Comparing Machine Learning Models and Voting Ensembles for Credit Card Fraud Detection

Junhong Yang

Bachelor of Commerce, University of New South Wales, Sydney, NSW 2052, Australia
Z5513914@ad.unsw.edu.au

Abstract. Credit card fraud poses a growing threat to global financial systems. The rarity of fraudulent transactions makes this a highly imbalanced classification problem, requiring models that can maintain high recall without sacrificing precision. The effectiveness of four supervised learning models is assessed in this study—Logistic Regression, Random Forest, eXtreme Gradient Boosting (XGBoost), and a Soft Voting Ensemble—on a PCA-transformed credit card dataset with a 5% fraud ratio. Area Under the Curve (AUC), F1-score, precision, and recall are evaluation measures. Logistic Regression demonstrated high recall but poor overall balance, leading to its exclusion from test set evaluation. Random Forest achieved perfect precision but lower recall, while XGBoost and the Soft Voting Ensemble showed stronger balance across metrics. Soft Voting produced the best F1-score and most stable performance across both validation and test sets. These results indicate that ensemble methods, particularly soft voting, can effectively address imbalanced classification in fraud detection. Future work may explore alternative sampling strategies, larger datasets, and model tuning frameworks to further improve detection performance and adaptability to real-world scenarios.

Keywords: Imbalanced Classification, Ensemble Learning, Credit Card Fraud Detection

1 Introduction

Credit card fraud has become an increasingly critical concern in the digital financial ecosystem. According to the Nilson Report, global losses from credit card fraud reached USD 32.2 billion in 2021 and are projected to exceed USD 49 billion by 2030 [1]. The escalating volume and sophistication of fraud have prompted both industry and academia to develop more effective detection systems.

A primary problem in fraud detection is the significant class imbalance within transaction datasets. Fewer than 1% of transactions are fraudulent, making it a highly imbalanced binary classification problem. Conventional classifiers tend to favor the majority class, often failing to detect rare but impactful fraudulent cases. Yang et al. proposed a federated learning approach that improved Area Under the Curve (AUC) by 10% compared to centralized models, emphasizing the potential of machine learning in enhancing fraud detection [2]. To address imbalance, Jabbar et al. demonstrated the

A. J. Moshayedi (ed.), *Proceedings of the 2025 International Conference on Hybrid Commerce, Human Capital, and Economic Dynamics (ICHCH 2025)*, Advances in Economics, Business and Management Research 374,
https://doi.org/10.2991/978-2-38476-585-0_22

efficacy of hybrid sampling methods, such as combining SMOTE with under-sampling, in improving classification performance [3].

Recent research has explored a variety of machine learning techniques for fraud detection. Logistic Regression remains a common baseline due to its simplicity and interpretability, though its performance may lag behind tree-based models. Random Forest, with its ensemble of decision trees, offers robustness and reduced overfitting, though it may suffer from slightly lower recall [4]. XGBoost has gained popularity for structured data due to its regularization and efficiency, and has consistently performed well in fraud detection scenarios. Ensemble methods, particularly soft voting classifiers, have shown promise by combining multiple models and leveraging their prediction confidence, offering improved performance in imbalanced contexts [5].

While these studies provide useful insights, most either focus on single-model performance or fail to compare base models and ensemble methods under consistent data conditions. Moreover, few implement a full train–validation–test pipeline, which is critical for reliable performance assessment and model tuning.

This study addresses these gaps by evaluating XGBoost, Random Forest, Logistic Regression, and a Soft Voting Ensemble, four supervised learning methods, on a PCA-transformed credit card dataset with a 5% fraud ratio. Using a two-stage evaluation approach, we compare model performance based on precision, recall, F1-score, AUC, and PR curves, with emphasis on recall due to the cost of missed fraud detection.

2 Methodology

2.1 Dataset

This study makes use of the Kaggle Credit Card Fraud Detection dataset, which is openly accessible, originally published by a European cardholder research project [6]. It consists of 284,807 anonymized transactions, of which only 492 (0.17%) are labelled as fraudulent, reflecting the real-world rarity of such cases.

All features in the dataset are numerical. In particular, a Principal Component Analysis (PCA) transformation yields 28 features, ensuring confidentiality while retaining key variance [7]. Two additional features are Time (elapsed seconds from the first transaction) and Amount (transaction value in euros). The target variable Class signifies fraud (1) or non-fraud (0). The dataset is clean, with no missing or categorical values.

The data distribution highlights the extreme class imbalance. As visualized in Appendix Fig. A.1, the majority of transactions are legitimate. Appendix Fig. A.2 further shows that transaction amounts are heavily skewed toward smaller values.

