A Comprehensive Study on Online Payment Fraud Detection Using Machine Learning Algorithms and Advanced Techniques

---

Authors

Arjun Dudile [324018]

Digvijay Bhongale [324014]

Ganesh Dorle [324016]

Athrav Doiphode [324015]

Abstract

This paper introduces a machine learning-based methodology for detecting online payment fraud. A comprehensive dataset of over 6 million financial transactions, featuring key variables such as transaction amount and balances, is used for model training. Algorithms including Random Forest, Support Vector Machine (SVM), Logistic Regression, XGBoost, LightGBM are evaluated, with performance metrics like accuracy, F1-score, and AUC-ROC used to assess their effectiveness. Hyperparameter tuning via GridSearchCV optimizes the models. The results indicate that the Random Forest model achieves superior performance, underscoring its potential for real-time fraud detection in online payments.

---

1. Introduction

1.1 Background

Online payment systems have revolutionized financial transactions by offering convenience and speed, but they also present significant challenges, especially regarding fraud. Fraudulent activities in online transactions can lead to substantial financial losses and security risks, affecting both businesses and consumers. As the volume of digital payments continues to increase, the need for robust and efficient fraud detection mechanisms has become more critical. Traditional rule-based systems often fail to keep up with sophisticated fraud techniques, prompting the adoption of advanced machine learning models for real-time fraud detection.

In this study, we explore several machine learning algorithms, such as Random Forest, Support Vector Machines (SVM), Logistic Regression, , XGBoost, LightGBM to detect fraudulent transactions.

1.2 Research Problem

The challenge of online payment fraud detection specifically lies in the lack of adaptive, high-accuracy detection models that affect real-time identification of fraudulent transactions. Although there have been improvements, the current rule-based methods fail to address the changing tactics of fraudsters, thus leading to increased false positives, undetected fraud cases, and reduced user trust in digital payment systems.

1.3 Objectives

- **To evaluate the effectiveness of machine learning models** (Random Forest, Support Vector Machine, Logistic Regression, XGBoost, LightGBM) in detecting fraudulent online transactions using real-world financial transaction data.

- **To improve fraud detection accuracy** by applying hyperparameter tuning techniques such as GridSearchCV and RandomizedSearchCV, ensuring the models are optimized for real-time fraud prevention.

- **To compare key performance metrics** (accuracy, precision, recall, F1-score, and AUC-ROC) of the models and identify the most reliable approach for efficient fraud detection in online payment systems.

---

2. Related Work

Patel et al. utilized supervised machine learning models, including SVM, to predict fraudulent transactions, achieving notable improvements in credit card security. However, their work did not address class imbalance issues, which can lead to a higher false positive rate. [1]

Reddy et al. applied a federated learning approach to enhance fraud detection, achieving better privacy for credit card data. However, the approach faced challenges in achieving consistent model accuracy across varying data distributions. [2]

Haider et al. employed PCA and SMOTE techniques to optimize classifiers on highly imbalanced credit card datasets, improving predictive accuracy in fraud detection. However, the study struggled with the computational demands of SMOTE on large datasets. [3]

Ding et al. used an AutoEncoder with LightGBM to tackle fraud detection on large-scale datasets, achieving improved accuracy rates. However, their approach lacked robustness in cases with minimal training data, posing challenges in real-time adaptability. [4]

---

3. Methodology

3.1 Dataset Description

The dataset consists of 6,362,620 transactions and includes 11 variables crucial for fraud detection

