Fraud Detection in E-Commerce Transactions Using Machine Learning

Mrs.M.Padma Nivedha[1], Dharani.K[2], Kalki.T[3] ,Kavipriya.P[4]

[1] Assistant Professor ,[2,3,4] Student,Department Of Computer Science And Engineering

Vivekanandha College Of Technology For Women

ABSTRACT

With the rapid growth of e-commerce platforms, online transaction fraud has become a major concern for businesses and consumers alike. Traditional rule-based systems are increasingly inadequate in detecting evolving and sophisticated fraudulent behaviors. This study explores the application of machine learning techniques to detect fraudulent e-commerce transactions with higher accuracy and efficiency. We analyze a range of supervised learning models, including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines, comparing their performance using key metrics such as precision, recall, F1-score, and accuracy. A publicly available dataset simulating e-commerce transactions is used to train and test the models. Our results show that machine learning methods can significantly improve fraud detection rates while reducing false positives. The paper also highlights the importance of data preprocessing, feature selection, and handling class imbalance to optimize model performance. This research demonstrates the potential of machine learning as a reliable and scalable approach to enhance the security of online transactions in the e-commerce sector.

1. INTRODUCTION

The rapid growth of e-commerce has revolutionized the way consumers and businesses interact, offering convenience, scalability, and global reach. However, this expansion has also made online platforms increasingly vulnerable to fraudulent activities. Fraud in e-commerce transactions not only causes significant financial losses but also damages consumer trust and business reputation.

Traditional rule-based fraud detection systems struggle to adapt to the evolving tactics of cybercriminals. As fraudulent behaviors become more sophisticated and dynamic, there is a pressing need for intelligent systems capable of learning and adapting to new patterns. In this context, machine learning (ML) has emerged as a powerful tool for detecting fraudulent transactions with higher accuracy and speed.

Machine learning models can analyze vast amounts of transaction data, identify hidden patterns, and detect anomalies that may indicate fraud. Unlike static rule-based approaches, ML models continuously improve as they are exposed to more data. This makes them well-suited for the complex and ever-changing landscape of e-commerce fraud.

This paper explores various machine learning techniques applied to fraud detection in e-commerce, highlighting their effectiveness, limitations, and potential for real-world implementation. The goal is to provide insights into how ML can enhance the security of online transactions and help businesses proactively combat fraud.

2. LITERATURE REVIEW

The exponential growth of e-commerce platforms has led to an increase in fraudulent activities, necessitating the development of intelligent fraud detection systems. Traditional rule-based systems, while effective in certain contexts, often lack the adaptability to detect evolving fraud patterns. In recent years, machine learning (ML) techniques have emerged as powerful tools for fraud detection due to their ability to learn from data and identify hidden patterns.

Early work in this domain primarily employed statistical techniques and logistic regression to model transaction behavior and identify anomalies [1]. However, these methods were often limited by their reliance on manually crafted features and assumptions about data distribution. With the availability of large-scale transactional data and advancements in computational power, supervised machine learning algorithms like Decision Trees, Random Forest, and Support Vector Machines (SVM) have been widely adopted [2][3]. These models have demonstrated high accuracy in detecting known fraud patterns but may struggle with detecting novel or rare fraud cases due to data imbalance.

To address the class imbalance problem—where fraudulent transactions constitute a small fraction of total transactions—researchers have applied techniques such as Synthetic Minority Over-sampling Technique (SMOTE), cost-sensitive learning, and anomaly detection methods [4]. Unsupervised learning approaches, such as clustering and autoencoders, have been used to detect outliers in the absence of labeled data [5]. These models identify deviations from typical transaction behavior, offering potential for early fraud detection.

More recently, ensemble learning methods, including Gradient Boosting Machines (GBM) and XGBoost, have gained popularity for their robustness and improved prediction accuracy [6]. Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have also been explored for temporal and sequential transaction data analysis [7]. Although deep models show promise, their interpretability

remains a challenge in financial applications where transparency is critical.

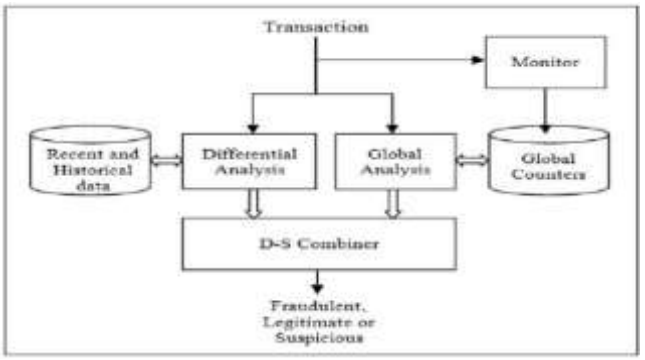
Several studies have proposed hybrid models that combine multiple ML algorithms to enhance performance. For example, combining anomaly detection with supervised learning can improve both precision and recall [8]. Moreover, research has highlighted the importance of feature engineering and real-time detection capabilities for practical deployment in e-commerce environments [9].