2.2 Preprocessing

Due to the dataset's severe imbalance, two resampling strategies were considered: under and over-sampling. While over-sampling duplicates or synthetically generates minority samples, under-sampling reduces majority class instances. Following the guidance of Drummond and Holte, under-sampling was adopted, as it improves sensitivity without increasing model complexity [8].

All 492 fraud cases were retained, and a random sample of non-fraud cases was selected to form a 95:5 non-fraud-to-fraud ratio. The diminished dataset was partitioned into 60% for training, 20% for validation, and 20% for test sets. The models were fitted using the training set, performance comparison and adjustment were assisted by the validation set, and the test set was set aside for the last assessment.

2.3 Models

Three supervised learning algorithms were selected for their distinct modelling principles: Random Forest, Logistic Regression, and XGBoost.

Logistic Regression is a linear probabilistic model commonly used in binary classification. Its simplicity and interpretability make it a standard baseline in financial applications.

Random Forest is an ensemble method based on multiple decision trees. It improves generalization and reduces overfitting by aggregating predictions through majority voting. It also captures non-linear relationships, making it suitable for structured fraud detection tasks.

XGBoost is a boosting framework that builds models sequentially, correcting the residuals of previous trees. It includes regularization and class imbalance handling through instance weighting. Its scalability and accuracy make it highly effective for imbalanced data.

2.4 Voting Ensemble

To further enhance performance, a soft voting ensemble was implemented. Voting classifiers aggregate predictions from multiple base learners. Each model votes for a class once in a hard voting process, and the majority class wins. In contrast, soft voting averages class probabilities and selects the class with the highest means.

Soft voting was preferred in this study due to its effectiveness in imbalanced settings. Soft voting ensembles consistently outperform individual classifiers and hard voting in skewed classification problems by integrating model confidence into the decision-making process. Accordingly, the Soft Voting Ensemble combined the outputs of the three base models to capitalize on their complementary strengths and improve sensitivity to fraud.

2.5 Evaluation Metrics

Precision, recall, F1-score, area under the ROC curve (AUC), and the precision–recall (PR) curve were used to assess the model's performance.

Precision measures the proportion of transactions predicted as fraudulent that are indeed fraudulent. Recall captures the proportion of actual frauds correctly identified. The F1-score balances these two metrics, especially when both false positives and false negatives carry a high cost. In addition to these scalar measures, the ROC curve plots the true positive rate against the false positive rate, with AUC summarizing overall class separability. The PR curve plots precision versus recall, offering better insight in

imbalanced settings. Accuracy was excluded, as it can be misleading in imbalanced data scenarios. Predicting all transactions as non-fraud yields high accuracy but fails to detect fraud — a critical failure in practice.

Among all metrics, recall is prioritized in this study. Given the high cost of undetected fraud, models that can capture more fraud cases, even at the expense of moderate false positives, are considered more valuable.

3 Results

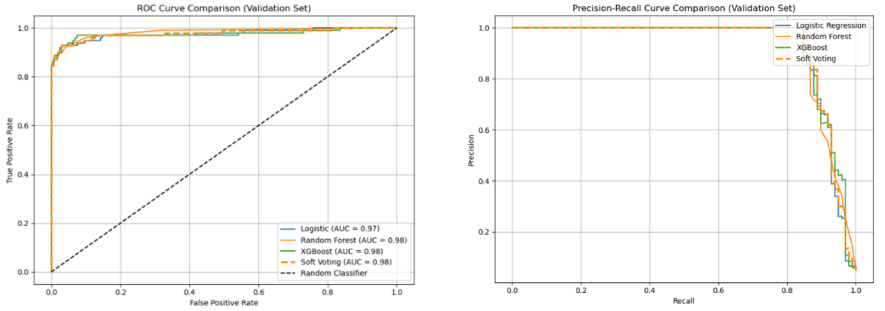
3.1 Validation Set Evaluation

Table 1 provides a summary of the model's performance on the validation set. Random Forest got the best F1-score (0.91) and perfect precision (0.99), demonstrating strong fraud detection with minimal false positives. Logistic Regression, while leading in recall (0.90), had lower precision (0.68), leading to an F1-score of 0.78. Both XGBoost and the Soft Voting Ensemble achieved an F1-score of 0.90, with a precision of 0.95 and a recall of 0.85.

