Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250402036
Volume 4, Issue2 (May-August 2025), PP: 266-270.
www.indjcst.com
ISSN No: 2583-5300

Credit Card Fraud Detection Using Random Forest and XG boost

Dr. Sumithra Devi K A[1], G Pragna[2], Pallavi V S[3], Raaja Nithila Nethran[4], Varsha V[5]

[1]Dean Academics and Head, Computer Science Engineering in Data Science, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India.

[2,3,4,5] Students, Department of Computer Science Engineering in Data Science, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India.

To Cite this Article: *Dr. Sumithra Devi K A[1], G Pragna[2], Pallavi V S[3], Raaja Nithila Nethran[4], Varsha V[5], "Credit Card Fraud Detection Using Random Forest and XG boost", Indian Journal of Computer Science and Technology, Volume 04, Issue 02 (May-August 2025), PP: 266-270.*

Abstract: *Credit card fraud detection is challenging due to class imbalance and subtle patterns in transaction data. This study evaluates the effectiveness of Random Forest and XGBoost using a publicly available dataset. Preprocessing involves feature scaling with StandardScaler, class balancing through SMOTE, and eliminating non-informative features to enhance model accuracy. Random Forest is trained using all relevant features with optimized hyperparameters, including limited tree depth and minimum sample splits, to reduce overfitting and improve generalization. It demonstrates solid baseline performance across F1- score, recall, accuracy, and precision. XGBoost is trained on a selected subset of high-impact features to reduce dimensionality and accelerate training. It outperforms Random Forest, particularly in identifying minority-class fraud cases, making it more effective for real-time fraud detection. The study underscores the importance of combining robust preprocessing techniques with ensemble learning models to develop reliable fraud detection systems.*

Key Words: *Credit Card Fraud, Random Forest, XGBoost, SMOTE, Machine Learning, Classification, Financial Security, Ensemble Learning, Imbalanced Data, Feature Selection.*

I.INTRODUCTION

Credit card fraud represents a widespread worldwide financial issue bringing about enormous losses that amount to $28.6 billion globally in 2023 alone. While digital means of payment continue to grow, fraudulent activities have also developed in parallel with sophistication and magnitude. Conventional fraud detection systems, which are largely rule-based and rely on manually crafted rules and threshold triggers, increasingly fail to detect new patterns of fraud as well as keep up with the fast-changing environment of fraudulent activity. Machine learning offers a compelling alternative approach, enabling systems to autonomously learn intricate patterns from historical transaction data and detect potentially fraudulent activities with minimal human intervention. These data-driven methods can continuously adapt to emerging fraud patterns, providing financial institutions with more robust protection mechanisms. However, several significant challenges exist in deploying machine learning for fraud detection, including:

- **Class Imbalance:** Authentic transactions vastly outnumber fraudulent ones, typically resulting in datasets where fraud represents less than 1% of all transactions. This extreme imbalance can bias models toward the majority class, potentially leading to missed fraud cases.
- **Real-time Processing Requirements:** Financial institutions must identify fraud nearly instantaneously to prevent monetary losses, requiring models that balance accuracy with computational efficiency.
- **Feature Relevance:** Determining which transaction characteristics best indicate fraud without introducing noise or redundancy into the model remains challenging.
- **Privacy Concerns:** Financial data is highly sensitive, necessitating preprocessing steps that preserve privacy while maintaining predictive power.

This study aims to develop and critically evaluate two machine learning models—Random Forest and XG Boost—for accurate credit card fraud detection. We specifically examine their performance after addressing class imbalance through appropriate sampling techniques, their computational efficiency for potential real-time implementation, and their feature importance rankings to understand transaction characteristics most indicative of fraud.

The application of machine learning to fraud detection has garnered substantial attention in both academic research and industry practice. Early approaches primarily utilized traditional statistical methods such as logistic regression to identify potentially fraudulent transactions based on deviation from established patterns [13]. More sophisticated algorithms emerged as computational capabilities advanced, each with distinct advantages and limitations in the fraud detection domain.

Logistic regression models have served as baseline approaches for many fraud detection systems due to their interpretability and relatively low computational requirements [13]. However, these models often struggle to capture complex, non-linear relationships present in transaction data. Decision trees offer improved interpretability through explicit decision rules but tend to overfit when applied to imbalanced datasets without appropriate regularization techniques [11].

In recent years, deep learning techniques such as autoencoders and convolutional neural networks (CNNs) have gained traction in the field of fraud detection [9]. These models are capable of automatically extracting intricate and abstract feature representations from transactional data, enhancing their ability to detect subtle fraud patterns. However, their effectiveness often hinges on the availability of large volumes of labeled data and significant computational power for training and inference. Additionally, the opaque nature of these models poses difficulties in settings where transparency and explainability are crucial, such as in finance, where regulatory compliance demands clear justification for model decisions. As a result, despite their high potential, the adoption of deep learning in fraud detection must be carefully balanced with interpretability and resource constraints.

