Article

Not peer-reviewed version

Comparative Analysis of Supervised Learning Models for Detecting Credit Card and Bank Account Fraud

Shruti Chandna *

Posted Date: 15 October 2025

doi: 10.20944/preprints202510.1007.v1

Keywords: fraud detection; machine learning; class imbalance; SMOTE; XGBoost; financial analytics

Article

Comparative Analysis of Supervised Learning Models for Detecting Credit Card and Bank Account Fraud

Shruti Chandna

Harrisburg University of Science and Technology; schandna@my.harrisburg.edu

Abstract

The purpose of this study is to investigate the efficacy of three supervised learning models, Logistic Regression, Random Forest, and XGBoost, on two datasets of financial fraud detection that were constructed differently with differing class distributions. The Credit Card Fraud Detection Dataset (Kaggle, 2023) is a synthetic dataset that has been artificially balanced to produce a 50:50 relative proportion of fraudulent and non-fraudulent observations to allow for the performance of the models to be evaluated under ideal conditions. On the other hand, the Bank Account Fraud Dataset (NeurIPS, 2022) reflects real-world monetary behavior and features extreme class imbalance characterized by only approximately 1% of the observations containing fraudulent behavior. (Jesus et al., 2022) A single pipeline was constructed using stratified 60 / 20 / 20 splits and SMOTE applied only to the training set, Evaluation metrics included F1-score and AUC-ROC. The results reflect close to perfect outcomes on the balanced synthetic dataset but large degradation in performance on the real-world imbalanced dataset. The model that consistently performed best on the imbalanced dataset was XGBoost as represented by the F1 (23.4%) and AUC (89.3%) values. These results are consistent with published benchmarks indicating that F1-scores in the 15 to 25% range represent excellent outcomes in practice in detection of fraudulent behavior. The results of the present study underscore the critical impact of data imbalance and real-world practicality of the dataset used in the performance of supervised models and indicate future study to apply techniques such as cost-sensitive learning, explainability and temporal modeling of financial data in operational settings in order to achieve generalization with the models tested.

Keywords: fraud detection; machine learning; class imbalance; SMOTE; XGBoost; financial analytics

1. Introduction

Fraud detection is a crucial field of financial analytics, where machine learning algorithms are used to identify suspicious activity and prevent losses. There are still significant challenges despite the progress in predictive modelling through advanced methods, related to data imbalance and a variety of specific fraud patterns. This study investigates how three machine learning algorithms, Logistic Regression, Random Forest, and XGBoost, behave over two datasets with significantly different statistical distributions. The preprocessing methodologies, sampling techniques, and evaluations are kept constant in order to study the effect that the structure of the datasets has on the behavior and stability of the models.

2. Datasets

2.1. Credit Card Fraud Detection Dataset (Kaggle, 2023)

This dataset consists of credit card transactions made by European credit cardholders in 2023. It contains over 550,000 records. Each record of transaction contains a unique identifier variable (id), a set of twenty-eight anonymized numerical variables (V1–V28) which give certain characteristics of the transaction (time, location, etc.) as well as other variables derived from them, the Amount variable giving the value of the transaction made, and a binary target variable Class which identifies whether or not the transaction was fraudulent (1) or non-fraudulent (0). In this investigation, a synthetic, balanced dataset composed of equal numbers of fraudulent and non-fraudulent records was used to assess the performance of models under idealized experimental conditions.

2.2. Bank Account Fraud Dataset (NeurIPS, 2022)

The NeurIPS 2022 dataset simulates fraudulent bank account openings using demographic and behavioral variables (e.g., age, income, failed_logins). Fraudulent accounts make up about 1.1% of all records, reflecting a highly imbalanced real-world scenario.

3. Methodology

3.1. Preprocessing

All features were numeric; thus, one-hot encoding was unnecessary. Data were standardized and split into training (60%), validation (20%), and test (20%) subsets using stratified sampling. SMOTE was applied only to the training portion of the Bank Fraud dataset to address imbalance while avoiding data leakage. (How Smote Oversampling Can Cause Data Leakage In Cross-Validation, n.d.)