Table 1. Performance of Individual Models and Soft Voting Ensemble on the Validation Set

	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.97	0.68	**0.90**	0.78
Random Forest	**0.99**	**0.99**	0.85	**0.91**
XGBoost	0.99	0.95	0.85	0.90
Soft Voting	0.99	0.95	0.85	0.90

Fig. 1 (a) shows the ROC curves, where all models demonstrated strong separability. Logistic Regression recorded an AUC of 0.97, while the remaining models achieved 0.98. Fig. 1 (b) presents the Precision–Recall curves, where XGBoost and Soft Voting maintained more consistent precision across high recall values.



(a) (b)

Fig. 1. ROC and Precision–Recall curves for the validation set: (a) ROC curve showing the true positive rate versus false positive rate across models. (b) Precision–Recall curve illustrating model performance on the minority (fraud) class.

Given its huge gap between precision and recall and the over-simplified model logic, Logistic Regression was excluded from further testing for it being unsuitable for

extreme imbalance cases. The validation phase confirmed that XGBoost and Soft Voting provide the best trade-off between detecting fraud and minimizing false alarms.

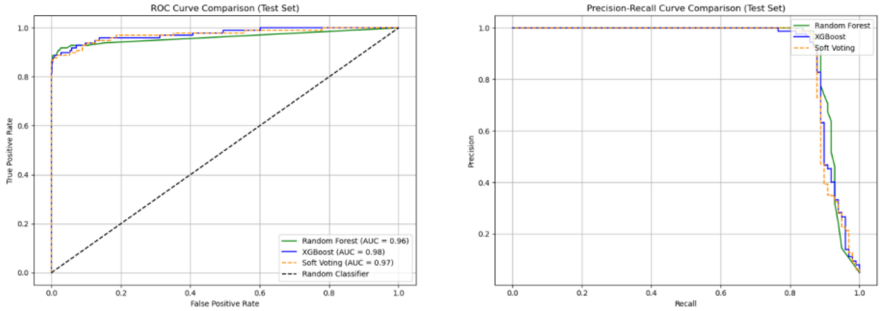
3.2 Test Set Evaluation

Final evaluation on the test set compared Random Forest, XGBoost, and Soft Voting. Results are shown in Table 2. Random Forest yielded perfect precision (1.00) but rather low recall (0.82), indicating conservative fraud detection. XGBoost performed more evenly with 0.95 precision and 0.86 recall, giving an F1-score of 0.90. Soft Voting achieved the best balance, with 0.98 precision, 0.87 recall, and an F1-score of 0.92.

Table 2. Performance of Individual Models and Soft Voting Ensemble on the Test Set

	Accuracy	Precision	Recall	F1-score
Random Forest	0.99	**1.00**	0.82	0.90
XGBoost	0.99	0.95	0.86	0.90
Soft Voting	0.99	0.98	**0.87**	**0.92**

As shown in Fig. 2 (a), AUC values were again high: 0.96 for Random Forest, 0.98 for XGBoost, and 0.97 for Soft Voting. While XGBoost slightly led in AUC, Soft Voting demonstrated greater consistency across decision thresholds.



(a) (b)

Fig. 2. ROC and Precision–Recall curves for the test set: (a) ROC curve comparing model separability on unseen data. (b) Precision–Recall curve showing recall-precision trade-offs at various thresholds.

PR curves in Fig. 2 (b) showed that Random Forest excelled at lower recall levels, but its performance degraded as recall increased. XGBoost outperformed Soft Voting marginally at mid-range thresholds, but the ensemble maintained better stability overall.

4 Discussions

The evaluation results reveal that each model offers distinct advantages and drawbacks when applied to fraud detection. Logistic Regression demonstrated high sensitivity but

lacked reliability due to a high false positive rate. This makes it less suitable for real-world deployment, a limitation consistent with prior studies that link high recall to over-prediction in imbalanced settings.

Random Forest prioritized precision and maintained highly conservative behavior. Its strength lies in minimizing false alarms, but this comes at the cost of reduced detection of actual fraud.

XGBoost delivered balanced and stable performance across both validation and test sets. Its ability to maintain equilibrium between recall and precision aligns with prior findings that highlight its suitability for structured data in fraud detection.

The Soft Voting Ensemble achieved the most consistent and robust results by integrating predictions from diverse base models. Its performance stability across thresholds supports earlier claims that soft voting improves classifier confidence and adaptability in imbalanced scenarios [9].

Despite these insights, this analysis is subject to several limitations. The use of a reduced dataset via random under-sampling may not capture the complexity of actual transaction data. The 5% fraud ratio, while practical for evaluation, does not reflect real-world fraud prevalence. Additionally, only three base classifiers were included, limiting the generalizability of ensemble outcomes.