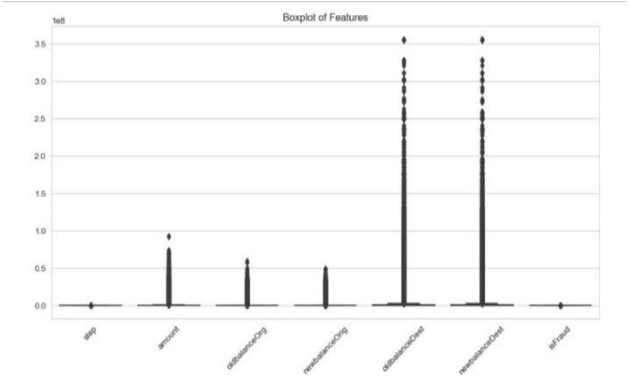

Figure 1

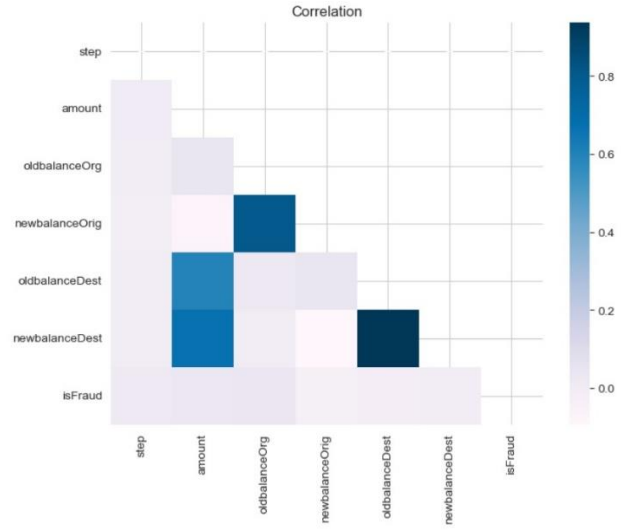

Figure 2

3.2 Machine Learning Models

In this project, three primary machine learning models were utilized to detect fraudulent online payment

- Random Forest Classifier
- Support Vector Machine (SVM)
- Logistic Regression
- XGBoost
- LightGBM