3.2. Model Development

Three models were trained under identical conditions:

1. Logistic Regression – interpretable linear baseline model.
2. Random Forest – an ensemble of decision trees capturing nonlinear relationships.
3. XGBoost – gradient-boosted decision trees optimized for imbalanced data.

Hyperparameters were tuned using the Optuna package with validation F1-score as the optimization target. Experiments were repeated under six random seeds to assess variance and model stability.

3.3. Evaluation Metrics

Performance was assessed using:

- F1-score: Harmonic mean of precision and recall.
- AUC-ROC: Area under the receiver operating characteristic curve.

4. Results

4.1. Credit Card Dataset (Balanced 50:50)

Model	F1	AUC-ROC	CV	Stability
Logistic Regression	99.84%	99.98%	0.018%	99.98%
Random Forest	99.97%	99.99%	0.003%	99.99%
XGBoost	99.98%	99.99%	0.005%	99.99%

All models performed nearly perfectly, demonstrating that when data are balanced and well-structured, even simpler algorithms can achieve excellent performance.

4.2. Bank Account Dataset (Imbalanced ≈1.1% Fraud)

Model	F1	AUC-ROC	CV	Stability
Logistic Regression	20.36%	86.95%	1.76%	98.9%
Random Forest	18.21%	83.02%	11.82%	89.5%
XGBoost	23.41%	89.34%	3.01%	98.3%

The sharp performance drop illustrates the complexity of detecting fraud in imbalanced data. Nevertheless, F1-scores between 15–25% align with real-world benchmarks for operational fraud detection systems (Gupta et al., 2023; Ryman & Lee, 2022; Singh et al., 2024).

4.3. Comparative Summary

Metric	Credit Card	Bank Fraud	Difference
Mean F1 (XGBoost)	99.97%	23.41%	↓76.56%
Mean AUC (XGBoost)	99.99%	89.34%	↓10.65%

These results confirm that class imbalance and weak signal-to-noise ratios in behavioral data are the primary drivers of degraded model accuracy and recall.

5. Discussion

This study shows that the nature of the data rather than the complexity of the models is most critical in the detection of fraudulent actions that involve money. The balanced data set produced a near perfect separation between all algorithms, whereas the great imbalance produced realistic but much lower F1-scores. The Random Forest model showed a large variance in the accuracy of said data, whereas the Logistic model, while constant and stable in behavior, had a limited recall. The XGBoost model gave the best balance of high precision and good recall validating the claim of the noise resistivity of these models. It should be noted that these "low" scores in the case of the imbalanced data should not be felt to indicate a poor performance of the statistical model employed. In reality the exposure of fraudulent transactions means that these operations would compose invariably less than 1% of the total records, whereby the inevitable lower F1-scores thus produced. For example, as divulged by Kulatilleke and Samarakoon (2022), the evaluation metrics F1 and precision, in a state of massive data imbalance and noisy data, would rapidly deteriorate where mass processing would be effective. (Kulatilleke et al., 2022) With Popova et al. (2025) also pointing to legitimate F1 scores between 15–25% being high-grade real-world operational scores for such models. (Alshameri et al., 2023) Provided, of course, the models yield some degree of increase in valid overall recall and precision over those of random and rule-based operational systems. The model of operation, where fraud detection was involved, was not to achieve full accuracy, but satisfactory detectability, and served well to achieve a high final recall percentage. Thus, making successful detection without significant loss of the machine time resources in processing millions of records, constant false alarms, and so many millions of wasted machine hours in management and summing of said statistics. The actual modelling fulfills these obligations in getting to let us see that the academic and operational bodies both have accepted in their research that in these data conditions, is needed the virtues of interpretability, generalizability, and instance stability rather than absolute F1 scores.

