WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Fraud Detection on Bank Payments Using Machine Learning

Dr. Lalitha. T[1], Lakshmi. V[2]

[1]Department of Computer Science Engineering and Information Science, Presidency University, Banglore, India
lalithasrilekha31@gmail.com
[2]Master of Computer Applications, Presidency University, Banglore, India lakshv18may@gmail.com

ABSTRACT:

In the evolving landscape of digital finance, fraudulent banking transactions pose a significant threat to financial institutions and customers alike. This paper presents a machine learning framework for detecting fraudulent banking transactions using real-world data. Algorithms such as Random Forest, K-Nearest Neighbors (KNN), and XG Boost are applied to transactional data after addressing severe class imbalance using SMOTE. The study includes in-depth data visualization, feature analysis, and model evaluation using metrics such as confusion matrix, precision, recall, and ROC-AUC. Experimental results show that ensemble learning models like XG Boost outperform others in accuracy and robustness, establishing their efficacy in real-world fraud detection systems.

Keywords: Fraud Detection, Machine Learning, SMOTE, Random Forest, XGBoost, Class Imbalance, Financial Security

I. Introduction:

With the rise of online banking and cashless transactions, fraudulent activities have become increasingly sophisticated. Traditional rule-based fraud detection systems are no longer sufficient. Machine learning offers a scalable, adaptive, and intelligent alternative capable of identifying hidden patterns in transactional data. This paper explores various machine learning algorithms and preprocessing strategies to build a robust fraud detection model.

II. Problem Statement

Financial fraud detection is a highly imbalanced classification problem where fraudulent cases are rare compared to legitimate ones. The challenge lies in accurately identifying fraudulent transactions with minimal false positives, thereby minimizing customer inconvenience and institutional losses.

III. Objectives

- To preprocess and explore real-world transaction data for fraud detection.
- To apply class balancing using SMOTE.
- To train and evaluate models including Random Forest, KNN, and XG Boost.
- To compare model performance using key evaluation metrics.
- To suggest future enhancements for improving fraud detection accuracy.

IV. Literature Review

Previous studies have highlighted various approaches to fraud detection on bank payments:

- **Dal Pozzolo et al. (2015)** explored the use of ensemble methods combined with undersampling techniques to detect fraudulent credit card transactions, achieving improved accuracy in highly imbalanced datasets.
- **Bahnsen et al. (2016)** applied cost-sensitive classification techniques to real-world banking data to minimize financial loss, reducing false positives in credit card fraud detection systems.
- **Carcillo et al. (2019** investigated hybrid fraud detection models that combined supervised and unsupervised learning approaches, achieving higher anomaly detection rates in transactional datasets.

These studies confirm that machine learning models, when effectively optimized and validated, can significantly improve fraud detection performance in financial systems.

V. Methodology

A. Dataset Description

The dataset used includes historical banking transaction data with attributes such as:

- Transaction Amount
- Transaction Category
- Time of Transaction
- Customer ID (anonymized)
- Merchant Information (anonymized)

Fraud Label (0 = Non-Fraud, 1 = Fraud)

Data was collected from a publicly available credit card fraud dataset designed for research and machine learning analysis.

B. Data Preprocessing

- Removal of missing values and validated data types.
- Visualized spending behaviour using boxplots for each transaction category.

Observed stark class imbalance (~1% fraud cases), requiring resampling.

C. Feature Engineering

- Category-wise fraud analysis highlighted risk-prone transaction types.
- Boxplots revealed significant variance in amount spent across categories.
- Statistical summaries provided insight into feature distributions.

D. Models Applied

- **Random Forest**: Captures non-linear interactions using ensemble decision trees.
- **K-Nearest Neighbors (KNN)**: Simple, instance-based learner effective for pattern detection.
- **XGBoost**: Advanced boosting algorithm well-suited for handling imbalanced classification.

E. Model Training and Tuning

- 70-30 split for Training and Testing sets.
- SMOTE applied to the training set to handle class imbalance.
- Hyperparameter tuning performed using Grid Search for Random Forest and XGBoost classifiers.
- Model performance evaluated using confusion matrix, precision, recall, F1-score, and ROC-AUC metrics.

VI. Results and Evaluation

A. Evaluation Metrics

- Confusion Matrix
- Precision
- Recall
- F1 Score
- ROC-AUC Score

Model	Precision	Recall	F1 Score	ROC-AUC
Random Forest	0.89	0.81	0.85	0.93
K-Nearest Neighbors (KNN)	0.75	0.68	0.71	0.85
XGBoost	**0.91**	**0.86**	**0.88**	**0.96**

B. Interpretation

- XGBoost consistently outperformed others across all metrics.
- Random Forest was a strong baseline with high recall, essential in minimizing false negatives.
- KNN performed reasonably well but lagged behind in ROC-AUC.

VII. Discussion

The results confirm that ensemble-based methods are particularly effective in fraud detection due to their ability to model complex patterns. SMOTE significantly improved model learning by providing better representation for the minority class. However, over-sampling may introduce noise if not handled carefully.

VIII. Challenges

- Severe data imbalance required careful resampling.
- KNN's performance was sensitive to feature scaling and distance metrics.
- Risk of overfitting with complex models like XGBoost on small synthetic data.

IX. Future Scope

- Incorporating **real-time streaming** with fraud alerts using APIs.
- Applying **Autoencoders** or **Isolation Forests** for unsupervised anomaly detection.
- Using **Graph Neural Networks** to detect fraud patterns across customer networks.
- Integrating sentiment analysis from user reviews or social media for additional context.

X. Conclusion

Machine learning algorithms, particularly ensemble models, show strong promise in identifying bank fraud with high precision and recall. With proper handling of data imbalance and feature engineering, such systems can enhance the security and integrity of digital financial transactions. Future integration with real-time data and unsupervised models could further elevate fraud detection systems.

References

[1] A. Dal Pozzolo, O. Caelen, Y. Le Borgne, S. Waterschoot, and G. Bontempi, "Calibrating Probability with Undersampling for Unbalanced Classification," in *2015 IEEE Symposium Series on Computational Intelligence*, 2015.

[2] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Cost-sensitive credit card fraud detection using Bayes minimum risk," in *2016 14th International Conference on Machine Learning and Applications (ICMLA)*, pp. 292–298, 2016.

[3] F. Carcillo, Y. Le Borgne, O. Caelen, and G. Bontempi, "Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection," *Information Sciences*, vol. 557, pp. 317–331, 2021.

[4N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.