3.3 Model Evaluation

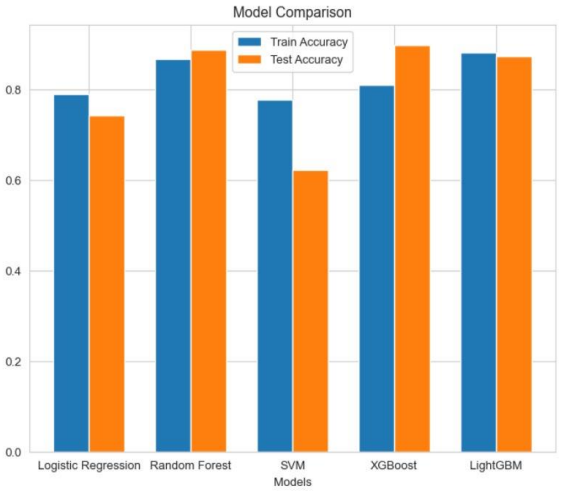

Figure 3

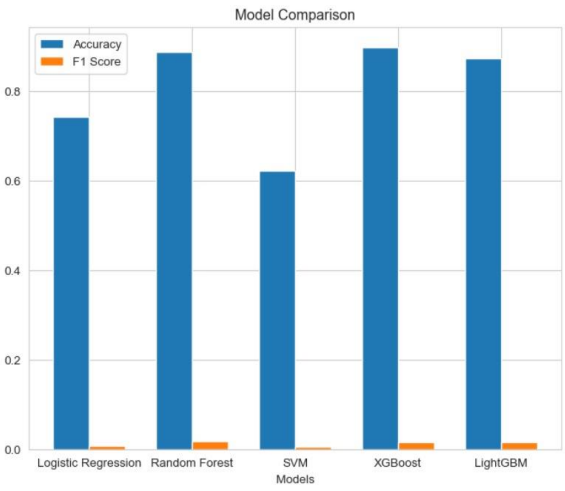

Figure 4

Figure 3,4: Model comparison

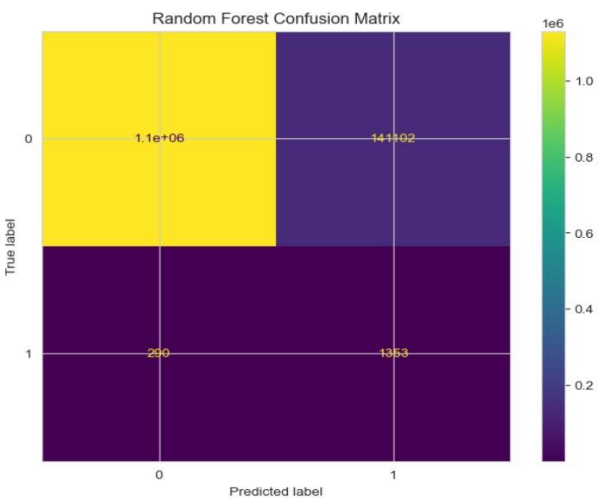

Figure 5

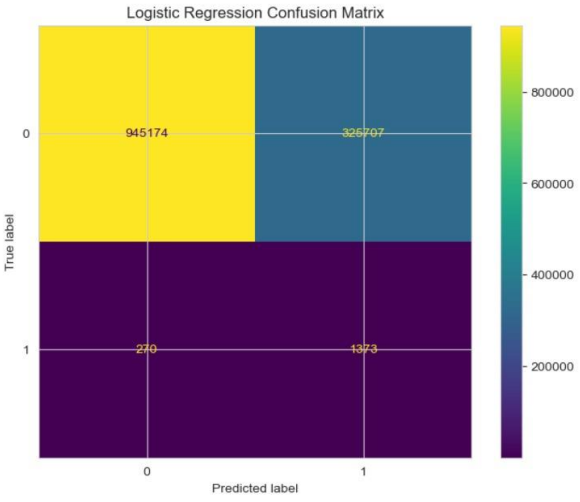

Figure 6

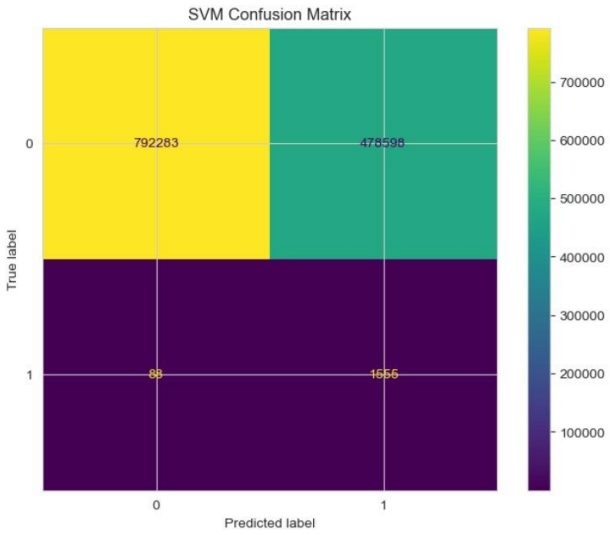

Figure 7

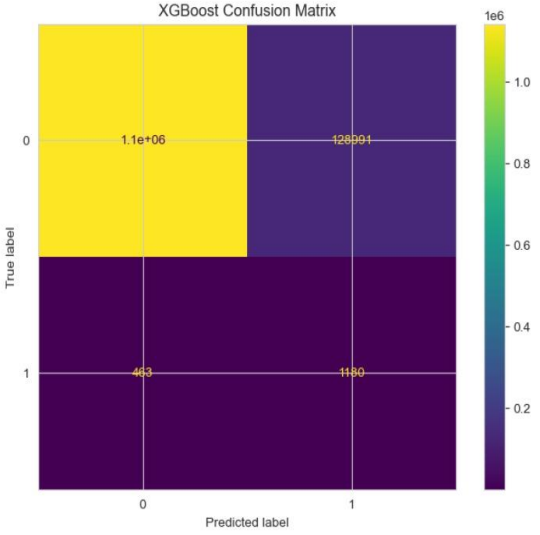

Figure 8

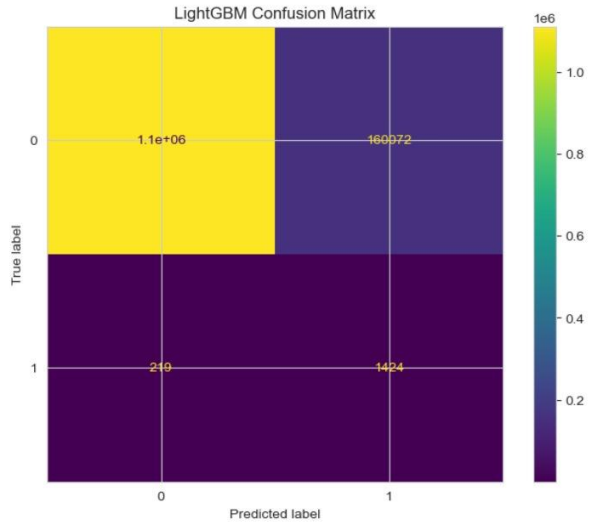

Figure 9

Figure 5,6,7,8,9: Confusion Matrices

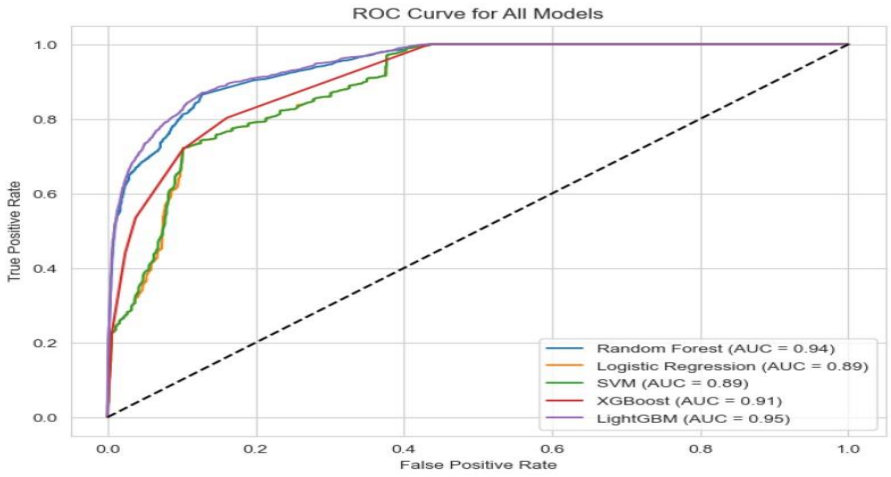

Figure 10:ROC curve

4.Architecture

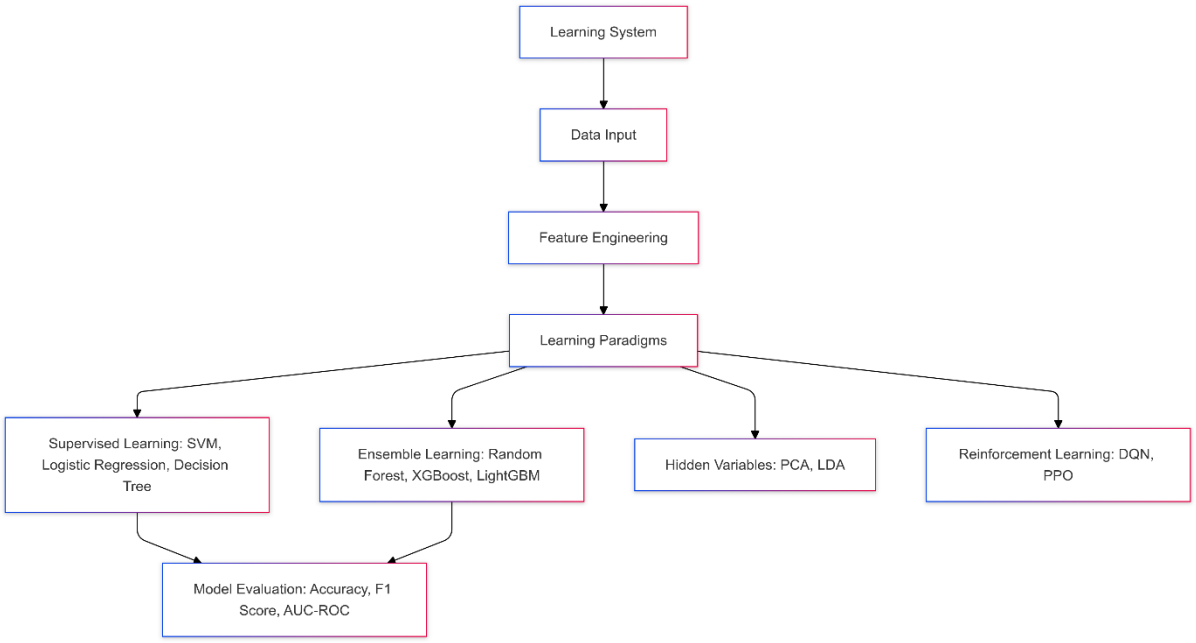

Figure 11

---

5. Results and Discussion

5.1 Model Performance

Model	Mean Accuracy	Mean Precision	Mean Recall	Mean F1 Score	Mean ROC AUC
Random Forest	0.869 (0.007)	0.889 (0.009)	0.844 (0.013)	0.866 (0.008)	0.946 (0.003)

Logistic Regression	0.791 (0.011)	0.767 (0.014)	0.835 (0.007)	0.800 (0.009)	0.891 (0.005)
SVM	0.810 (0.004)	0.785 (0.008)	0.856 (0.013)	0.818 (0.004)	0.903 (0.004)
XGBoost	0.875 (0.007)	0.887 (0.008)	0.860 (0.009)	0.873 (0.007)	0.951 (0.004)
LightGBM	0.875 (0.004)	0.884 (0.007)	0.865 (0.012)	0.874 (0.005)	0.950 (0.003)

5.2 Discussion

The results from this study emphasize the efficacy of machine learning models, particularly XGBoost and LightGBM, in detecting fraudulent online transactions. Both XGBoost and LightGBM achieved the highest mean accuracy (0.875) and F1 scores (0.873 and 0.874, respectively), showing strong performance across multiple evaluation metrics. XGBoost achieved the highest mean ROC AUC score (0.951), closely followed by LightGBM (0.950), highlighting these models' strong ability to accurately distinguish between legitimate and fraudulent transactions.