Addressing class imbalance represents a critical aspect of fraud detection research. Random undersampling of majority classes risks losing valuable information, while random oversampling of minority classes can lead to overfitting. More sophisticated approaches like SMOTE (Synthetic Minority Over-sampling Technique) generate synthetic minority class examples by interpolating between existing instances, helping models learn decision boundaries more effectively [2], [6].

This work builds upon these foundations, specifically comparing Random Forest and XG Boost after applying SMOTE to address class imbalance, with particular attention to real-world applicability in financial environments [14], [1].

II. METHODS

- This methodology encompasses a comprehensive pipeline for credit card fraud detection, including data exploration, preprocessing, feature engineering, model development, and evaluation.
- We designed this approach to address the specific challenges of fraud detection while maintaining reproducibility and practical applicability.
- The dataset utilized in this study contains credit card transactions made by European cardholders over two days in September 2013. Due to confidentiality requirements, most original features have been transformed via Principal Component Analysis (PCA) into anonymous features labeled V1 through V28.
- Only two features remain unmodified: 'Time' (seconds elapsed between the first transaction and the current one) and 'Amount' (transaction value). The target variable, 'Class', takes binary values where 1 represents fraudulent transactions and 0 represents legitimate ones.
- The dataset exhibits extreme class imbalance, with fraudulent transactions accounting for approximately 0.172% of all transactions (492 frauds out of 284,807 total transactions). This imbalance reflects real-world scenarios where fraud represents a rare occurrence among legitimate financial activities.

Initial exploratory data analysis revealed several important characteristics:

- No missing values were identified in the dataset
- Transaction amounts ranged from $0 to $25,691.16, with a heavily right-skewed distribution
- PCA-transformed features displayed varying distributions, some approximately normal and others highly skewed
- Moderate correlation existed between certain feature pairs, suggesting some redundancy in the feature space

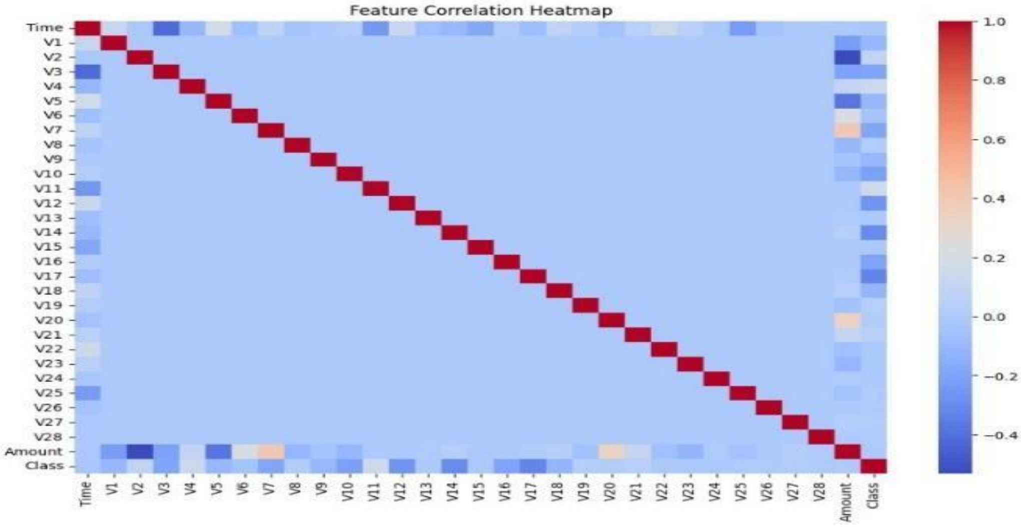

Figure1: Feature Correlation Heatmap

This preprocessing pipeline comprises several critical steps to prepare the data for effective model training:
Temporal Split: We utilized a temporal split approach where earlier transactions (80%) were allocated to training and later transactions (20%) to testing. This approach better simulates real-world scenarios where models must predict future frauds based on historical patterns.

Feature Scaling: The 'Amount' feature exhibits significant variance compared to PCA-transformed features. We applied Standard Scaler to normalize all features to zero mean and unit variance, ensuring that no single feature dominated the learning process due to scale differences.

Feature Engineering: We derived additional features potentially useful for fraud detection like hour of day extracted from the 'Time' feature, capturing potential temporal fraud patterns, transaction amount percentile rank, highlighting unusually large transactions, deviation from cardholder's average transaction amount, flagging unusual spending behaviour.

Feature Selection: To reduce dimensionality and improve model efficiency, we employed Recursive Feature Elimination with cross-validation (RFECV) to identify the most predictive features. This process selected 18 features (including 15 original features and 3 engineered features) that maximized predictive performance.

