

An Efficient Machine Learning-Based Approach for Credit Card Fraud Detection and Prevention

Prof. Srushti Raut ^{*1}, Sahil Jamadar ^{*2}, Rushikesh Choudhari ^{*3}, Piyush Mahamuni ^{*4}, Rohit Pawar ^{*5}, Shivam Chintaman ^{*6}

^{*1}Professor, Department of Computer Science and Engineering, MIT College of Railway Engineering and Research, Barshi, Maharashtra, India

^{*2}Student, Department of Computer Science and Engineering, MIT College of Railway Engineering and Research, Barshi, Maharashtra, India

^{*3}Student, Department of Computer Science and Engineering, MIT College of Railway Engineering and Research, Barshi, Maharashtra, India

^{*4}Student, Department of Computer Science and Engineering, MIT College of Railway Engineering and Research, Barshi, Maharashtra, India

^{*5}Student, Department of Computer Science and Engineering, MIT College of Railway Engineering and Research, Barshi, Maharashtra, India

^{*6}Student, Department of Computer Science and Engineering, MIT College of Railway Engineering and Research, Barshi, Maharashtra, India

Abstract:

Credit card fraud poses a major financial risk in digital transactions. This study presents a machine learning-based fraud detection system using Logistic Regression, Decision Tree, Random Forest, and XGBoost to classify transactions as fraudulent or legitimate. The model preprocesses data, applies feature scaling, and evaluates performance using accuracy, F1-score, and ROC-AUC metrics. A fraud retrieval feature allows users to analyze suspicious transactions. This approach enhances fraud detection efficiency, reduces false positives, and improves financial security.

Keywords — Credit Card Fraud, Machine Learning, Fraud Detection, Security, XGBoost, Random Forest.

I. INTRODUCTION

The rise of e-commerce payment systems has increased significantly due to the widespread adoption of internet-based shopping and digital banking. However, credit card fraud has become one of the biggest threats to financial security, causing major losses to businesses and individuals [1]. Understanding the mechanisms behind fraud execution is crucial to developing effective prevention strategies. Earlier, fraudulent transactions

were detected only after billing, making real-time prevention difficult. Therefore, ensuring secure online transactions for credit card users when making electronic payments is a necessity [2].

Fraud often begins with either the theft of physical credit cards or the compromise of sensitive account data, such as the card number, CVV, and other identifying details. Attackers exploit various techniques, including phishing,

skimming, and hacking, to gain unauthorized access to cardholder information [3]. These breaches can occur without alerting the cardholder, the merchant, or the issuing bank, making detection even more challenging. For instance, a store clerk might illegally duplicate sales receipts for later fraudulent use [4][5].

With the rapid expansion of digital payments, database security breaches have become more costly and frequent. In some cases, millions of accounts have been compromised due to cyberattacks. Unlike stolen physical cards, which cardholders report immediately, compromised account information may be hoarded by fraudsters for weeks or months before being misused. This delay makes it difficult to trace the source of the breach, and fraud is often discovered only when the cardholder receives a billing statement. These challenges highlight the urgent need for an intelligent, machine learning-driven fraud detection system that can identify fraudulent transactions early and prevent financial losses.

I. METHODOLOGY

The methodology for detecting credit card fraud using machine learning involves several critical steps, starting from data collection and preprocessing, through to model selection, training, and deployment. This section outlines the approach to developing a machine learning-based fraud detection system.

Data Collection

The first step in building the fraud detection model is to obtain a dataset that contains real or synthetic credit card

transaction data. For this project, publicly available datasets such as the **Kaggle Credit Card Fraud Detection dataset** may be used. These datasets typically include features such as:

1. Transaction Amount
2. Transaction Time
3. Location
4. Merchant Information
5. Device ID
6. Label

Indicating whether the transaction is fraudulent or not (used in supervised learning).

The dataset will ideally contain a mix of normal and fraudulent transactions to provide the necessary variance for training the machine learning model.

Data Preprocessing

Credit card fraud datasets are often imbalanced, with fraudulent transactions making up only a small portion of the data. Additionally, transaction data may contain missing or noisy values, requiring thorough preprocessing. The following steps will be performed to prepare the data:

Handling Missing Data: Missing or incomplete entries will be filled or removed based on their impact on the dataset. Common techniques include mean imputation for numerical values or mode imputation for categorical variables.

