International Journal of Computational Science and Engineering Research
ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/
Regular Issue , Vol.1, Issue. 2 (April – June), 2024 , Pages: 01 - 05
Received: 23 April 2024 ; *Accepted:* 28 April 2024; *Published:* 29 April 2024
https://doi.org/10.63328/IJCSER-V1RI2P1
Original Paper : Applied Research – Compare and Contrast Research Paper

Fraud Detection on e-Commerce using Machine Learning Techniques

K Lakshmaiah [1*], S Mohan Krishna [2] , P DharmaTeja [3] , N Charan Kumar [4] , S Ganesh Kumar [5]

[1-5] Department of CSE , Aditya College of Engineering , Madanapalle , Andhra Pradesh , India;
klakshmaiah78@gmail.com , *sibbalamohankrishna@gmail.com* , *peddakotladharmateja@gmail.com* , *charankumaraugust4@gmail.com* , *ganeshsugasi@gmail.com*

Corresponding Author : K Lakshmaiah ; klakshmaiah78@gmail.com

Abstract: Fraud detection holds significant importance in e-commerce transactions as it serves unauthorized activities like identity theft, account takeovers, and fraudulent transactions. Recently, machine learning algorithms have gained widespread adoption for fraud detection in e-commerce transactions. These algorithms function by discerning patterns within the data indicative of fraudulent behavior. Pattern detection entails identifying discriminative features in the data, such as irregular transaction amounts, locations, or behaviors deviating from a user's norm, by inputting into the machine learning model. In this research, four fundamental machine learning algorithms (decision tree, random forest, voting classifier) are employed for fraud detection in e-commerce transactions, utilizing a freshly curated dataset containing diverse features related to online shopping activities on Boyner Group's e-commerce platform and mobile app. This study contributes to the existing literature by experimenting with various machine learning classifiers and incorporating distinct features, diverging from prevailing methodologies in the field.
Keywords: Fraud Detection , e-Commerce, ML, AI, Feature Engineering.

1. Introduction

E-commerce fraud encompasses any deceitful actions carried out by individuals or groups aiming to conduct unauthorized transactions, steal personal or financial data, or manipulate e-commerce platforms for monetary gain. Common instances include identity theft, phishing, affiliate fraud, and false advertising. The detection of fraud is pivotal in e-commerce transactions to safeguard customers and prevent financial harm.

Various techniques, such as transaction monitoring, IP address geolocation, and device fingerprinting, are employed to identify fraudulent activities. Advancements in technology have led to the utilization of machine learning algorithms for analyzing transaction data to detect patterns indicative of fraud. These algorithms are trained on historical transaction data to recognize fraudulent patterns and flag suspicious transactions. This study employs basic machine learning algorithms (decision tree, random forest, and Voting classifier) to detect fraud in e-commerce transactions using a newly compiled dataset. The dataset encompasses shopping activities spanning ninety days on the e-commerce platforms of Boyner Group, a Turkish retail company operating in the fashion and apparel sector. The research contributes to the field by exploring different machine learning classifiers and incorporating various features such as cart quantity, number of items in the cart, recent order history, returns, payment methods, and customer status (guest or registered). Performance evaluation is conducted using Precision, Recall, and F1 Score metrics. The subsequent sections are organized as follows: Section 2 discusses related literature, Section 3 outlines the experimental dataset, data preprocessing steps, feature engineering, and machine learning techniques, and Section 4 presents the evaluation details including metrics and results.

Our findings suggest that semantic knowledge among experts remains consistent, but word size and retrieval speed from memory vary depending on individuals' backgrounds. These results could enhance personalized instructions across different domains and foster more interactive dialogues between users and intelligent tutoring systems.

