Company Bankruptcy Prediction

EE708 Course Project, Group 29

I. Objective

The objective of this study is to develop a predictive model capable of assessing the likelihood of corporate bankruptcy based on various financial and business-related attributes. By leveraging machine learning techniques, the model aims to enhance financial risk assessment and contribute to informed decision-making for economic stability.

II. Preprocessing

The outlier capping process applies the Interquartile Range (IQR) method to limit extreme values in numerical features. It calculates Q1 (25th percentile) and Q3 (75th percentile), determines the IQR ($Q_3 - Q_1$), and caps values below $Q_1 - 1.5 \cdot IQR$ and above $Q_3 + 1.5 \cdot IQR$ to these bounds revealing a class imbalance of 5301 non-bankrupt vs. 154 bankrupt cases.

The dataset is split into 80% training and 20% testing while preserving class distribution using stratification.

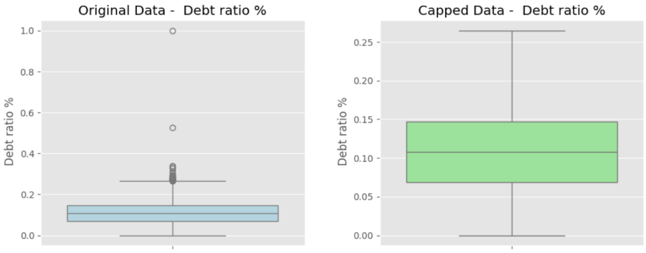


Fig. 1. Example of capping outliers on a feature

III. Refining Training Data

A. Eliminating Class Imbalance

The dataset exhibited a severe class imbalance (5301 non-bankrupt vs. 154 bankrupt cases), which could lead to biased predictions favoring the majority class. Traditional machine learning models tend to perform poorly on highly imbalanced datasets. To address this issue, existing research on resampling techniques was referred to, and SMOTE (Synthetic Minority Over-sampling Technique) was identified as a suitable method for mitigating class imbalance. After experimentation, the best ratio was found to be 10%, adjusting the distribution from 5301 non-bankrupt vs. 154 bankrupt to 4241 non-bankrupt vs. 424 bankrupt.

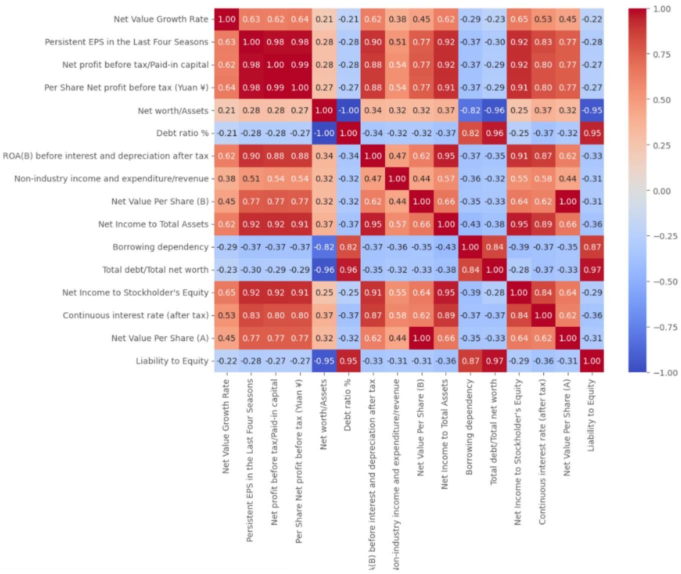


Fig. 2. Correlation Heatmap to select final features

B. Feature Selection

In addition to handling class imbalance, feature selection was performed to enhance model interpretability and efficiency. Out of the initial 95 features, only 15 were retained after filtering based on importance scores using a threshold of 0.02. Then, referring to Fig. 2, the correlation heatmap, the features were further reduced to 10 based on correlation values. As a result, dimensionality was reduced from 95 to 10 features (89.47%), minimizing the influence of less informative variables and leading to a more robust predictive model.

IV. Model Architecture

A. Model Selection

To identify the most effective bankruptcy prediction model, several standard machine learning methods were evaluated after referring to existing literature on financial risk assessment and classification algorithms:

- Logistic Regression (LR): A linear classifier estimating bankruptcy probability.
- K-Nearest Neighbors (KNN): A distance-based classifier sensitive to feature scaling.
- Decision Tree (DT): A rule-based model that splits data efficiently.
- Random Forest (RF): An ensemble of decision trees improving generalization.

- XGBoost (XGB): A gradient boosting model optimizing weak learners.
- LightGBM (LGBM): A faster gradient boosting alternative for large datasets.
- Neural Networks (NN): A deep learning model capturing complex patterns.