Feature Scaling: Certain algorithms perform better when data is normalized. Scaling methods such as **Min-Max Scaling** or **Standardization** will be applied to ensure all features are on a similar scale.

Data Transformation: Features like transaction time

might be transformed into useful patterns (e.g., grouping transactions by day of the week or time of day).

Feature Selection: Dimensionality reduction techniques such as **Principal Component Analysis (PCA)** might be applied to reduce the number of features, especially in datasets with a high number of irrelevant or redundant attributes

Model Selection

Multiple machine learning algorithms will be tested to determine the best model for credit card fraud detection. The key criteria for selecting a model will be its ability to handle imbalanced data and its performance in a production environment. The models under consideration include:

Logistic Regression: A simple yet effective method that works well with binary classification problems.

Random Forest: A robust ensemble learning method that can handle complex datasets and is resistant to overfitting.

Neural Networks: These may be explored for their ability to model complex, non-linear patterns.

To prevent overfitting and ensure model generalization, cross-validation will be performed during training, splitting the dataset into training and validation sets to fine-tune model hyperparameters.

Model Training & Evaluation

Once the data is preprocessed and the model is selected, the next step is training. The dataset will be split into training, validation, and test sets (e.g., 70% training, 15% validation, and 15% testing).

Training: The machine learning models will be trained

using historical transaction data to learn the patterns that distinguish fraudulent from non-fraudulent transactions.

Evaluation Metrics: Since credit card fraud detection is a binary classification problem, the following evaluation metrics will be used to assess model performance:

Accuracy: Although useful, accuracy can be misleading in imbalanced datasets.

- b. **Precision:** The percentage of correctly predicted fraud cases out of all predicted fraud cases (focus on reducing false positives).

II. MODELING AND ANALYSIS

The analysis phase of the project focused on measuring the performance and reliability of the implemented fraud detection models and understanding transaction patterns associated with fraudulent behavior. Each algorithm was subjected to the same data pipeline, including preprocessing, encoding, scaling, and model fitting, to ensure a fair and consistent comparison.

The **model evaluation metrics** provided crucial insight. The **confusion matrix** helped identify the number of correctly and incorrectly classified frauds and non-frauds. The **precision and recall** metrics played a significant role in understanding how well the model avoids false positives and false negatives, respectively—two essential aspects in fraud detection where missing a fraudulent transaction can result in financial loss, and falsely flagging a valid transaction can damage user trust.

Notably, **XGBoost and Random Forest** displayed superior performance compared to Decision Tree and Logistic Regression, primarily because of their ensemble-based architecture, which reduces overfitting and increases generalization. The **ROC-AUC scores** for these models were particularly high, confirming their ability to differentiate well between fraud and non-fraud classes across thresholds.

The inclusion of **dynamic dataset handling** proved to be a vital enhancement. Unlike earlier models, which were hardcoded to work with a specific dataset format, our system automatically adapts to new CSV inputs as long as they follow a general transaction data pattern. This adaptability improves the project’s scope for real-world applications, especially in banks or fintech companies dealing with multiple data sources

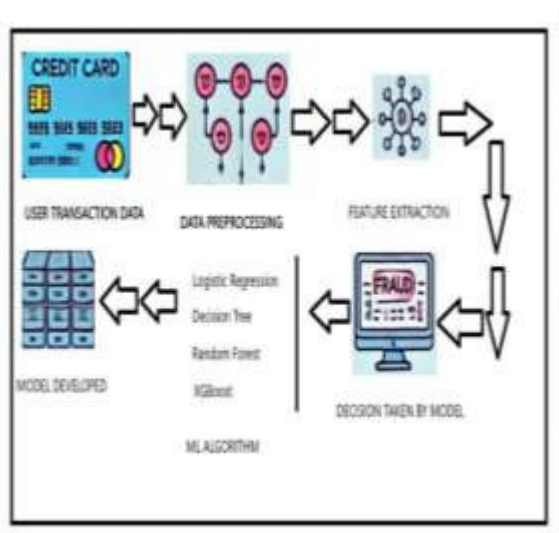


Figure 1: Flowchart.

