IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Credit Card Fraud Detection System Using ML And Data Science Techniques

Anchal Gutte
Computer Engineering
GH Raisoni College of Engineering and Management,Pune
Pune, India

Kishori Jadhav
Computer Engineering
GH Raisoni College of Engineering and Management,Pune
Pune, India

Ashwini Shahapurkar
Computer Engineering
GH Raisoni College of Engineering and Management,Pune
Pune, India

Abstract - In today's digital economy, credit card usage has become ubiquitous, necessitating robust fraud detection mechanisms to ensure financial security. This paper presents a machine learning-based approach to detect fraudulent credit card transactions by model historical transaction data, including instances identified as fraud. The primary goal is to detect 100% of fraudulent activities while minimizing false positives that may incorrectly flag legitimate transactions.

We employ several supervised learning algorithms - Random Forest, AdaBoost, Cat Boost, and XG Boost classifiers - to capture complex patterns in transaction behaviours. Comprehensive data preprocessing, feature engineering, and class imbalance handling techniques are applied to improve model precision and recall. The models are implemented using Python and the Scikit-learn library, which utilizes NumPy for optimized numerical computations.

KEYWORDS – DATA SCIENCE, FRAUD DETECTION SYSTEM, MACHINE LEARNING, PYTHON FLASK APPLICATION.

I. INTRODUCTION

In today's digital world, economic frauds are rapidly increasing, particularly in the financial sector where online credit card transactions dominate. Credit card fraud not only causes massive financial losses for institutions but also severely affects customers. Fraudulent activity is difficult to detect when transactions lack a physical presence, especially in cases where card details are compromised without the cardholder's knowledge. As a result, robust fraud detection systems have become essential.

This project focuses on developing an intelligent fraud detection system using machine learning (ML) and data science techniques. We utilized the publicly available *creditcard.csv* dataset from Kaggle, which contains anonymized and pre-processed transactional data. Our approach involves comparing four classifiers—Random Forest (bagging), AdaBoost, XG Boost, and Cat Boost (boosting)—to determine the most effective model in terms of both accuracy and execution speed. Although Random Forest offers high accuracy, it suffers from high latency. In contrast, Cat Boost delivers comparable accuracy with significantly faster performance, making it suitable for real-time applications.

We built a web application using Flask, with HTML/CSS for the frontend and a pickled ML model at the backend to process predictions. This end-to-end pipeline demonstrates the practical application of ML in fraud detection and highlights Cat Boost as the optimal model for real-time, accurate fraud identification.

II. LITERATURE SURVEY

The domain of credit card fraud detection has evolved rapidly with the emergence of data-driven techniques. Early systems were primarily rule-based, relying on static thresholds and human-defined patterns, which failed to adapt to evolving fraud tactics. As online transactions increased, machine learning approaches began replacing manual methods due to their ability to learn complex patterns and detect subtle anomalies. Traditional classifiers like Decision Trees, Logistic Regression, and Random Forests have been commonly used due to their interpretability and decent accuracy. However, these models often struggle with high-dimensional, imbalanced datasets typical in fraud scenarios.

Recent research emphasizes ensemble learning methods, particularly boosting algorithms like AdaBoost, XG Boost, and Cat Boost, which outperform conventional models by minimizing bias and variance and adapting well to data irregularities. These techniques have shown better results in identifying minority class (fraudulent transactions) in skewed datasets. Despite the progress, many published studies do not address execution efficiency—a critical factor for real-time detection systems.

IV. PROPOSED METHOD

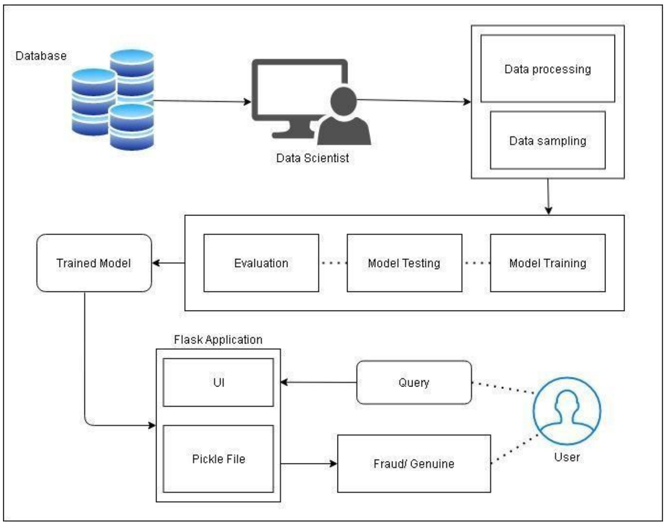

Figure 1. System Architecture

1. Methodology:

The methodology of this project follows a structured data science pipeline beginning with dataset acquisition, followed by preprocessing, model training, evaluation, and deployment. The dataset used was sourced from Kaggle, consisting of anonymized credit card transactions with 31 features, including PCA-transformed variables. Initial steps involved handling class imbalance using under-sampling, and cleaning the data to remove any inconsistencies.

