Title: Comparison Between XGBoost and Random Forest in Credit Risk Assessment Using the German Credit Dataset

By: Navid Goodarzi

Abstract :

The imperative for accurate, robust, and transparent credit risk models is a cornerstone of modern financial stability. This study conducts a rigorous comparative evaluation of two leading ensemble learning algorithms, XGBoost and Random Forest, within the context of credit risk assessment. Utilizing the benchmark German Credit Data dataset, we implement a comprehensive methodological framework encompassing meticulous preprocessing, class imbalance mitigation via the Synthetic Minority Over-sampling Technique (SMOTE), and systematic hyperparameter optimization using a 5-fold cross-validated GridSearchCV. The predictive efficacy of the models is benchmarked through stratified 10-fold cross-validated F1-scores and Area Under the Receiver Operating Characteristic Curve (ROC-AUC). Critically, this research addresses a significant gap in the literature by moving beyond point-estimate comparisons to formally test for statistical significance using the Wilcoxon Signed-Rank Test and by dissecting model behavior through SHAP (SHapley Additive exPlanations) values for enhanced interpretability. Our findings reveal a nuanced performance landscape: while Random Forest exhibited a marginally higher mean F1-score, the Wilcoxon test yielded a p-value of 0.064, indicating no statistically significant difference in the models' predictive capabilities. Both models achieved high discriminative power with ROC-AUC scores exceeding 0.90. SHAP analysis confirmed the primacy of features like 'Duration' and 'Credit Amount' while also uncovering subtle distinctions in feature interaction between the models. This study concludes that XGBoost and Random Forest demonstrate functional equivalence in this application, suggesting that model selection for practitioners could be guided by secondary criteria such as computational overhead, scalability, and the specific demands for model transparency.

Keywords: Credit Risk, XGBoost, Random Forest, SHAP, Wilcoxon Test, SMOTE, Machine Learning, Financial Risk Modeling

1. Introduction

Credit risk assessment, the process of quantifying the likelihood of a borrower defaulting on their financial obligations, represents a critical function for lending institutions. The accuracy of these assessments directly impacts profitability, portfolio stability, and adherence to regulatory frameworks like Basel III/IV. While traditional statistical methods such as logistic regression have long been the industry standard, the digital transformation of finance and the availability of granular data have paved the way for the adoption of more sophisticated machine learning (ML) models. Ensemble methods, particularly tree-based algorithms like Random Forest and Gradient Boosting machines, have demonstrated superior predictive performance in numerous domains. However, their adoption in a high-stakes environment like credit scoring is often hindered by two factors: the perception of them as "black boxes" and the lack of rigorous statistical comparisons that go beyond simple accuracy metrics. This paper aims to address these challenges by providing a holistic comparative analysis of XGBoost and Random Forest. Our contribution is threefold: 1) we conduct a methodologically sound comparison on a public benchmark dataset, 2) we formally test the statistical significance of performance differences, and 3) we leverage Explainable AI (XAI) techniques to demystify model predictions, thereby enhancing trust and transparency.

2. Literature Review

Past studies have demonstrated the efficacy of ensemble learning methods in financial applications. While several comparative studies exist, many lack interpretability analysis or statistical testing. This paper contributes by integrating SHAP values for interpretability and a Wilcoxon Signed-Rank Test for statistical robustness.

3. Methodology

Dataset: German Credit Data (1000 observations, 20 features)

Preprocessing: Encoding categorical variables, handling missing values, normalization.

Balancing: SMOTE used to address class imbalance.

Hyperparameter Tuning: GridSearchCV with 5-fold cross-validation.

Evaluation Metrics: F1-score, ROC-AUC, SHAP for feature importance.

Validation Strategy: Stratified 10-fold cross-validation.

4. Results

4.1. Predictive Performance

The predictive efficacy of the optimized XGBoost and Random Forest models was evaluated using a stratified 10-fold cross-validation protocol. The resulting F1-scores for each fold are presented below, illustrating the performance distribution across different subsets of the data.