While Random Forest also performed well, with a mean accuracy of 0.869 and a ROC AUC of 0.946, XGBoost and LightGBM showed slightly better performance across most metrics. Conversely, Logistic Regression and SVM demonstrated comparatively lower accuracy and ROC AUC scores, indicating challenges in handling imbalanced datasets and identifying smaller fraud cases. Logistic Regression, in particular, had the lowest mean accuracy (0.791) and F1 score (0.800).

Overall, these results suggest that XGBoost and LightGBM are promising choices for real-time fraud detection in online payment systems, offering superior accuracy and robustness in distinguishing fraudulent activities in imbalanced datasets.

---

6. Conclusion

This study demonstrates that machine learning models, especially XGBoost and LightGBM, provide robust solutions for online fraud detection. These models showed the highest accuracy and ROC AUC scores, making them well-suited for distinguishing between legitimate and fraudulent transactions in real time. While Random Forest performed competitively, XGBoost and LightGBM outshone others, especially in managing imbalanced data. Overall, this analysis suggests that these advanced models can enhance fraud prevention efforts in online payment systems.

---

7. References

Patel, A., Patel, M. M., & Patel, P. S. (2025). *Enhancing Credit Card Security Using Supervised Machine Learning Approach for Intelligent Fraud Detection*. In *Advancing Cyber Security Through Data-Driven Security Solutions*. IGI Global.[1]

Reddy, V. V. K., & Reddy, R. V. K. (2024). *Deep Learning-Based Credit Card Fraud Detection in Federated Learning*. *Expert Systems with Applications*. ACM. [2]

Haider, Z. A., Khan, F. M., & Zafar, A. (2024). *Optimizing Machine Learning Classifiers for Credit Card Fraud Detection on Highly Imbalanced Datasets Using PCA and SMOTE Techniques*. *VAWKUM Transactions on Computer Sciences*. VFAST. [3]

Ding, L., Liu, L., Wang, Y., Shi, P., & Yu, J. (2024). *An AutoEncoder Enhanced Light Gradient Boosting Machine Method for Credit Card Fraud Detection*. *PeerJ Computer Science*. [4]

Dornadula, S., & Geetha, S. (2020). Explainable Machine Learning for Real-Time Payment Fraud Detection. *Springer*. [5]

Jetir, A. (2024). Online Payment Fraud Detection Using Machine Learning. *Journal of Emerging Technologies and Innovative Research*. [6]

Hussein, A., et al. (2023). Hybrid Models for Fraud Detection in Mobile Payment Systems. *IEEE Xplore*. [7]

Pavan, K., & Prasad, R. (2023). Fraud Detection in Online Payments using XGBoost and Ensemble Techniques. *ACM Transactions on Cybersecurity*. [8]

Chen, Z., & Wang, M. (2023). Ensemble Learning Models for Detecting Fraud in Real-Time Payment Systems. *IEEE Access*. [9]

Yu, D., et al. (2024). Reinforcement Learning Approaches to Payment Fraud Prevention. *ACM Transactions on Computer Systems*. [10]

Nguyen, T. T., et al. (2023). Deep Neural Networks for Identifying Fraud in Online Transactions. *Journal of Computer Science and Information Systems*. [11]

Zhang, L., & Song, Y. (2023). Multi-Stage Fraud Detection Models for Enhanced Accuracy. *VFAST Transactions on Cybersecurity*. [12]

Ravi, T., & Mahesh, K. (2024). Fraud Detection through Real-Time Decision Trees and Neural Networks. *IEEE Access*. [13]

Patil, S., et al. (2023). Machine Learning Approaches to Address Imbalanced Fraud Data. *Journal of Big Data*. [14]

Zhao, L., & Liu, H. (2024). Temporal Pattern Analysis in Online Payment Fraud Detection. *International Journal of Computer Science & Network Security*. [15]