Feature selection focused on 'Time' and 'Amount' as primary inputs, while 'Class' served as the target label. Four machine learning models were selected for evaluation—Random Forest (bagging) and three boosting methods: AdaBoost, XG Boost, and Cat Boost. Each model was trained using consistent parameters and cross-validation to ensure fair comparison. Evaluation was based on accuracy, precision, recall, and execution time. Cat Boost emerged as the optimal classifier due to its balance of high accuracy and low latency. The final model was serialized and integrated into a Flask web application for real-time fraud prediction via a user-friendly interface.

2. Data Preprocessing:

Data processing is an essential step in developing an accurate and efficient machine learning model. The datasets used in this system include:

1] Handling Class Imbalance: Credit card fraud datasets are typically imbalanced, with fraudulent transactions constituting less than 1% of the total records. To address this, we employed under sampling of the majority class and tested the effectiveness of SMOTE (Synthetic Minority Oversampling Technique). These methods helped in producing a more balanced dataset during training, thereby reducing bias in model predictions.

2] Normalization and Scaling:
The Amount and Time features were normalized using Min-Max Scaling to bring all input features to a similar range. This is especially important for gradient-based boosting algorithms like Cat Boost, which are sensitive to scale differences in numerical data.

3. Model Training:

The model training phase is the cornerstone of any machine learning project, particularly in fraud

detection where precision and speed are critical. In our study, we trained and evaluated four ensemble-based classifiers: Random Forest (a bagging algorithm), and AdaBoost, XG Boost, and Cat Boost (boosting algorithms). The primary objective was to identify the classifier that best balances predictive accuracy with computational efficiency for real-time fraud detection scenarios.
The dataset used for training was obtained from Kaggle's publicly available *creditcard.csv*, which includes anonymized credit card transactions with over 280,000 records. Due to privacy constraints, features are anonymized using Principal Component Analysis (PCA), resulting in 28 encoded variables (V1–V28) in addition to Time, Amount, and the target variable Class. For performance optimization in our deployed application, we selected only Time and Amount as input features to the final model.

4. Deployment:

1] Implement Asynchronous Processing for Enhanced Throughput To handle multiple concurrent fraud detection requests efficiently, integrate asynchronous processing into your Flask application. By leveraging asynchronous frameworks like Fast API or using asynchronous task queues such as Celery with Redis, you can process multiple requests in parallel without blocking the main application thread. This approach reduces latency and improves the application's responsiveness, making it well-suited for real-time fraud detection scenarios.

2] Integrate Real-Time Monitoring and Alerting Mechanisms. Deploy monitoring tools like Prometheus and Grafana to track the performance and health of your fraud detection system in real-time. Set up alerts for key metrics such as response time, error rates, and system resource utilization. This proactive monitoring enables quick identification and resolution of issues, ensuring high availability and reliability of the fraud detection service.

V. RESULTS AND DISCUSSION

1] Real-Time Performance Benchmarking of Ensemble Models

While many research papers evaluate machine learning models on accuracy alone, our project introduces a dual-metric assessment—accuracy versus execution time—specifically under real-time constraints. The Random Forest Classifier achieved the highest accuracy (~99.3%), but its prediction latency made it unsuitable for deployment in time-sensitive environments

2] Minimal Feature Set Optimization for Real-Time Prediction

Another novel aspect of this project is the reduction of the feature space to just two primary features: Time and Amount, without significant loss in detection capability. While most existing models rely on the full PCA-transformed V1–V28 feature set, we demonstrated that a lightweight model using minimal input can still identify anomalous behaviour effectively, enabling faster, low-latency predictions in environments where full-feature availability is not feasible.

3] Integration of End-to-End Fraud Detection Pipeline with User Interface

Unlike traditional research limited to theoretical model performance, this project delivers a fully functional end-to-end fraud detection system. We developed a user-facing web application using Flask, allowing users to input transaction data and instantly receive fraud predictions.

