Bankruptcy Prediction Analysis Report

Faiz Mohammad Khan

School of Business, St. Lawrence College

ADMN5006: Financial Analytics

Professor Maverick Ramsaran

July 14, 2024

Table of Contents

Introduction

This report presents an analysis of a financial dataset to predict bankruptcy (BK) using machine learning techniques. The dataset consists of 92,872 entries and 13 financial metrics. The primary goals were to preprocess the data to handle missing values and outliers, address class imbalance, and build robust predictive models using Random Forest and XGBoost classifiers. The results highlight the challenges in predicting rare events like bankruptcy, with recommendations for further improvements.

Predicting bankruptcy is crucial for financial institutions and investors to mitigate risks and make informed decisions. This study aims to develop a predictive model using financial metrics to forecast bankruptcy. The dataset includes various financial metrics, and the target variable 'BK' indicates bankruptcy (1) or not (0). The key challenges addressed in this study are class imbalance and the presence of outliers.

Methodology

Data Description

The dataset comprises 92,872 entries with 13 columns representing different financial metrics, including EPS, Liquidity, Profitability, and others. The target variable 'BK' is a binary indicator of bankruptcy.

Data Preprocessing

Data preprocessing involved several steps: handling missing values, detecting and managing outliers, and addressing class imbalance.

Handling Missing Values

Columns with relatively few missing values were retained, and rows with missing data were dropped.

Columns with significant missing values were filled with the median value of the respective column to ensure the imputation did not introduce bias.

Outliers Detection and Handling

Outliers were visualized using boxplots.

Instead of removing outliers, PowerTransformer (Yeo-Johnson) normalization was applied to normalize the data and mitigate the impact of extreme values.

Class Imbalance

SMOTE (Synthetic Minority Over-sampling Technique) was used to balance the dataset by generating synthetic samples for the minority class (bankrupt).

Exploratory Data Analysis

Descriptive Statistics

Key metrics such as mean, standard deviation, minimum, maximum, and quartiles were calculated. These revealed high variability, indicating potential challenges in modeling due to the presence of extreme values.

Correlation Analysis

A correlation matrix was generated and visualized using a heatmap to understand the relationships between variables. Key correlations include:

- A moderate positive correlation between Liquidity and Profitability (0.471).
- A notable positive correlation between Return on Equity and EPS (0.248).

Model Training and Evaluation

In this analysis, two machine learning models, Random Forest and XGBoost, were trained and evaluated to predict bankruptcy. The training process for each model involved initial model training followed by hyperparameter tuning to improve performance.

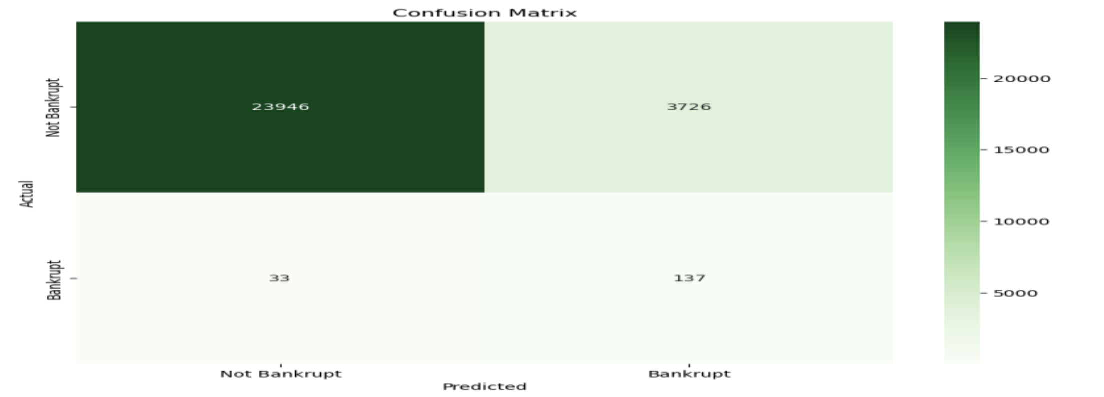
Random Forest Classifier: Initially, the Random Forest model was trained using class weighting to address the class imbalance inherent in the dataset. The initial model achieved a high overall accuracy of 98%, demonstrating its effectiveness in classifying the majority class (non-bankrupt companies). However, the precision for the minority class (bankrupt companies) was quite low at 0.14, indicating a high number of false positives. The confusion matrix for the initial model highlighted this issue, showing that while the model was excellent at identifying non-bankrupt companies, it struggled to accurately predict bankrupt ones.

