AITA
ARTIFICIAL INTELLIGENCE
THEORY AND APPLICATIONS

■ RESEARCH ARTICLE

Fraud Detection on E-commerce Transactions Using Machine Learning Techniques

Murat Gölyeri[a], **Sedat Çelik**[a], **Fatma Bozyiğit**[bc†], **Deniz Kılınç**[d]

[a] Boyner Group, Istanbul, Turkey
[b] Department of Computer Science, University of Antwerp, Antwerp, Belgium
[c] AnSyMo/CoSys corelab, Flanders Make, Leuven, Belgium
[d] Department of Computer Engineering, İzmir Bakırçay University, İzmir, Turkey
[†] fatma.bozyigit@uantwerpen.be

RECEIVED MARCH 30, 2023
ACCEPTED APRIL 29, 2023

CITATION Gölyeri, M., Çelik, S., Bozyiğit, F., & Kılınç. D. (2023). Fraud detection on e-commerce transactions using machine learning techniques. *Artificial Intelligence Theory and Applications, 3*(1), 45-50.

Abstract

Fraud detection is an important aspect of e-commerce transactions as it helps to prevent fraudulent activities such as unauthorized transactions, identity theft, and account takeovers. Recently, machine learning algorithms have been widely used in the literature to detect fraud in e-commerce transactions. These algorithms work by learning patterns in the data that indicate fraudulent activity. Pattern detection involves discovering the discriminative features in the data, such as unusual transaction amounts, locations, or behaviors that are out of the normal range for a particular user, to feed the machine learning method. In this study, four basic machine learning algorithms (decision tree, logistic regression, random forest, and extreme gradient boosting) are used to detect fraud in e-commerce transactions using a newly created dataset including various features about online shopping activities on Boyner Group's e-commerce website and mobile application. The study contributes to the literature by trying different machine learning classifiers and utilizing different features that differ from current approaches in the literature.

Keywords: fraud detection; e-commerce; machine learning; feature engineering

1. Introduction

E-commerce fraud refers to any fraudulent or dishonest activity conducted by individuals or groups performing unauthorised transactions, steal personal or financial information, or manipulate e-commerce systems for financial gain. Some common examples of e-commerce fraud include identity theft, phishing, chargeback fraud, affiliate fraud and false advertising.

Fraud detection is a crucial aspect of e-commerce transactions as it helps protect the customers and prevent financial losses [1]. There are some techniques that can be used to detect fraud in e-commerce transactions, such as transaction monitoring, IP address geolocation and device fingerprinting. With evolving technology, machine learning algorithms can be used to analyze transaction data and identify patterns indicative of fraudulent activity. These algorithms can be trained on historical transaction data to detect fraudulent patterns and flag suspicious transactions [2].

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN : 978-605-69730-2-4 © 2023 University of Bakırçay

In this study, four basic machine learning algorithms (decision tree [3], logistic regression [4], random forest [5] and extreme gradient boosting [6]) are used to detect fraud in e-commerce transactions using a newly created dataset. The dataset includes shopping activities during ninety days on the e-commerce website and mobile application of Boyner Group[1], a Turkish retail company operating in the fashion and apparel industry. The study contributes to the literature by trying different machine learning classifiers using various features such as cart quantity, number of items in the cart, number of successful orders in the last 24 hours, number of failed orders in the last 24 hours, number of returns in the last 24 hours, number of returns in the last week, order ID, payment method and customer status (guest or registered). The performance of the classifiers are then compared using the metrics of Precision, Recall and F1 Score. The remaining parts of the study are structured as follows. Section 2 discusses related work. Section 3 provides information about the experimental dataset, data preprocessing steps, feature engineering and machine learning methods used in the study. Section 4 presents the details of the evaluation with metrics and results. Our research shows that experts' knowledge does not differ semantically, but rather the information is interconnected. The experiments conducted have shown that the size of words and the retrieval speed of words from memory varies between individuals with different background knowledge. The results of our study could also help to provide better, personalized instructions to users in different areas and to build a more interactive dialogue between the user and an intelligent tutoring system.

2. Related Work

E-commerce has grown rapidly in recent years, making it an attractive target for fraudulent activity. Fraudsters use sophisticated techniques to bypass security measures and steal money from e-commerce businesses. Detecting and preventing fraud in e-commerce transactions is a challenging task, and researchers have explored various machine learning and data mining techniques to address this problem.