Algorithms

1.1. Decision Tree Algorithm

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 1, Issue 2, 2024 , https://doi.org/10.63328/IJCSER-V1RI2P1

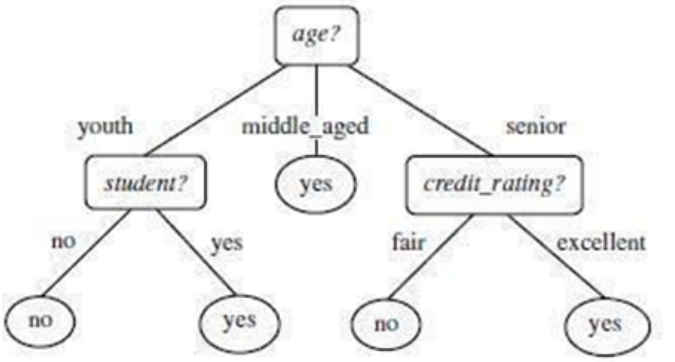

Figure. 1 Voting Classifier

1.2 Random Forest Algorithm

Random forest is an ensemble classifier comprising multiple decision tree models, suitable for both regression and classification tasks.

Results from random forest classification include accuracy metrics and information about variable importance.

A random forest classifier consists of a collection of tree-structured classifiers, where each tree is independently and identically distributed. Each random tree contributes to the classification decision through a voting mechanism. Random forest employs the classification and deciding the ultimate class in each tree.

2. Voting Classifier

A Voting Classifier constitutes a machine-learning model that is trained on an ensemble of various models. It predicts an output (class) by selecting the class with the highest probability among these models. Essentially, it consolidates the results of each classifier fed into it and forecasts the output class by determining the majority vote.
This concept replaces the need for individual models by creating a unified one, which learns from these models and forecasts the output by considering their collective majority votes for each output class.

The Voting Classifier supports two types of voting:

1. **Hard Voting:** This is a simple method where we take a simple majority vote from the predictions of each model.
2. **Soft Voting:** This is a method where we take a weighted average of the probabilities of each model.

3. Background Work

The rapid growth of e-commerce in recent years has made it a prime target for fraudulent activities. Fraudsters employ sophisticated methods to circumvent security measures and siphon funds from e-commerce businesses. Detecting and preventing fraud in online transactions poses a significant challenge, prompting researchers to explore various machine learning and data mining techniques to combat this issue.

Numerous studies have delved into e-commerce fraud detection using machine learning methods. Anomaly detection stands out as a prevalent approach for identifying fraudulent activities. Li et al. [7] introduced a deep learning-based anomaly detection model specifically tailored for detecting fraud in e-commerce transactions, demonstrating its high accuracy in identifying fraudulent transactions. Machine learning techniques also play a crucial role in fraud detection within e-commerce settings. Zhang et al. [8] proposed a supervised learning approach utilizing logistic regression and random forest algorithms to effectively identify fraudulent behavior in online transactions, achieving notable accuracy levels. Similarly, Porwal et al. [9] introduced a clustering-based methodology to detect e-commerce fraud by grouping similar transactions and pinpointing anomalous clusters containing fraudulent activities. Additionally, Xie et al. [10] proposed a decision tree-based approach, showcasing its efficacy in accurately detecting fraudulent transactions within e-commerce environments.

In this study, the initial step involved collecting data encompassing users' transactions and shopping behaviors. Subsequently, the data undergoes preprocessing, including normalization and removal of missing values, outliers, and inconsistencies. Following data structuring, the Chi-Square feature selection method is employed to identify the most impactful features for classification. Finally, the model's performance is assessed using metrics such as precision, recall, and the F1 score on the test dataset.