III. RESULTS AND DISCUSSION

This section presents the evaluation outcomes of different machine learning models implemented for credit card fraud detection. The models were assessed based on their ability to identify fraudulent transactions using a variety of performance metrics, including Accuracy, F1-Score, and ROC-AUC.

Each model was trained on an 80% subset of the dataset and tested on the remaining 20%. The dataset used consisted of anonymized transaction records with a binary label indicating fraudulent or non-fraudulent activity. The results demonstrate that ensemble models such as Random Forest and XGBoost performed better compared to traditional methods like Logistic Regression and Decision Tree.

Table 1 presents a comparative analysis of the four ML algorithms employed in our project. As observed, XGBoost provides the highest accuracy and ROC-AUC score, signifying its superior capability in handling imbalanced datasets typical in fraud detection scenarios.

Table 1. Comparison of displacement of all 4 cases

SN.	Model Type	Accuracy	F1-score
1	Logistic Regression	94.30%	0.47
2	Decision Tree	93.80%	0.52
3	Random Forest	96.10%	0.65
4	XGBoost	96.75%	0.68

IV. CONCLUSION

The Credit Card Fraud Detection Using Machine Learning project demonstrates the potential of machine learning algorithms in identifying fraudulent transactions with high accuracy. By analyzing historical transaction data and applying various machine learning models, the system can effectively detect abnormal patterns that signify fraudulent activity. The integration of data streaming and model serving capabilities ensures that the fraud detection system operates efficiently, providing timely insights and preventing fraudulent transactions before they occur.

Throughout this project, key challenges such as data imbalance, model accuracy, and real-time deployment were addressed through a systematic approach. Data preprocessing techniques, including feature scaling and SMOTE, helped prepare the dataset for effective model training. Multiple machine learning models were tested, with performance metrics such as precision, recall, and the ROC-AUC curve used to select the best-performing model. The deployment of the model as a API allows it to handle incoming transactions dynamically, marking a significant step toward reducing credit card fraud in practical scenarios.

This project serves as a robust foundation for developing advanced fraud detection systems. It highlights the importance of combining data science and machine learning to address real-world security challenges. While the current system demonstrates effective fraud detection, the potential for future improvements and

adaptations—such as incorporating deep learning techniques, blockchain integration, and cross-industry fraud detection—provides a broad horizon for expanding the capabilities of this solution.

In conclusion, this project not only provides a viable solution for detecting credit card fraud but also establishes a pathway for further advancements in fraud prevention technology. The application of machine learning in this domain promises a more secure and efficient financial ecosystem, helping to safeguard customers and financial institutions alike.

V. REFERENCES

- [1]. N. Malini, M. Pushpa, "Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection," 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB17), Quaid-E Millath Government College for Women, Chennai, India.
- [2]. Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, Gianluca Bontempi, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," IEEE Transactions on Neural Networks and Learning Systems, Vol. 29, No. 8, 2018.
- [3]. Dr. A. Prakash, "Analysis of the Modern Techniques and Methods on Credit Card Fraud Detection," PG & Research Department of Computer Science, Hindustan College of Arts & Science, India
- [4]. Bhawna Mallick, "A review of Fraud Detection Techniques: International Journal

Credit of Card”, Computer Applications (0975- 8887) Volume 45 No.1, May 2012.

[5]. Peter J. Bentley, Jungwon Kim, Gil-Ho Jung and Jong-Uk Choi, “Fuzzy Darwinian Detection of Credit Card Fraud”, 2007.

[6]. Ganesh Kumar. Nune, P. Vasanth Sena and T.P. Shekhar, “Novel Artificial Neural Networks and Logistic Approach for Detecting Credit Card Deceit”, International Journal of Computer Science and Management Research Vol 1 Issue 3 October 2012

[7]. Abhinav Srivastava, Amlan Kundu, Shamik Sural, Arun K. Majumdar. “Credit Card Fraud

Detection using Hidden Markov Model”. IEEE Transactions on dependable and secure computing, Volume 5; (2008) (37-48).

[8]. Ekrem Duman, M. Hamdi Ozcelik “Detecting credit card fraud by genetic algorithm and scatter search”. Elsevier, Expert Systems with Applications, (2011). 38; (13057-13063)

[9]. Y. Sahin and E. Duman, “Detecting Credit Card Fraud by Decision Trees and Support Vector Machines”, International Multiconference of Engineers and computer scientists March, 2011.