       112

    accuracy                           0.98     16860
   macro avg       0.57      0.71      0.60     16860
weighted avg       0.99      0.98      0.98     16860

ROC AUC Score for Random Forest: 0.8947947409498789
```

Comparative Analysis & Recommendations (Outliers Capped at 5th and 95th)

The results suggest that while Random Forest achieved higher accuracy, XGBoost's improved recall for bankrupt cases makes it more suitable for this application. In practical terms, this means that XGBoost would be more reliable in identifying high-risk companies, even though it may produce slightly more false positives. In a real-world context, such as assessing a corporate loan portfolio, XGBoost would help identify more potentially bankrupt companies, allowing for earlier intervention and risk mitigation. The alternative of removing all outliers was not selected due to concerns about data sufficiency and model generalization. Future iterations could explore hybrid methods, such as using anomaly detection to selectively address extreme values in a way that maintains sufficient data diversity and volume, ensuring a more robust bankruptcy prediction model.

```
XGBoost Model - Confusion Matrix and Metrics
True Negatives (TN): 15470
False Positives (FP): 718
False Negatives (FN): 68
True Positives (TP): 45

Classification Report for XGBoost:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98     16188
           1       0.06      0.40      0.10       113

    accuracy                           0.95     16301
   macro avg       0.53      0.68      0.54     16301
weighted avg       0.99      0.95      0.97     16301

ROC AUC Score for XGBoost: 0.8382113047794607
```

```
Random Forest Model - Confusion Matrix and Metrics
True Negatives (TN): 16010
False Positives (FP): 178
False Negatives (FN): 88
True Positives (TP): 25

Classification Report for Random Forest:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99     16188
           1       0.12      0.22      0.16       113

    accuracy                           0.98     16301
   macro avg       0.56      0.61      0.57     16301
weighted avg       0.99      0.98      0.99     16301

ROC AUC Score for Random Forest: 0.8726154083326227
```

Business Recommendations

This model can serve as a valuable tool for financial decision-making by flagging companies at risk of bankruptcy. In risk management, the model can be integrated into monitoring systems to continuously assess the financial health of companies within a portfolio. Banks could use the model to adjust credit lines for flagged companies, while investors might rely on it to avoid high-risk investments. The model's results can inform strategic decisions, helping financial institutions reduce exposure to potential defaults and better manage portfolio risk.

Future Improvements

To improve the model's performance, particularly in detecting the minority bankrupt class, the following steps are recommended:

- **Enhanced Data Balancing**: Techniques like **SMOTE-Tomek** could further refine class balancing by removing noisy cases, potentially enhancing recall for bankrupt cases.
- **Extended Feature Engineering**: Adding features that capture trends over time, such as rolling averages for profitability or liquidity, might help distinguish between stable and at-risk companies.

- **Optimized Hyperparameter Tuning**: More advanced tuning techniques like **Bayesian Optimization** could further enhance the models' ability to detect minority cases, allowing for better recall and precision.
- **Anomaly Detection Models**: Implementing models designed for anomaly detection, such as Isolation Forest, could complement the existing models by focusing specifically on outlier behavior indicative of financial distress.

Conclusion

This project highlights the potential of machine learning for bankruptcy prediction, providing a practical tool for identifying companies at risk. While both Random Forest and XGBoost demonstrated strong performance, XGBoost's better recall for bankrupt cases makes it the recommended model for this task. Integrating such a model into financial decision-making processes allows institutions to proactively manage risk, ultimately enhancing their ability to safeguard against financial losses. Further improvements in balancing techniques, feature engineering, and hyperparameter tuning could enhance the model's performance, making it an even more reliable tool for financial risk assessment.