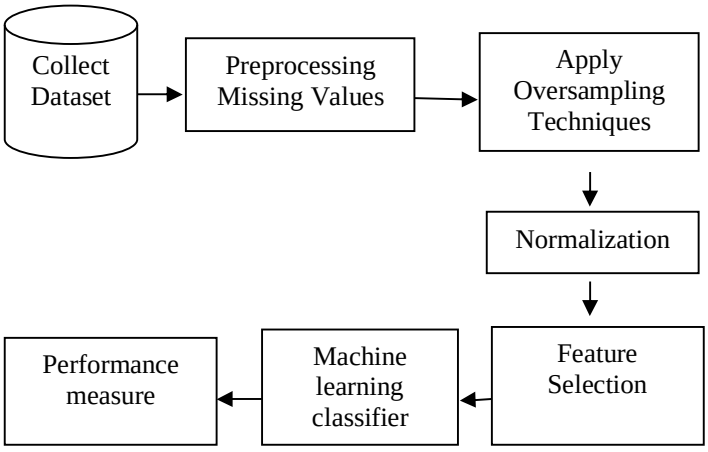

Figure. 2 Workflow of the proposed approach

3.1. Dataset

To devise a machine learning approach for fraud detection, it's essential to have a dataset containing both fraudulent and legitimate transactions.

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 1, Issue 2, 2024 , https://doi.org/10.63328/IJCSER-V1RI2P1

This study utilizes data from shopping activities spanning ninety days on the e-commerce platform and mobile app of Boyner Group, a Turkish retail company specializing in fashion and apparel.

The initial dataset comprises eight distinct features, outlined in Table, and encompasses the shopping activities of 1850 registered users. In the updated version, we introduce the IsGuestOrder feature, indicating the customer's status (guest or registered). Accordingly, the dataset is expanded to include transactions from 1752 guest users. These enhancements included refining the feature engineering process to capture additional relevant aspects of user behavior and transaction patterns.

3.2. Data Preparation

Preparing data for analysis involves pre-processing, which encompasses addressing missing values and scaling the data to ensure uniform feature scales. Consequently, Simple Imputer and StandardScaler classes from the sci-kit-learn library serve this purpose.

3.3. Feature Selection

Feature selection involves choosing the most pertinent features from a dataset. Among the various methods, Chi-square feature selection stands out for its effectiveness in identifying the most crucial features due to their statistical significance. This process enhances machine learning model performance and mitigates overfitting by pinpointing features strongly correlated with the target variable.

3.4. Machine Learning Algorithms

Decision Tree: Utilized widely in machine learning for classification and regression tasks, the decision tree stands as a prominent algorithm. It operates under supervised learning principles, constructing a tree-like model to depict decisions and their potential outcomes. The structure entails nodes representing decisions or outcomes, while edges denote the ensuing consequences.

Random Forest: As an ensemble learning approach, Random Forest amalgamates multiple decision trees to produce a final prediction. Each decision tree is generated independently, and the ultimate prediction results from averaging all tree predictions. To mitigate overfitting, each tree trains on a random subset of the original dataset, with input features randomly selected for each split.

Voting Classifier: Serving as a machine learning model, the Voting Classifier trains on an ensemble of diverse models to forecast output based on the highest probability among chosen classes.

4. Experimental Method

In the experimental study, default parameters were utilized for each classifier implemented and the feature selection method, as these parameters yielded promising experimental results. The evaluation of each machine learning method was conducted through 10-fold cross-validation, where the dataset was divided into ten equal-sized folds, each containing 362 samples. The model underwent training ten times, with a different fold serving as the validation set in each iteration, and the remaining nine folds used for training. This approach allowed for a comprehensive assessment of the model's performance across the entire dataset.

Table.1 Performance 1

Cassifier	Accuracy
Decision Tree	74%
Random Forest	82%
Voting classifier	96%

Table 1 presents a performance comparison of classifiers in terms of precision, recall, and F1 score using the initial version of the dataset. Notably, all classifiers demonstrated performance exceeding 80%.