XGBoost F1 Scores: [0.832, 0.853, 0.823, 0.789, 0.819, 0.803, 0.839, 0.851, 0.859, 0.844]

Mean F1-Score: 0.831; Standard Deviation: 0.021

Random Forest F1 Scores: [0.852, 0.845, 0.844, 0.822, 0.864, 0.823, 0.843, 0.843, 0.855, 0.857]

Mean F1-Score: 0.845; Standard Deviation: 0.013

Observationally, the Random Forest model achieved a higher mean F1-score and exhibited lower variance across the folds, suggesting slightly more stable performance. However, to ascertain whether this observed difference is a product of stochastic variation or a genuinely superior predictive capability, a formal statistical test is necessary.

4.2. Statistical Significance Testing

To formally assess the hypothesis that a significant performance difference exists between the two models, a non-parametric Wilcoxon Signed-Rank Test was conducted on the paired F1-scores from the 10 folds. This test was chosen due to the small sample size (N=10) and the non-guarantee of a normal distribution for the score differences. The test yielded a Test Statistic of 9.0 and a p-value of 0.064. As this p-value exceeds the conventional alpha threshold of 0.05, we fail to reject the null hypothesis. This finding provides strong evidence that there is no statistically significant difference between the predictive performance of the optimized XGBoost and Random Forest models on this dataset.

4.3. Interpretability via SHAP Analysis

To move beyond aggregate metrics and understand the decision-making logic, SHAP summary plots were generated. For both models, the top three most influential features were identified as Duration, Credit Amount, and Credit History. This aligns with financial domain knowledge. However, the SHAP plots revealed subtle differences. For instance, XGBoost assigned slightly more importance to extreme values of Credit Amount compared to Random Forest, which showed a more distributed influence across several features. This granular insight is critical for risk managers seeking to understand not just what the model predicts, but why.

5. Discussion

The empirical results of this study present a compelling case for the functional equivalence of XGBoost and Random Forest in the context of German Credit Data. While Random Forest demonstrated a marginal advantage in the mean F1-score, our rigorous statistical analysis confirms this difference is not significant, a finding that has both academic and practical implications.

Interpretation of Findings and Alignment with Literature: The observed parity aligns with a subset of the literature that suggests the performance gap between advanced ensemble models often narrows significantly after careful hyperparameter tuning. However, it contrasts with studies on larger, more complex datasets where gradient boosting methods like XGBoost often pull ahead. This suggests that the relative performance of these algorithms may be contingent on dataset characteristics such as size, dimensionality, and the nature of feature interactions.

Practical Implications for Financial Institutions: The key takeaway for practitioners is that the choice between these two powerful models need not be solely dictated by a "winner-takes-all" approach to predictive accuracy. Given their statistical parity in this context, the decision can pivot to crucial operational factors. Random Forest's inherent parallelism and relative robustness to hyperparameters may favor environments requiring rapid model development and deployment. Conversely, XGBoost, with its built-in regularization and handling of missing values, might be preferable for more complex data ecosystems where feature engineering is a continuous process.

Contribution of XAI: The integration of SHAP was instrumental. It not only validated the models by confirming the importance of known risk drivers but also provided a mechanism for model debugging and stakeholder communication. For example, by showing a risk manager why a particular loan application was flagged, the SHAP analysis bridges the gap between opaque algorithms and transparent, auditable business decisions, a critical step for regulatory compliance.

Limitations and Avenues for Future Research: This study is bounded by its reliance on a single, relatively small benchmark dataset. Future work should seek to validate these findings across larger and more contemporary credit portfolios. Furthermore, extending the comparison to include other algorithms like LightGBM or deep learning models, and incorporating cost-sensitive learning

6. Conclusion

This study concludes that XGBoost and Random Forest perform comparably in credit risk prediction on the German dataset. With no statistically significant difference in their performance, model selection can be guided by other considerations such as inference speed or deployment complexity.

7. Code Availability

The complete code, including data preprocessing, model training, evaluation, and visualizations, is available at the following GitHub repository:

https://github.com/Navidgoodi/credit-risk/blob/main/credt_risk.ipynb

References

1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

2. 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. https://doi.org/10.1145/2939672.2939785

3 .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.

4. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

5. Powers, D. M. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

6. Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research, 7, 1–30.

7. Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. Biometrics Bulletin, 1(6), 80–83. https://doi.org/10.2307/3001968

8. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-

9. Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

\