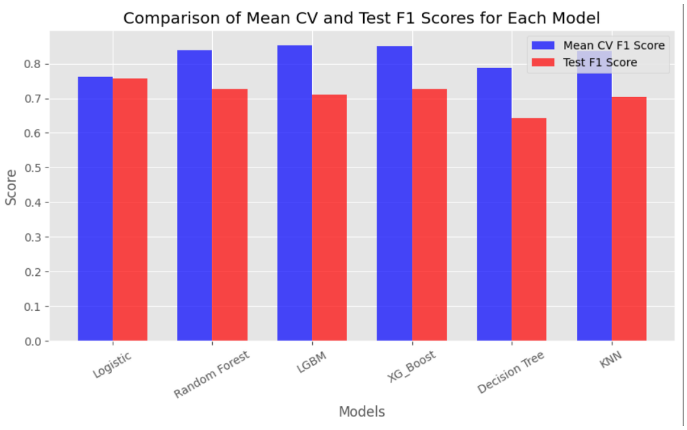


Fig. 3. Plot of CV F1 scores on training set and F1 scores on test set

B. Metric Selection

To evaluate model performance, various metrics were considered, including accuracy, precision, recall, and F1-score. Given the dataset's severe class imbalance (5301 non-bankrupt vs. 154 bankrupt cases), F1 score was chosen as the most suitable metric. F1 score calculates recall and precision for each class separately and then averages the values, ensuring that performance on the minority class is not overshadowed by the majority class. The graph in Fig.3 comparing Mean CV F1 Scores and Test F1 Scores across various models shows that Logistic Regression achieves the highest Test F1 Score. While its Mean CV F1 Score is slightly lower than some other models (LGBM and XGBoost), the superior performance on the test set, reflecting better generalization, led us to select Logistic Regression for our final model.

C. Model Tuning

To enhance the performance of the model, a Grid Search was conducted to systematically explore a range of hyperparameters, optimizing the performance of the Logistic Regression model. The optimal regularization strength (λ) and other hyperparameters like class weight, maximum iterations, regularization type etc. —were obtained to maximize the accuracy while maintaining model stability. Based on the results obtained, $L1$ (Lasso) regularization was applied to enhance generalization and prevent overfitting.

V. Final Accuracies

The hyperparameter tuning process resulted in an improvement in the model's accuracy and F1-score, as evidenced by $Fig.4$ and $Fig.5$ in the Final Accuracies section. The confusion matrix $(Fig.4)$ demonstrates the model's ability to correctly classify instances, while the classification report $(Fig.5)$ reflects enhanced precision, recall, and F1-score. These results highlight the effectiveness of the applied tuning strategies in optimizing model performance.

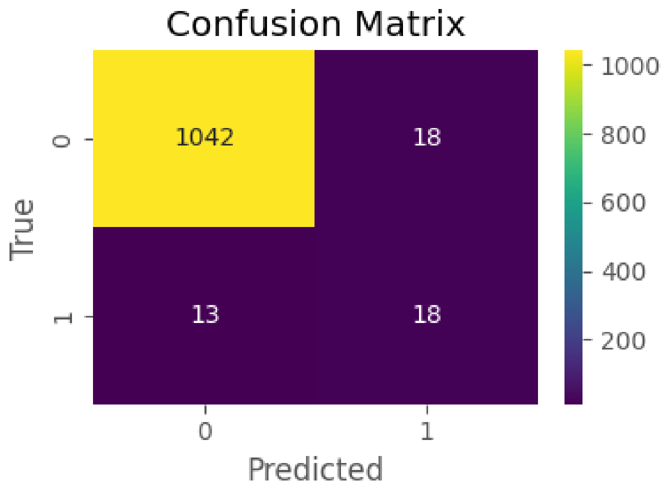
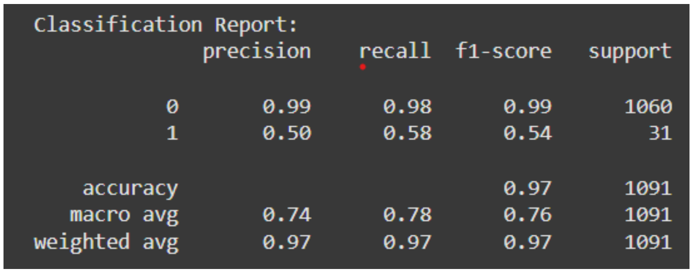


Fig. 4. Confusion Matrix of the tuned logistic regression model.



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

           0       0.99      0.98      0.99      1060
           1       0.50      0.58      0.54        31

    accuracy                           0.97      1091
   macro avg       0.74      0.78      0.76      1091
weighted avg       0.97      0.97      0.97      1091
```

Fig. 5. Classification Report of the tuned logistic regression model.

VI. Conclusion

This concludes the analysis, emphasizing the role of SMOTE, feature selection, and hyperparameter tuning in enhancing model performance.

References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *SMOTE: Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[2] I. Guyon and A. Elisseeff, *An Introduction to Variable and Feature Selection*, Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.

[3] G. Chandrashekar and F. Sahin, *A Survey on Feature Selection Methods*, Computers & Electrical Engineering, vol. 40, no. 1, pp. 16–28, 2014.

[4] E. I. Altman, *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy*, Journal of Finance, vol. 23, no. 4, pp. 589–609, 1968.

[5] J. A. Ohlson, *Financial Ratios and the Probabilistic Prediction of Bankruptcy*, Journal of Accounting Research, vol. 18, no. 1, pp. 109–131, 1980.

[6] T. Shumway, *Forecasting Bankruptcy More Accurately: A Simple Hazard Model*, Journal of Business, vol. 74, no. 1, pp. 101–124, 2001.

[7] R. Tibshirani, *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.