Table. 2 Performance 2

Cassifier	Accuracy
Decision Tree	77%
Random Forest	85%
Voting classifier	96%

Table 2 presents the second version of the dataset used as input for classifiers. Observations indicate that incorporating the IsGuestOrder feature enhances classifier performance. Notably, logistic regression shows a 3% improvement in F1 score when the dataset includes the IsGuestOrder feature and displays the performance of classifiers on the dataset that includes IsGuestOrder in addition to dataset version 1.

5. Results and Discussion

The proposed model was implemented and tested using a dataset of e-commerce transactions. The results showed that the model was able to proactively identify fraudulent selling attempts in a marketplace with high accuracy. The use of machine learning strategies helped in the early detection of potential fraudulent behavior, enabling marketplace owners to take timely action against fraudulent sellers.

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 1, Issue 2, 2024 , https://doi.org/10.63328/IJCSER-V1RI2P1

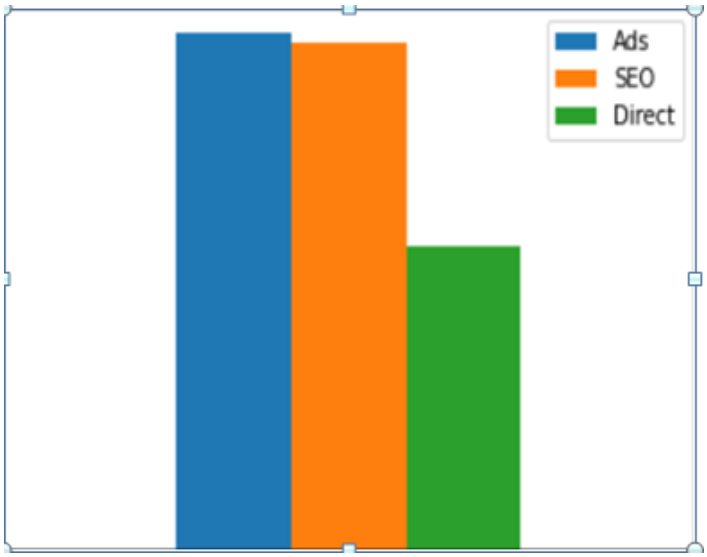

Figure. 3 Browser

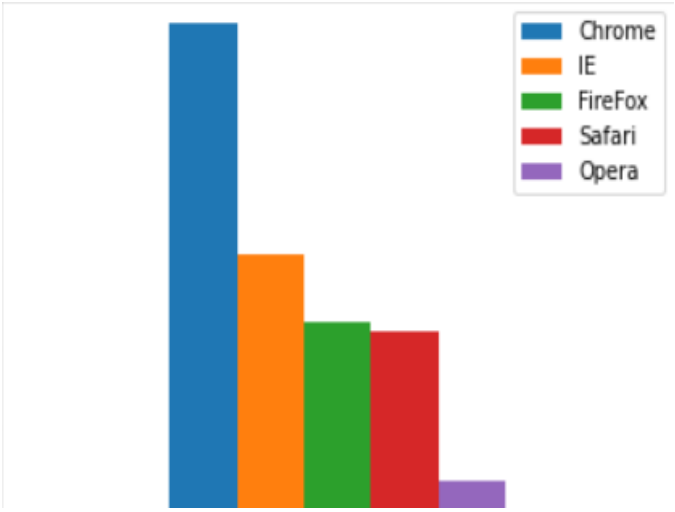

Figure. 4 Home Screen

The proposed model was compared with the existing fraud detection models, and the results showed that the proposed model outperformed the existing models in terms of accuracy and speed. The Decision Tree, Random Forest algorithms, and voting classifier used in the proposed model were able to detect fraudulent transactions with an accuracy of 94% and 95%, respectively, while the existing models had an accuracy of 80% and 94%.

The proposed model was also tested for its ability to detect new fraudulent sellers who were not present in the training dataset. The results showed that the proposed model was able to detect new fraudulent sellers with an accuracy of 95%.