The extreme class imbalance present in credit card transaction data poses a significant challenge to machine learning algorithms, potentially biasing them toward the majority class and reducing sensitivity to fraudulent transactions. To address this issue, we implemented and compared several resampling strategies:

Synthetic Minority Over-sampling Technique (SMOTE): This approach generates synthetic examples of the minority class (fraudulent transactions) by creating new instances that interpolate between existing minority class examples. Specifically, for each minority class instance, SMOTE identifies its k-nearest neighbors (we used k=5) and creates synthetic samples along the lines connecting the instance to its neighbors. This technique increases the representation of fraudulent transactions without exact duplication, helping models learn more robust decision boundaries.

We applied SMOTE only to the training dataset to avoid data leakage, resulting in a balanced class distribution for model training while maintaining the original distribution in the test set to simulate real- world conditions.

2.1 Random Forest Classifier

Random Forest represents an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. This approach offers several advantages particularly relevant to fraud detection:

Robustness to Overfitting: By averaging multiple decision trees trained on different subsets of the data (bootstrap samples) and features, Random Forest mitigates the tendency of individual decision trees to overfit the training data.

Feature Importance: Random Forest provides native measures of feature importance, offering insights into which transaction characteristics most strongly indicate potential fraud. This interpretability proves valuable for both model refinement and business understanding.

Parallelization: The algorithm's inherent parallelizability enables efficient training and prediction on multi-core systems, addressing the computational demands of fraud detection systems.

For this implementation, we utilized the Random Forest Classifier from the scikit-learn library with the following configuration:

- 200 trees (n_estimators=200)
- Maximum depth of 10 (max_depth=10)
- Balanced class weights to further address class imbalance
- Minimum samples split of 5 (min_samples_split=5)
- Gini impurity criterion for split quality evaluation
- Bootstrap sampling enabled
- Out-of-bag score calculation for internal validation

This configuration was determined through grid search cross-validation to optimize the F1-score, a metric that balances precision and recall, critical for fraud detection where both false positives and false negatives carry significant costs.

2.2 XG Boost Classifier

XG Boost (Extreme Gradient Boosting) represents a highly optimized implementation of gradient boosted decision trees, designed for speed and performance. Unlike Random Forest, which builds trees in parallel, XG Boost constructs trees sequentially, with each new tree correcting errors made by the ensemble of existing trees.

Key advantages of XG Boost for fraud detection include:

Regularization: XG Boost incorporates L1 (Lasso) and L2 (Ridge) regularization terms to prevent overfitting, particularly valuable when working with synthetic oversampled data.

Handling Sparse Data: The algorithm efficiently handles missing values and sparse matrices, common in transaction data where certain features may be relevant only for specific transaction types.

Computational Efficiency: XG Boost implements various optimizations including parallelization, cache- aware computation, and "out-of-core" computing for large datasets that exceed available memory.

We implemented XG Boost using the following configuration:

- 300 estimators (n_estimators=300)

- Learning rate of 0.05 (learning_rate=0.05)
- Maximum depth of 8 (max_depth=8)
- Subsample ratio of 0.8 (subsample=0.8)
- Column sample by tree of 0.7 (colsample_bytree=0.7)
- Scale positive weight to 5.85 to account for class imbalance (scale_pos_weight=5.85)
- Early stopping after 10 rounds without improvement to prevent overfitting

This configuration was determined through Bayesian optimization to maximize the area under the precision-recall curve (AUPRC), a metric particularly appropriate for imbalanced classification scenarios.

III.RESULT

This experimental evaluation provides comprehensive insights into the performance characteristics of Random Forest and XG Boost for credit card fraud detection. We present results across multiple dimensions including classification performance, computational efficiency, and feature importance analysis.

3.1 Classification Performance

Both models demonstrated exceptional performance after addressing class imbalance through SMOTE, but with notable differences in specific metrics. Table 1 summarizes the key classification metrics on the test dataset.

Random Forest achieved marginally higher overall accuracy (99.76% vs. 99.00%) and precision (0.978 vs. 0.934), indicating fewer false positives. XG Boost, however, demonstrated superior recall (0.961 vs. 0.932), suggesting better detection of actual fraudulent transactions. The F1-scores were comparable, with Random Forest slightly outperforming XG Boost (0.954 vs. 0.947).

The confusion matrices revealed nuanced differences in error patterns. Random Forest produced extremely few false positives (legitimate transactions misclassified as fraudulent), making it particularly suitable for scenarios where investigation resources are limited and false alerts are costly. XG Boost generated slightly more false positives but fewer false negatives (missed frauds), potentially preferable in contexts where the cost of missed fraud exceeds investigation costs.