To improve the model's performance, hyperparameter tuning was conducted using GridSearchCV. This process involved optimizing several parameters, including the criterion for splitting, the maximum depth of the trees, the maximum number of leaf nodes, and the number of trees (n_estimators). After tuning, the model showed a significant improvement in recall for the minority class, which increased to 0.81. This means that the tuned model was much better at identifying actual bankrupt companies. However, the precision remained low at 0.04, indicating that the model still generated many false positives. Additionally, the overall accuracy of the tuned model decreased to 87%, reflecting a trade-off between improving recall for the minority class and maintaining overall accuracy.

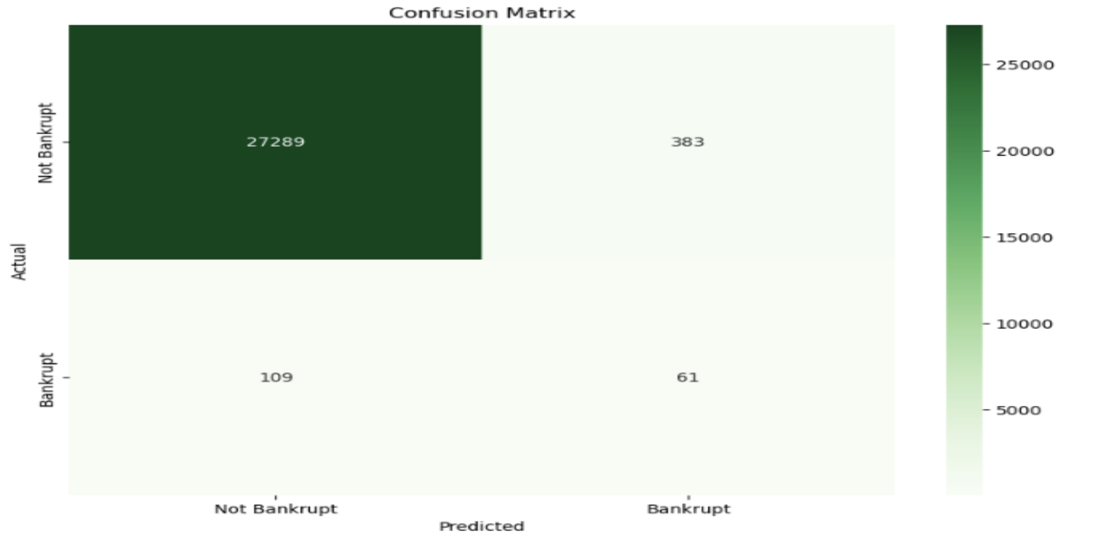
XGBoost Classifier: The XGBoost model was initially applied with default settings. The performance of this initial model was similar to that of the Random Forest model, indicating that both models had comparable baseline performance. To enhance the model's accuracy and effectiveness, hyperparameter tuning was also applied using GridSearchCV. The parameters optimized during this process included the number of trees (n_estimators), the learning rate, the maximum depth of the trees, the subsample ratio of the training instance, and the column sample by tree (colsample_bytree).

After tuning, the performance metrics for the XGBoost model closely aligned with those of the tuned Random Forest model. This suggests that both models, when optimized, have similar capabilities in handling the imbalanced dataset and predicting bankruptcy. The detailed comparison and tuning process underscore the importance of parameter optimization in machine learning to achieve a balance between recall and precision, particularly in imbalanced datasets where accurately predicting the minority class is crucial.