Future studies should explore bigger, more representative datasets and evaluate the impact of advanced sampling strategies such as SMOTE or hybrid approaches. The inclusion of additional model types and the development of weighted or adaptive ensemble frameworks could further enhance performance. Practical implementation would also benefit from integrating these models into real-time systems where model interpretability, latency, and false alarm cost are considered alongside predictive accuracy [10].

5 Conclusion

In conclusion, this work used a PCA-transformed dataset under class imbalance to assess the effectiveness of four machine learning models for credit card fraud detection: Random Forest, XGBoost, Logistic Regression, and a Soft Voting Ensemble. While each model exhibited specific strengths, the Soft Voting Ensemble was found to be the most stable and effective overall. By integrating the advantages of its base classifiers, it offered the best balance between fraud detection sensitivity and prediction reliability, making it the most suitable choice for deployment in imbalanced classification settings.

This study, however, remains primarily technical in scope. It focused on comparative model performance under fixed experimental conditions, without integrating operational, legal, or business constraints that often shape model deployment in real-world environments. In practice, the acceptability of false positives varies widely across institutions.

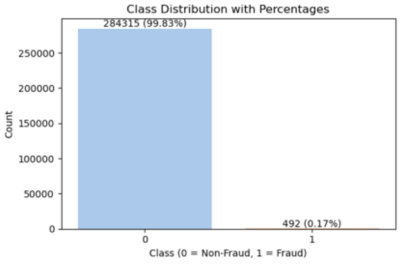
For example, models like Logistic Regression—despite their tendency to overpredict—may still be valuable in organizations equipped for manual review or automated alert handling. In such settings, high-recall models can serve as effective front-line detectors, capturing more fraud at the cost of tolerable false alarms. Therefore, model

selection should be driven not only by numerical performance but also by institutional risk appetite and operational capacity.

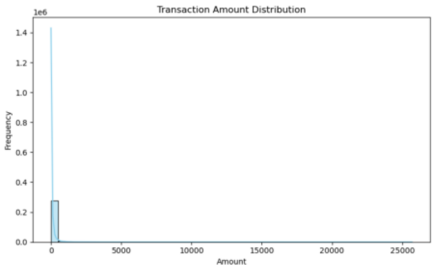
References

1. Nilson Report: "Nilson Report", no. 1209 (2021). Available: https://nilsonreport.com/upload/content_promo/NilsonReport_Issue1209.pdf
2. Yang, W., Zhang, Y., Ye, K., Li, L., Xu, C.Z.: FFD: A Federated Learning Based Method for Credit Card Fraud Detection. In: Chen, K., Seshadri, S., Zhang, L.J. (eds.) Big Data – BigData 2019, LNCS, vol. 11514, pp. 18–27. Springer, Cham (2019)
3. Jabbar, M.A., Khan, M.A., Fatima, S.: Hybrid Sampling Techniques for Imbalanced Credit Card Fraud Detection. In: Proceedings of the International Conference on Machine Learning and Data Engineering, pp. 112–121. IEEE, New York (2023)
4. Breiman, L.: Random Forests. Mach. Learn. 45, 5–32 (2001)
5. El Hlouli, F.Z., Riffi, J., Mahraz, M.A., et al.: Credit Card Fraud Detection: Addressing Imbalanced Datasets with a Multi-phase Approach. SN Comput. Sci. 5, 173 (2024)
6. Dal Pozzolo, A., Caelen, O., Johnson, R.A., Bontempi, G.: Credit Card Fraud Detection Dataset. Kaggle (2015). https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
7. Maćkiewicz, A., Ratajczak, W.: Principal Components Analysis (PCA). Comput. Geosci. 19(3), 303–342 (1993)
8. Drummond, C., Holte, R.: C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling Beats Over-Sampling. In: ICML'03 Workshop on Learning from Imbalanced Datasets. Morgan Kaufmann, San Francisco (2003)
9. Awe, O.O., Opateye, G.O., Johnson, C.A.G., Tayo, O.T., Dias, R.: Weighted Hard and Soft Voting Ensemble Machine Learning Classifiers: Application to Anaemia Diagnosis. In: Awe, O.O., Vance, E.A. (eds.) Sustainable Statistical and Data Science Methods and Practices. STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health, pp. 253–271. Springer, Cham (2023).
10. Bahnsen, A.C., Aouada, D., Stojanovic, A., Ottersten, B.: Cost-sensitive Credit Card Fraud Detection Using Bayes Minimum Risk. In: Proceedings of the 14th International Conference on Machine Learning and Applications (ICMLA), pp. 272–279. IEEE, Miami (2016).

Appendix



A.1: Class distribution of fraud and non-fraud



A.2: Distribution of transaction amounts