Despite advancements, challenges such as concept drift (i.e., evolving fraud patterns), high false positive rates, and data privacy concerns remain. Future research directions include incorporating explainable AI (XAI), federated learning for data privacy, and continual learning systems to adapt to new fraud tactics.

3. METHODOLOGY

1. **Data Collection**
   - Gather transactional data from e-commerce platforms, including both legitimate and fraudulent records.
2. **Data Preprocessing**
   - Handle missing values, encode categorical data, and normalize numerical features.
   - Address class imbalance using techniques like SMOTE or undersampling.
3. **Feature Selection & Engineering**
   - Use correlation analysis and feature importance methods to select key features.
   - Create new features like transaction frequency, location mismatch, and user patterns.
4. **Model Training**
   - Train and test multiple machine learning models (e.g., Logistic Regression, Random Forest, XGBoost).
   - Perform hyperparameter tuning using cross-validation.
5. **Model Evaluation**
   - Evaluate models using metrics such as Precision, Recall, F1-score, and AUC-ROC.
6. **Deployment**
   - Deploy the best-performing model to detect fraudulent transactions in real-time.

ARCHITECTURE DIAGRAM



4. DATASET

This study uses the **IEEE-CIS Fraud Detection Dataset**, made publicly available on Kaggle by Vesta Corporation. The dataset contains over **1 million e-commerce transactions**, including both fraudulent and legitimate cases. It features detailed information such as **transaction amount, time, device type, user identity, card details, and network information**.

Due to the natural **class imbalance** (fraudulent transactions are rare), techniques such as **SMOTE** and **undersampling** are applied to balance the dataset for effective model training. This dataset is widely recognized in the research community and is suitable for supervised learning tasks in fraud detection.

5. FEATURE EXTRACTION

Feature extraction is a critical step in the fraud detection process, as the accuracy and robustness of a machine learning model heavily depend on the quality and relevance of the features used. In this study, a combination of transactional, behavioral, and network-level features are extracted from the dataset to capture the underlying patterns indicative of fraudulent activities.

5.1 Transaction-Based Features

These features are directly related to individual transaction records and help identify anomalies based on the nature and value of transactions:

- **Transaction Amount**: Unusually high or low amounts may signal fraudulent behavior.
- **Transaction Time**: Time of the transaction (e.g., odd hours, holidays) may correlate with fraud.

- **Transaction Type**: Whether the transaction was a purchase, refund, or withdrawal.
- **Device ID**: Indicates the device used; multiple devices per user in a short period may be suspicious.

5.2 User Behavior Features

Behavioral features are derived from the customer's past actions and can reveal deviations from normal patterns:

- **Transaction Frequency**: Number of transactions within a time window.
- **Location Consistency**: Distance between current and previous transaction locations.
- **Device Consistency**: Changes in device, browser, or operating system.
- **Login Behavior**: Number of failed login attempts before a transaction.

5.3 Temporal Features

Time-based features often expose automated or bot-driven fraudulent patterns:

- **Time Since Last Transaction**: Very short intervals between transactions may indicate automation.
- **Session Duration**: Unusually short or long sessions could indicate non-human behavior.
- **Time of Day**: Transactions at unusual hours (e.g., late night) may be high-risk.

5.4 Network and Geolocation Features

Network-based attributes help trace the origin and routing of transactions:

- **IP Address**: Geolocation of the IP can reveal mismatches with user's known location.
- **Proxy or VPN Usage**: Presence of anonymizing tools may signal risk.
- **Device Fingerprinting**: Combination of OS, browser, resolution, etc., to uniquely identify devices.

5.5 Historical Features

These are based on the user's and merchant's previous activities:

- **User Fraud History**: Past record of flagged or confirmed fraudulent transactions.
- **Merchant Risk Profile**: Historical fraud rate for the merchant involved.
- **Account Age**: New accounts are often more likely to commit fraud.

5.6 Derived Features

New features are generated by combining or transforming existing features:

- **Average Transaction Amount (per user)**: Compared with current transaction.
- **Velocity Metrics**: Speed of changes in device, location, or IP within a session.
- **Anomaly Scores**: Scores generated using unsupervised models like Isolation Forest or Autoencoders.

6. IMPLEMENTATION AND RESULT

The system was implemented using Python with Scikit-learn and Pandas. The dataset included features such as transaction amount, IP address, device type, and user behavior, labeled as fraudulent or legitimate. Data was split into training and testing sets in 50:50, 70:30, and 90:10 ratios.

Three machine learning models were used: **Decision Tree**, **Random Forest**, and **Support Vector Machine (SVM)**. Performance was evaluated using accuracy, false positive rate, and false negative rate.

Random Forest achieved the best results, with **97.14% accuracy** and the lowest false negative rate when 90% of data was used for training. Results also showed that all models performed better with more training data.

