LONDON INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND DATA SCIENCE · Vol. 1 No. 1 (2026)

Robust Feature Selection for Imbalanced Tabular Datasets Using Ensemble Gradient Boosting

Daniel Whitfield
Department of Computer Science, University College London, London, United Kingdom
d.whitfield@example.org

Mei Ling Tan
School of Computing, National University of Singapore, Singapore
m.tan@example.org

DOI: 10.5281/zenodo.0000000

Abstract *— Imbalanced tabular datasets remain a persistent obstacle for supervised learning pipelines deployed in domains such as fraud detection, industrial defect screening, and rare disease diagnosis, where minority class instances carry disproportionate operational significance. This study proposes a feature selection framework that couples recursive elimination with an ensemble of gradient boosting learners trained under class-aware resampling, aiming to identify predictive variables that remain stable across resampled folds rather than variables that merely correlate with majority class noise. We evaluate the approach across several benchmark tabular collections spanning finance, manufacturing, and healthcare, comparing selected feature subsets against filter-based and wrapper-based baselines using precision-recall area, minority class recall, and selection stability indices. Results indicate that ensemble-guided selection consistently yields smaller, more interpretable feature subsets while preserving minority class detection performance relative to using the full feature space. We further examine how resampling ratio and boosting depth interact with selection stability, offering practical guidance for practitioners tuning these hyperparameters jointly rather than independently. The findings suggest that treating feature selection and imbalance correction as a joint optimization problem, rather than sequential preprocessing steps, produces more robust and auditable machine learning pipelines for high-stakes tabular prediction tasks.*

Index Terms *— feature selection, class imbalance, gradient boosting, ensemble learning, tabular data*

I. Introduction

Supervised learning pipelines deployed on tabular data routinely encounter severe class imbalance in domains such as fraud detection, industrial defect screening, and rare disease diagnosis, where the minority class carries the greatest operational cost when misclassified [1]. Feature selection is often applied as a preprocessing step to reduce dimensionality and improve interpretability, yet conventional selection criteria are typically optimized against majority class statistics, which can silently discard variables that are informative only for the minority class [2]. This mismatch between selection objective and deployment objective produces feature subsets that appear statistically sound but underperform where it matters most.

Ensemble gradient boosting methods have demonstrated strong predictive performance on imbalanced tabular problems, but their internal feature importance scores are rarely examined for stability across resampled training folds [3]. A feature that ranks highly on one resampled fold may vanish entirely on another, undermining practitioner confidence in the resulting subset even when aggregate performance metrics look acceptable [4]. Treating feature selection and class-imbalance correction as sequential, independent preprocessing steps therefore risks producing pipelines that are difficult to audit and fragile to distributional drift.

This paper proposes a feature selection framework that couples recursive elimination with an ensemble of gradient boosting learners trained under class-aware resampling, explicitly optimizing for cross-fold selection stability alongside predictive performance. We evaluate the framework on benchmark tabular collections spanning finance, manufacturing, and healthcare, comparing it against filter-based and wrapper-based baselines.

II. Related Work

Filter-based feature selection methods, including mutual information and minimum redundancy maximum relevance criteria, remain popular for their computational efficiency but are largely agnostic to class distribution, ranking features primarily by their association with the majority class signal. Wrapper-based approaches that incorporate a downstream classifier into the selection loop can partially compensate for this by evaluating subsets against task-relevant metrics, though they are typically far more computationally expensive and sensitive to the specific resampling ratio chosen for the wrapped classifier. Recent work on ensemble feature importance has examined aggregation across bootstrap samples, but few studies jointly vary resampling ratio and boosting depth to characterize how these hyperparameters interact with selection stability, leaving practitioners with limited guidance on how to tune them jointly rather than independently.

III. Methodology

The proposed framework performs recursive feature elimination guided by an ensemble of gradient boosting learners, each trained on an independently resampled version of the training data using class-aware undersampling of the majority class. Feature importance is aggregated across ensemble members at each elimination round, and features are removed only when they consistently rank in the lowest importance quantile across a majority of ensemble members rather than on a single model’s ranking.

Table 1. Comparison of feature selection methods across evaluation metrics

Method	PR-AUC	Minority Recall	Stability Index
Filter (mRMR)	0.68	0.61	0.54
Wrapper (RFE-SVM)	0.74	0.69	0.62
Full Feature Set	0.76	0.71	—
Proposed (Ensemble-GB)	0.77	0.70	0.81

A. Ensemble-Guided Elimination Procedure

At each elimination round, the ensemble is trained on K resampled folds, and a selection stability index is computed to determine whether a candidate feature subset F is retained. We define stability as the average pairwise Jaccard similarity of the top-ranked feature sets F_i obtained across folds:

$$S = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \frac{|F_i \cap F_j|}{|F_i \cup F_j|} \tag{1}$$

Elimination proceeds only while S remains above a practitioner-defined threshold, preventing the procedure from converging on a subset that performs well on average but varies substantially across resampled folds.

IV. RESULTS

Table 1 summarizes performance across the evaluated tabular collections, reporting precision-recall area, minority class recall, and the stability index defined above. The proposed ensemble-guided method achieved smaller feature subsets than the full feature space while maintaining comparable minority class recall, and produced consistently higher stability scores than both baseline approaches.

Selection stability improved most sharply as resampling ratio approached balance, but further increases in boosting depth beyond a moderate value yielded diminishing stability gains while increasing computational cost.

V. DISCUSSION

The results support treating feature selection and imbalance correction as a joint optimization problem rather than as sequential, independently tuned preprocessing steps. Practitioners tuning resampling ratio and boosting depth in isolation risk arriving at locally optimal configurations that appear satisfactory on held-out performance metrics yet exhibit poor selection stability, complicating downstream auditing and regulatory review in high-stakes deployment settings such as fraud detection and clinical screening. The interaction observed between resampling ratio and stability suggests that moderate undersampling, rather than aggressive balancing, tends to preserve more reliable minority-relevant signal.

VI. CONCLUSION

This study presented an ensemble-guided feature selection framework for imbalanced tabular data that jointly optimizes predictive performance and cross-fold selection stability. Across benchmark collections spanning finance, manufacturing, and healthcare, the approach produced smaller, more interpretable feature subsets without sacrificing minority class detection performance relative to baseline methods. Future work should examine how these findings generalize to streaming and concept-drift settings where class balance itself shifts over time.

REFERENCES

[1] H. He and E. A. Garcia, ”Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263-1284, 2009.

[2] I. Guyon and A. Elisseeff, ”An introduction to variable and feature selection,” J. Mach. Learn. Res., vol. 3, pp. 1157-1182, 2003.

[3] T. Chen and C. Guestrin, ”XGBoost: A scalable tree boosting system,” in Proc. ACM SIGKDD, 2016, pp. 785-794.

[4] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, ”SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321-357, 2002.

[5] L. Breiman, ”Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

[6] U. Kamath and J. Liu, Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning. Springer, 2021.

[7] S. Nogueira, K. Sechidis, and G. Brown, ”On the stability of feature selection algorithms,” J. Mach. Learn. Res., vol. 18, no. 1, pp. 6345-6398, 2017.