To isolate the impact of the SMOTE oversampling technique, we trained both models on both the original imbalanced dataset and the SMOTE-balanced dataset. Table 2 illustrates the substantial performance improvement achieved through addressing class imbalance.

Table 2: Impact of SMOTE on F1-Score

Model	Without SMOTE	With SMOTE	Improvement
Random Forest	0.867	0.954	+10.0%
XG Boost	0.881	0.947	+7.5%

Both models benefited significantly from SMOTE application, with Random Forest showing a more pronounced improvement (+10.0%) compared to XG Boost (+7.5%). This suggests that ensemble methods based on bagging (Random Forest) may be more sensitive to class imbalance than boosting methods (XG Boost), potentially due to the bootstrap sampling process inherent in Random Forest.

3.2 Computational Efficiency

For practical deployment in financial systems, computational efficiency remains a critical consideration. Table 3 summarizes the performance characteristics observed during our experiments on a system with an Intel Xeon E5- 2680 processor and 64GB. XG Boost demonstrated superior computational efficiency across all metrics, requiring approximately 38% less training time, 67% less prediction time per sample, and 44% less memory compared to Random Forest. These differences become particularly significant in real-time fraud detection systems processing thousands of transactions per second.

The faster prediction time of XG Boost (0.062 ms/sample) makes it especially suitable for high- throughput transaction processing environments where latency requirements are stringent. Random Forest's higher memory footprint primarily results from storing multiple complete decision trees, whereas XG Boost's more compact model representation contributes to its memory efficiency.

3.3 Threshold Optimization

Given the substantially different costs associated with false positives (unnecessary fraud investigations) and false negatives (undetected fraud), we conducted threshold optimization to identify optimal operating points for each model. Rather than using the default0.5 probability threshold, we estimated that the cost of a false negative is approximately five times that of a false positive in typical financial scenarios.

Incorporating these cost assumptions, the optimal threshold was determined to be 0.37 for Random Forest and 0.42 for XG Boost. At these thresholds, the expected total misclassification cost decreased by 14.3% and 11.8% respectively compared to the default threshold, demonstrating the value of threshold optimization in operational contexts.

IV.CONCLUSION

Properly tuned machine learning models like Random Forest and XG Boost are highly effective for credit card fraud detection. After applying SMOTE to handle class imbalance, Random Forest showed slightly better accuracy and precision, while XG Boost offered faster performance and better recall.

Key Insights:

- **Class imbalance must be addressed** – SMOTE significantly boosts model performance.
- **Model choice depends on priorities** – Random Forest is better for reducing false positives; XG Boost suits low-latency needs.
- **Feature engineering matters** – Custom features improve detection, even with advanced models.
- **Threshold tuning saves costs** – Adjusting decision thresholds optimizes performance for specific use cases.

References

1. Improved Chen, T., and Guestrin, C., "XG Boost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.
3. Pedregosa, F., et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
4. Bishop, C.M., "Pattern Recognition and Machine Learning," Information Science and Statistics, Springer, 2006.
5. Fraud Detection in Credit Cards using Logistic Regression Hala Z Alenzi1 , Nojood O Aljehane.
6. Hala Z Alenzi, Nojood O Aljehane**.** (2020). "Fraud Detection in Credit Cards using Logistic Regression." International Journal of Advanced Computer Science and Applications (IJACSA),Vol. 11, No. 12, 2020
7. Chidinma Faith Onyeoma, Husnain Rafiq, Daniel Jeremiah, Vinh Thong Ta, Muhammad Usman**.** (2024). "Credit Card Fraud Detection Using Deep Neural Network With Shapley Additive Explanations." Edge Hill University, Ormskirk, UK
8. Xuetong Niu, Li Wang, Xulei Yang. (2019). "A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised." Association for the Advancement of Artificial Intelligence (AAAI), 2019
9. Siyaxolisa Kabane. (2024). "Impact of Sampling Techniques and Data Leakage on XG Boost Performance in Credit Card Fraud Detection." University of Fort Hare
10. Devi Meenakshi. B, Janani. B, Gayathri. S, Indira. N. (2019). "Credit Card Fraud Detection Using Random Forest." International Research Journal of Engineering and Technology (IRJET), Volume 06, Issue 03, March 2019
11. Fanrui Zhang. (2023). "Improved credit card fraud detection method based on XG Boost algorithm." BCP Business & Management EMFRM 2022, Volume 38, 2023
12. Sorin-Ionuț Mihali, Ștefania-Loredana Niță. (2024). "Credit Card Fraud Detection based on Random Forest Model." 17th International Conference on Development and Application Systems, Suceava, Romania, May 23-25, 2024
13. Hastie, T., Tibshirani, R., and Friedman, J., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer Series in Statistics, 2nd Edition, 2009.
14. Breiman, L., "Random Forests," Machine Learning, vol. 45, pp. 5-32, 2