6. Conclusions and Future Work

This research illustrates that data structure, rather than total algorithm complexity, is the driving problem for success in financial fraud detection. Although idealized balanced synthetic datasets produce nearly perfect results, realistic of these are rare. The very dramatic fall in F1 from a nearly 100% score in the balanced Credit Card dataset to a mere 23% in the imbalanced Bank Fraud dataset illustrates clearly how data imbalance, feature realism, and class noise restrict the performance limits of models.

Based on the empirical studies made in these situations, the paper recommends a two-tier benchmark framework for future fraud work: (1) idealized datasets for tuning algorithms, and (2) realistic datasets for operation in real systems. In addition, it would change the meaning of success in analytics, as it has been shown that moderate F1 performance is nevertheless consistent with high importance in business value if realistic class imbalance is assumed.

In future work, it is suggested that this area might be extended by looking at adaptive resampling, dynamic thresholds, and temporal pattern learning as methods of bridging the gap much more closely between laboratory performance and results in live deployment.

To bridge the gap between lab performance and actual performance, future directions should involve:

- Adaptive Resampling and Dynamic Thresholds, whereby sample fractions and cut off points will be changed concerning the evolving distribution of fraud.
- Cost Sensitivity and Focal Loss Training will improve recall on rare fraud classes.
- Temporal and Sequential Modelling (LSTM, Transformers) will resolve to model dynamic transaction structures.
- Explainable AI methods (SHAP, LIME) so sectors are transparent and confident.
- Semi-supervised and On-Line Learning whereby the model will be capable of changes/augmentations with sensitivity to technology applications.
- Two Tier Benchmarks which actually condition the models on both ideal and realistic data to ensure meaningful application.
- Implementation of the aforementioned avenues will help to dampen the existing gulf between experimental results and actual application to ensure enhancement of actual application ensuring technological robustness and business benefit for the counterskills in fraud prevention.

Use of AI Tools: The author used OpenAI's ChatGPT to assist in language refinement and formatting during manuscript preparation. The author reviewed and approved all content and takes full responsibility for the analysis and conclusions.

Code Availability: The Python scripts implementing data preprocessing, model training, and evaluation (including SMOTE balancing and F1 optimization pipeline) are available as supplementary material. The code can be freely reused under an open academic license for educational and research purposes.

Data Availability: Both datasets used in this study are publicly available: Credit Card Fraud Detection Dataset (Kaggle, 2023): https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023**;** Bank Account Fraud Dataset (NeurIPS 2022): https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD Conference*, 785–794.

3. Gupta, A., Kumar, N., & Yadav, S. (2023). Evaluating machine-learning models for financial fraud detection under severe class imbalance. *Expert Systems with Applications, 228*, 120579. https://doi.org/10.1016/j.eswa.2023.120579
4. Ryman, A., & Lee, D. (2022). Real-world fraud detection benchmarks with extreme imbalance: A comparative study. *IEEE Access, 10*, 62493–62508. https://doi.org/10.1109/ACCESS.2022.3179235
5. Singh, P., Raj, R., & Chatterjee, S. (2024). Understanding performance limits of supervised fraud detection under class imbalance. *Information Sciences, 658*, 120992. https://doi.org/10.1016/j.ins.2024.120992
6. Kulatilleke, S., & Samarakoon, M. (2022). *Empirical study of machine learning classifier evaluation metrics behavior in massively imbalanced and noisy data*. arXiv preprint arXiv:2208.11904. https://arxiv.org/abs/2208.11904
7. Popova, E., Dubrova, A., & Thalheim, L. (2025). *Credit card fraud detection: Model evaluation under class imbalance*. arXiv preprint arXiv:2509.15044. https://arxiv.org/pdf/2509.15044
8. Credit Card Fraud Detection Dataset (Kaggle, 2023): https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023
9. Bank Account Fraud Dataset (NeurIPS 2022): https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.