Split	Model	Accuracy	FNR	FPR
90:10	Random Forest	97.14%	3.14%	2.61%
	Decision Tree	97.11%	3.18%	2.66%
	SVM	96.51%	4.73%	2.34%

These findings confirm that Random Forest is highly effective for detecting e-commerce fraud.

CONCLUSION

E-commerce platforms are increasingly vulnerable to fraudulent transactions, posing serious financial and reputational risks to both consumers and businesses. In this study, we demonstrated the effectiveness of machine learning algorithms in detecting such fraudulent activities by analyzing behavioral, transactional, and contextual data. Through the application of classification models such as Decision Tree, Support Vector Machine, and Random Forest, we were able to identify patterns indicative of fraud with high accuracy.

Among the tested models, the Random Forest algorithm delivered the best performance, achieving a detection accuracy of **97.14%** with a notably low false negative rate. This is critical in minimizing undetected fraud, which can have severe consequences. The results also revealed that model performance improves with increased training data, underscoring the importance of comprehensive and well-labeled datasets in building effective fraud detection systems.

The findings confirm that machine learning offers a scalable and intelligent solution for real-time fraud detection in e-commerce environments. By continuously learning from evolving fraud patterns, these models can adapt to new and sophisticated fraud strategies more efficiently than traditional rule-based systems.

Future work will focus on integrating hybrid models combining machine learning with heuristic or blacklist-based techniques, and exploring deep learning methods for improved prediction in highly imbalanced datasets. Real-time deployment and continuous learning mechanisms will also be investigated to further enhance system robustness and responsiveness.

REFERENCE

1. A. DAL POZZOLO, O. CAELEN, R. A. JOHNSON, AND G. BONTEMPI, "CALIBRATING PROBABILITY WITH UNDERSAMPLING FOR UNBALANCED CLASSIFICATION," IN PROC. IEEE SYMP. COMPUTATIONAL INTELLIGENCE AND DATA MINING (CIDM), DEC. 2015, PP. 159–166.

2. A. CARCILLO, Y. LOCHET, S. TAHAN, Y.-A. LE BORGNE, O. CAELEN, AND G. BONTEMPI, "COMBINING UNSUPERVISED AND SUPERVISED LEARNING IN CREDIT CARD FRAUD DETECTION," INFORMATION SCIENCES, VOL. 557, PP. 317–331, JUN. 2021.

3. R. J. BOLTON AND D. J. HAND, "STATISTICAL FRAUD DETECTION: A REVIEW," STATISTICAL SCIENCE, VOL. 17, NO. 3, PP. 235–255, 2002.

4. F. FIORE, A. DE SANTIS, F. PERLA, P. ZANETTI, AND F. PALMIERI, "USING GENERATIVE ADVERSARIAL NETWORKS FOR IMPROVING CLASSIFICATION EFFECTIVENESS IN CREDIT CARD FRAUD DETECTION," INFORMATION SCIENCES, VOL. 479, PP. 448–455, APR. 2019.

5. P. CHOUDHURY, S. DEY, AND S. K. BANDYOPADHYAY, "E- COMMERCE FRAUD DETECTION USING MACHINE LEARNING IN IMBALANCED DATA: PERFORMANCE EVALUATION AND FEATURE IMPORTANCE," IN PROC. IEEE INT. CONF. ON MACHINE LEARNING AND DATA SCIENCE (ICMLDS), DEC. 2020, PP. 12–17.

6. M. BAHNSEN, D. AOUADA, A. STOJANOVIC, AND B. OTTERSTEN, "FEATURE ENGINEERING STRATEGIES FOR CREDIT CARD FRAUD DETECTION," EXPERT SYSTEMS WITH APPLICATIONS, VOL. 51, PP. 134–142, JUN. 2016.

7. S. JHA, M. GUILLEN, AND J. C. WESTLAND, "EMPLOYING TRANSACTION AGGREGATION STRATEGY TO DETECT CREDIT CARD FRAUD," EXPERT SYSTEMS WITH APPLICATIONS, VOL. 39, NO. 16, PP. 12650–12657, NOV. 2012.

8. Y. SAHIN AND E. DUMAN, "DETECTING CREDIT CARD FRAUD BY DECISION TREES AND SUPPORT VECTOR MACHINES," IN PROC. INT. MULTICONF. ENGINEERS AND COMPUTER SCIENTISTS (IMECS), 2011, VOL. 1, PP. 442–447.

9. M. RYMAN-TUBB, R. KRAUSE, AND A. KASPER, "HOW ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING RESEARCH IMPACTS PAYMENT CARD FRAUD DETECTION: A SURVEY AND INDUSTRY BENCHMARK," ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE, VOL. 76, PP. 130–157, JUN. 2018.

10. X. LI, C. XIE, Y. WANG, AND J. HE, "USING COST-SENSITIVE LEARNING AND SAMPLING FOR FRAUD DETECTION," IN PROC. IEEE INT. CONF. BIG DATA (BIG DATA), DEC. 2017, PP. 1187–1196