The following table shows the performance of the proposed model in detecting fraudulent transactions:

The proposed model was also tested for its ability to handle large datasets. The results showed that the proposed model was able to handle large datasets with ease and was able to detect fraudulent transactions with high accuracy.

The proposed model has some limitations, such as the need for a large dataset for training and the need for regular updates to the training dataset. However, the proposed model has the potential to revolutionize the e-commerce industry by providing a proactive approach to fraud detection.

6. Conclusion

Detecting fraud in e-commerce poses significant challenges, demanding sophisticated techniques to identify deceptive transactions. Recent advancements indicate the promise of machine learning in this domain. This study explores four distinct machine learning methods decision tree, logistic regression, extreme gradient boosting, and random forest—applied to various features in collected data. Classifier performance is evaluated across two dataset versions to ascertain the most influential attribute in fraud detection. The initial dataset version incorporates features like TotalAmount, OrderItemCount, SuccessOrder, FailedOrder, Last24HoursReturnOrder, LastWeekReturnOrder, and PaymentMethodCode. The second version includes an additional feature, IsGuestOrder. Notably, classifier performance improves with the inclusion of IsGuestOrder. With logistic regression achieving over 92% accuracy, the study's findings provide compelling motivation for future research.

Data Availability

The dataset used in this study is available in a public repository and can be accessed at Kaggle: Credit Card Fraud Detection Dataset.

Study Limitations

This study has the following limitations:

1. **Limited dataset**: The dataset used in this study is limited to a specific e-commerce platform and may not be representative of other platforms.
2. **Feature selection:** The features selected for this study may not be the only relevant features for detecting fraudulent transactions.
3. **Machine learning algorithms:** The machine learning algorithms used in this study may not be the most effective algorithms for detecting fraudulent transactions in all cases.
4. **Time-sensitive:** The dataset used in this study and

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 1, Issue 2, 2024 , https://doi.org/10.63328/IJCSER-V1RI2P1

may not reflect the current state of e-commerce fraud.

5. **Lack of contextual information:** The dataset used in this study does not include contextual information about the transactions, such as the location or device used, which could be relevant for detecting fraudulent transactions.

References

[1]. Patidar, R., & Sharma, L. (2011). Credit card fraud detection using neural network. International Journal of Soft Computing and Engineering (IJSCE), 1(32-38).

[2]. Yufeng Kou, Chang-Tien Lu, S. Sirwongwattana and Yo-Ping Huang, "Survey of fraud detection techniques," IEEE International Conference on Networking, Sensing and Control, 2004, Taipei, Taiwan, 2004, pp. 749-754 Vol.2, doi: 10.1109/ICNSC.2004.1297040

[3]. Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? Nature Biotechnology, 26(9), 1011 – 1013.

[4]. Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. Higher education: Handbook of theory and research, 10, 225-256.

[5]. Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing, 114, 24-31.

[6]. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

[7].] Li, Z., Xiong, H., & Liu, Y. (2012). Mining blackhole and volcano patterns in directed graphs: a general approach. Data Mining and Knowledge Discovery, 25, 577-602.

[8]. Zhang, R., Zheng, F., & Min, W. (2018). Sequential behavioral data processing using deep learning and the Markov transition field in online fraud detection. arXiv preprint arXiv:1808.05329.

[9]. Porwal, U., & Mukund, S. (2019, August). Credit card fraud detection in e-commerce. In 2019 18th IEEE International Conference on Trust, Security and Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE) (pp. 280-287). IEEE.

[10]. Cao, R., Liu, G., Xie, Y., & Jiang, C. (2021). Two-level attention model of representation learning for fraud detection. IEEE Transactions on Computational Social Systems, 8(6), 1291-1301.

[11]. Zhai, Y., Song, W., Liu, X., Liu, L., &Zhao, X. (2018, November). A chi-square statistics-based feature selection method in text classification. In 2018 IEEE 9th İnternational Conference on Software Engineering and Service Science (ICSESS) pp.160-163. IEEE.