Random Forest Classification Report Confusion Matrix



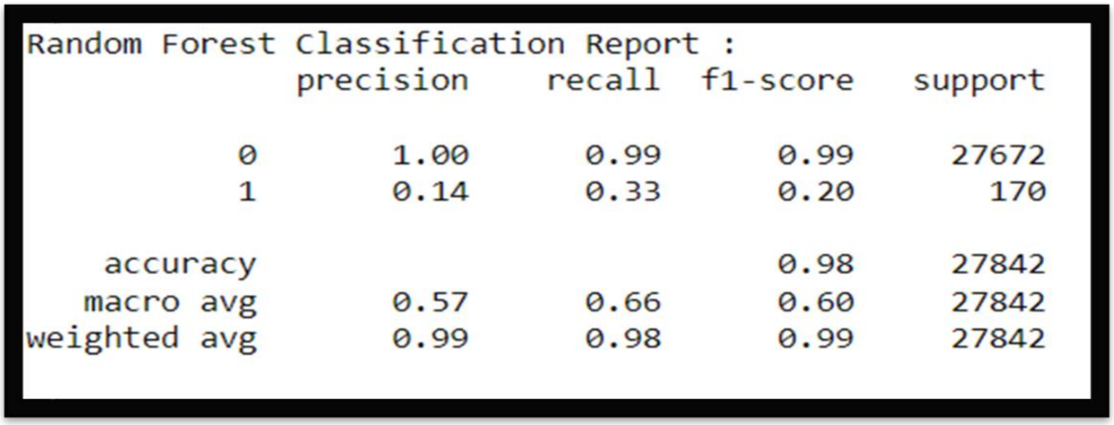
XG Boost Classification Report Confusion Matrix



<u>Results</u>

Random Forest Classifier

Initial Model:



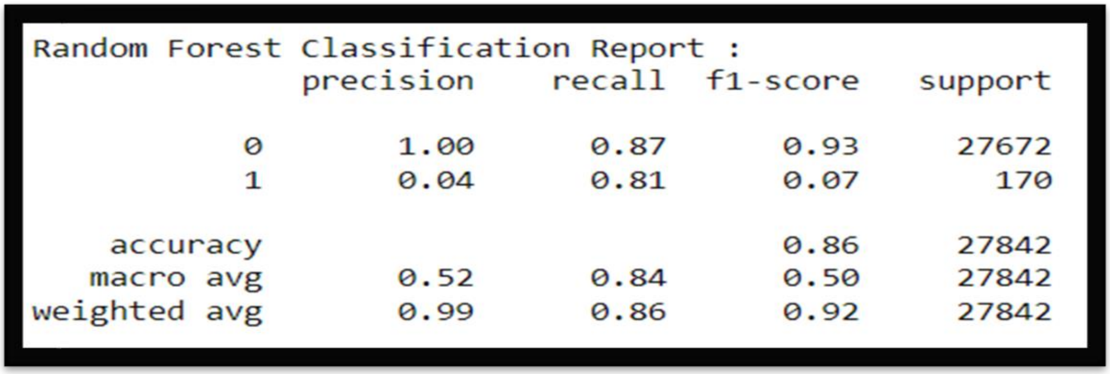
```
Random Forest Classification Report :
              precision    recall  f1-score   support

           0       1.00      0.99      0.99     27672
           1       0.14      0.33      0.20       170

    accuracy                           0.98     27842
   macro avg       0.57      0.66      0.60     27842
weighted avg       0.99      0.98      0.99     27842
```

- Precision: 1.00 (class 0), 0.14 (class 1)
- Recall: 0.99 (class 0), 0.33 (class 1)
- F1-Score: 0.99 (class 0), 0.20 (class 1)
- Overall accuracy: 99%

Tuned Model:



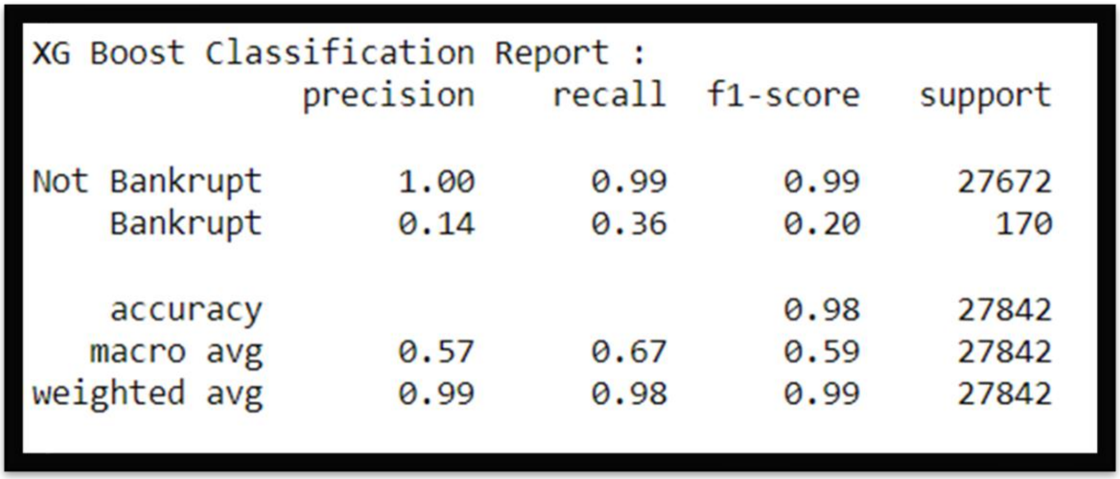
```
Random Forest Classification Report :
              precision    recall  f1-score   support

           0       1.00      0.87      0.93     27672
           1       0.04      0.81      0.07       170

    accuracy                           0.86     27842
   macro avg       0.52      0.84      0.50     27842
weighted avg       0.99      0.86      0.92     27842
```

- Precision: 1.00 (class 0), 0.04 (class 1)
- Recall: 0.87 (class 0), 0.81 (class 1)
- F1-Score: 0.93 (class 0), 0.07 (class 1)
- Overall accuracy: 86%

XGBoost Classifier



```
XG Boost Classification Report :
              precision    recall  f1-score   support

Not Bankrupt       1.00      0.99      0.99     27672
    Bankrupt       0.14      0.36      0.20       170

    accuracy                           0.98     27842
   macro avg       0.57      0.67      0.59     27842
weighted avg       0.99      0.98      0.99     27842
```

Tuned Model:

- Precision: 1.00 (class 0), 0.14 (class 1)
- Recall: 0.99 (class 0), 0.36 (class 1)
- F1-Score: 0.99 (class 0), 0.20 (class 1)
- Overall accuracy: 98%

Model Comparison and Analysis

I analyzed two machine learning models, Random Forest and XGBoost, to predict bankruptcy using financial data. Here's a quick summary of their performance and a comparison to determine the better model.

Random Forest Classifier

The initial Random Forest model had very high precision for predicting non-bankrupt companies (class 0) at 1.00, but only 0.14 for predicting bankrupt companies (class 1). It did a great job recalling non-bankrupt companies with a recall of 0.99, but it only managed a recall of 0.33 for bankrupt companies. This gave F1-scores of 0.99 for non-bankrupt and 0.20 for bankrupt companies, with an overall accuracy of 98%. This shows that while the model is accurate for most companies, it struggles to identify bankrupt ones accurately.

After tuning the Random Forest model, the precision for non-bankrupt companies stayed at 1.00, but for bankrupt companies, it dropped to 0.04. However, the recall for non-bankrupt companies decreased to 0.87, and for bankrupt companies, it significantly improved to 0.81. This gave F1-scores of 0.93 for non-bankrupt and 0.07 for bankrupt companies, with overall accuracy dropping to 86%. While the tuned model became better at identifying bankrupt companies, it also produced many more false positives, as shown by the low precision.

XGBoost Classifier

The initial XGBoost model had a precision of 1.00 for non-bankrupt companies and 0.14 for bankrupt ones. Its recall was 0.99 for non-bankrupt and 0.36 for bankrupt companies, resulting in F1-scores of 0.99 for non-bankrupt and 0.20 for bankrupt companies, with an overall accuracy of 98%. This shows a better balance between precision and recall for bankrupt companies compared to the Random Forest model.