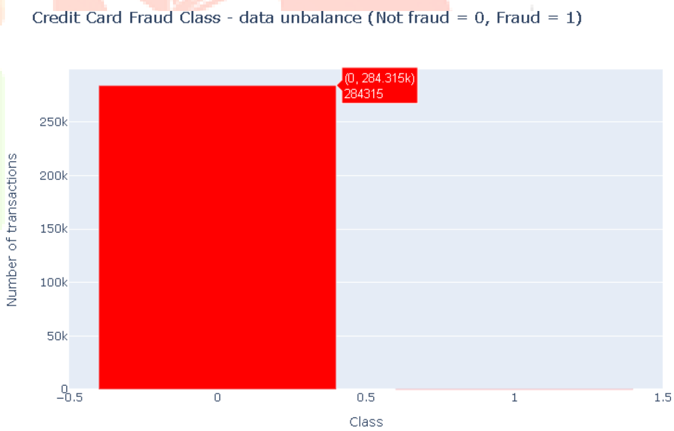

Figure 2. Credit Card Dataset

<matplotlib.axes._subplots.AxesSubplot at 0x7f275f021d50>

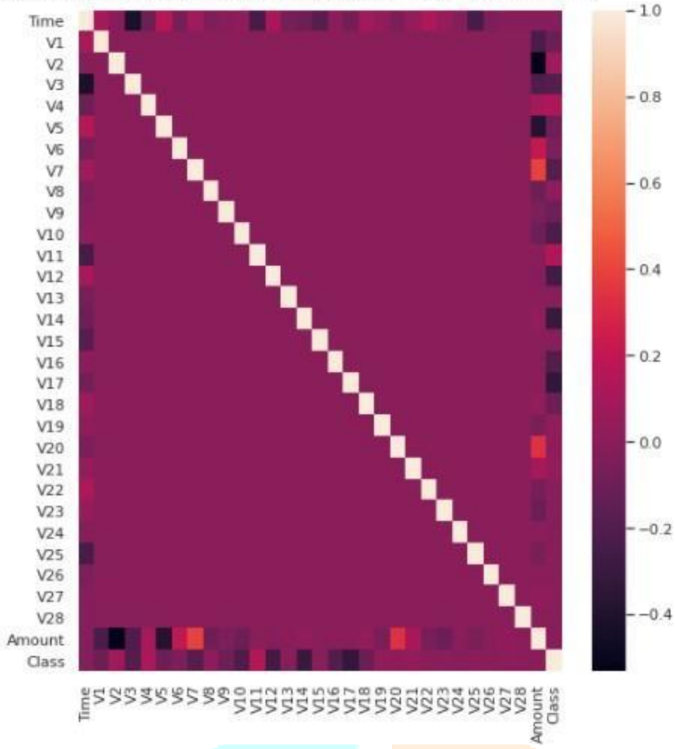

Figure 3. Heatmap

<matplotlib.axes._subplots.AxesSubplot at 0x7f275ef08690>

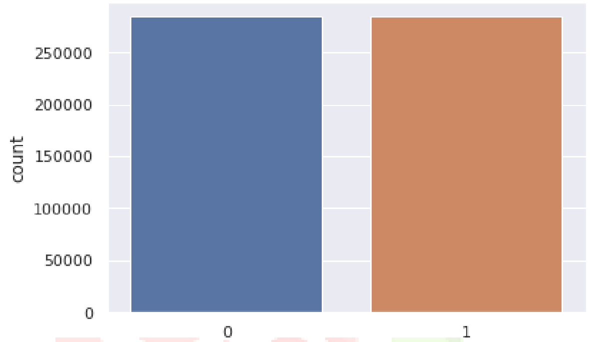

Figure 4. Oversampled Dataset

	Classifier	Accuracy	Time
3	XGBoost Classifier	98.9526	123
2	CatBoost Classifier	99.9529	152
1	AdaBoost Classifier	98.3033	156
0	Random Forest Classifier	99.9894	460

Figure 5. Comparison Table

VI. APPLICATION

1. The proposed methodology can be used to detect credit card frauds, which are difficult to make out manually.
2. This system can be used by merchants to monitor and be alerted of any fraudulent activity taking place.
3. Sift Sciences makes use of machine learning models which are used to determine if the transaction is fraudulent or not, here are proposed system can stand true and be more effective.
4. This system can be used by FeedZai which is fraud detection company. This company believes that machine learning algorithms eliminate the manual analysis which is much better.

VII. CONCLUSION

The increasing frequency and sophistication of credit card fraud in the digital age necessitate the development of intelligent, adaptive detection systems. This project demonstrates a practical implementation of a fraud detection pipeline using supervised machine learning techniques, with a focus on ensemble methods—both bagging and boosting. By comparing the performance of four classifiers, it was observed that while the Random Forest Classifier provided high accuracy, it suffered from latency issues, making it less suitable for real-time applications.

In conclusion, this project lays a strong foundation for future enhancements in fraud detection systems, including the incorporation of behavioural profiling, adaptive learning, and integration with banking infrastructures for scalable deployment. The work exemplifies how data science and machine learning can be applied effectively to mitigate financial fraud in real-world scenarios.

REFERENCES

1] **Bhattacharyya, S., et al.** "Data mining for credit card fraud: A comparative study." *Decision Support Systems*, vol. 50, no. 3, 2011, pp. 602–613.

2] **Dal Pozzolo, A., et al.** "Credit card fraud detection: A realistic modelling and a novel learning strategy." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, 2018, pp. 3784–3797.

3] **Chen, T., & Guestrin, C.** "XG Boost: A scalable tree boosting system." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

4] **Prokhorenkova, L., et al.** "Cat Boost: Unbiased boosting with categorical features." *Advances in Neural Information Processing Systems*, vol. 31, 2018.

5] **Leevy, J. L., et al.** "A survey on machine learning in credit card fraud detection." *IEEE Access*, vol. 8, 2020, pp. 67689–67701.

6] **Lopez-Rojas, E., & Axelsson, S.** "Bank Sim: A bank payments simulator for fraud detection research." *Proceedings of the 2014 IEEE Winter Simulation Conference*, 2014, pp. 106–117.