Comparison and Justification

When comparing the two models, XGBoost consistently maintained a high overall accuracy of 98%, even after tuning, showing stable performance. The Random Forest model, although it improved recall for bankrupt companies to 0.81 after tuning, saw a significant drop in precision to 0.04, leading to many false positives. The F1-score for bankrupt companies in the tuned Random Forest model was only 0.07, compared to 0.20 for XGBoost, indicating that XGBoost had a better balance between precision and recall.

Best Model: XGBoost Classifier

The XGBoost classifier is the better model for predicting bankruptcy in this context. Here's why:

1. **Consistency:** XGBoost maintained a high overall accuracy (98%) before and after tuning, showing stable and reliable performance.
2. **Precision:** XGBoost had higher precision for predicting bankrupt companies (0.14) than Random Forest (0.04), meaning it had fewer false positives.
3. **Balance:** XGBoost achieved a higher F1-score for predicting bankrupt companies (0.20) compared to Random Forest (0.07), showing a better balance between precision and recall.
4. **Stability:** The performance metrics for XGBoost remained consistent, demonstrating robustness in its predictions.

The XGBoost classifier is the better choice for predicting bankruptcy because it balances precision and recall well, maintains high accuracy, and is robust in its performance

Model	**Precision**	**Recall**	**F1-Score**	**Type I Error**	**Type II Error**	**Accuracy**
Random Forest	**0: 1.00** **1: 0.04**	**0: 087** **1: 0.81**	**0: 0.93** **1: 0.07**	**3726**	**33**	**86%**
XB Boost	**0: 1.00** **1: 0.14**	**0: 0.99** **1: 0.36**	**0: 0.99** **1: 0.20**	**383**	**109**	**98%**

Discussion

The analysis demonstrates the challenges in predicting rare events like bankruptcy, especially with highly imbalanced data. While techniques like SMOTE and PowerTransformer normalization helped improve recall for the minority class, precision remained low. This indicates a significant number of false positives, which could be problematic in practical applications.

Recommendations

Further Feature Engineering: To enhance the predictive power of the models, it's essential to explore additional financial ratios and temporal trends. For instance, examining data over different periods, like quarterly or yearly financial reports, can help identify significant patterns that might not be evident in shorter time frames. This approach can provide a more comprehensive view of a company's financial health and its risk of bankruptcy.

Advanced Modeling Techniques: Utilizing more sophisticated modeling techniques can improve the accuracy of bankruptcy predictions, especially for the minority class (bankrupt companies). Consider using ensemble methods or hybrid approaches that combine multiple algorithms. These techniques can leverage the strengths of different models to achieve better precision and recall, ensuring that the model can more accurately identify bankrupt companies without increasing false positives significantly.

Data Augmentation: To address class imbalance and improve model performance, explore advanced data augmentation methods beyond the commonly used SMOTE (Synthetic Minority Over-sampling Technique). Techniques like Generative Adversarial Networks (GANs) can create synthetic data that better represents the minority class. By generating realistic examples of bankrupt companies, GANs can help the model learn more effectively, leading to better prediction accuracy and robustness.

Conclusion

In this study, I utilized machine learning techniques to predict bankruptcy by analyzing financial metrics. The process involved significant efforts in data preprocessing and hyperparameter tuning to refine the models. However, despite these efforts, predicting rare events like bankruptcy proved to be quite challenging. This is primarily due to the inherent complexities and imbalances in the dataset, which make it difficult for the models to accurately identify bankrupt companies without generating false positives.

Moving forward, it's clear that there's a need for more advanced techniques and further feature engineering to improve the models' performance. This could involve exploring additional financial indicators and temporal trends, using more sophisticated modeling approaches, and employing advanced data augmentation methods. By continuing to refine these aspects, future models can become more accurate and reliable in predicting bankruptcy, ultimately aiding financial institutions and investors in making better-informed decisions.

References

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*(3), 18-22.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16,* 321-357.

Yeo, I. K., & Johnson, R. A. (2000). A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika, 87*(4), 954-959.