Several studies have investigated e-commerce fraud detection using machine learning techniques. Anomaly detection is a commonly used technique for fraud detection. In their study, Li et al [7] proposed a deep learning-based anomaly detection model to detect fraud in e-commerce transactions. They showed that their model can detect fraudulent transactions with high accuracy. Machine learning is another popular technique for fraud detection in e-commerce. In their study, Zhang et al [8] proposed a supervised learning approach to detect fraudulent behavior in e-commerce transactions. They used logistic regression and random forest algorithms to train their model and achieved high accuracy in detecting fraudulent transactions. Porwal et al [9] proposed a clustering-based approach for detecting fraud in e-commerce transactions. They used clustering to group similar transactions together and identified anomalous clusters that contained fraudulent transactions. In another study, Xie et al [10] proposed a decision tree-based approach for e-commerce fraud detection. They showed that their approach can detect fraudulent transactions effectively and with high accuracy.

[1] https://www.boyner.com.tr/

3. Materials and Methods

In this study, the first step is to collect data consisting of users' transactions and shopping activities. Then the data is prepared using pre-processing methods by normalizing and removing missing values, outliers and other inconsistencies. After structuring the data, the ChiSquare [11] feature selection method is applied to determine which feature is most effective in classification. Finally, the performance of the model is evaluated using metrics such as precision, recall and F1 score in the test set.

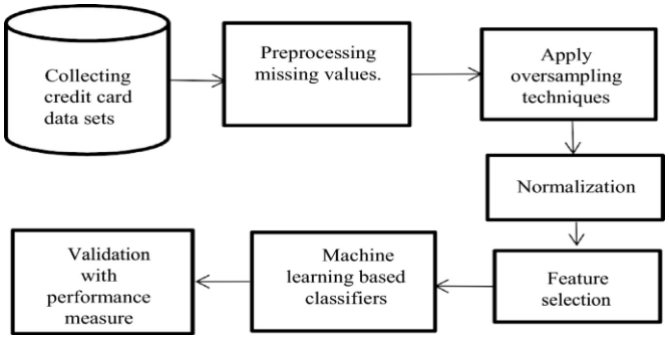

Figure 1. Workflow of the proposed approach

3.1. Dataset

To develop a machine learning method for fraud detection, a dataset containing both fraudulent and legitimate transactions is needed. In this study, the dataset includes shopping activities during ninety days on the e-commerce website and mobile application of Boyner Group, a Turkish retail company operating in the fashion and apparel industry.

In this study, we included eight different features, explained in Table 1, in a first version of the dataset. In this version, shopping activities of 1850 registered users are included. In the second version of the dataset IsGuestOrder feature is included. This feature shows the situation of the customer (guest or registered). Consideringly, 1752 guest users' transactions are added to dataset.

Table 1. Features in the dataset

Feature	Feature Name	Feature Description
Feature 1	TotalAmount	basket amount
Feature 2	OrderItemCount	number of items in the basket
Feature 3	SuccessOrder	number of successful orders in the last 24 hours
Feature 4	FailedOrder	number of failed orders in the last 24 hours
Feature 5	Last24HoursReturnOrder	number of returns in the last 24 hours
Feature 6	LastWeekReturnOrder	number of returns in the last week
Feature 7	OrderID	order ID
Feature 8	PaymentMethodCode	payment method

3.2. Data Preperation

Pre-processing is a necessary step to prepare the data for analysis. This includes dealing with missing values and scaling the data so that all features are at a similar scale.

Consideringly, SimpleImputer and StandardScaler classes from the scikit-learn library are used for this purpose.

3.3. Feature Selection

Feature selection is a technique used to select the most relevant features from a data set. Chi-square feature selection is one of the most effective techniques for selecting the most important features based on their statistical significance. By identifying the features that are most strongly associated with the target variable, it can help improve the performance of machine learning models and reduce the risk of overfitting.

3.4. Machine Learning Algorithms

Decision Tree

Decision tree is one of the widely used machine learning algorithms for classification and regression tasks. It is a supervised learning algorithm that builds a tree-like model of decisions and their possible consequences. The tree structure consists of nodes and edges, where the nodes represent the decision or outcome, and the edges represent the possible consequences of the decision.

Logistic Regression

It is a type of regression analysis used when the dependent variable is binary or dichotomous. The aim of logistic regression is to model the probability of a particular outcome based on one or more predictor variables.

Extreme Gradient Boosting

Extreme Gradient Boosting is a tree-based ensemble method that combines the predictions of multiple decision trees to produce a final prediction. The algorithm is very effective in dealing with high-dimensional data and has the ability to model non-linear relationships between variables.

Random Forest

Random Forest is a type of ensemble learning method in which multiple decision trees are created and combined to make a final prediction. In Random forest, each decision tree is created independently and the final prediction is made by averaging the predictions of all the trees. To prevent overfitting, each tree is trained on a random subset of the original dataset and a random subset of the input features is used for each split in the tree.

4. Experimental Study

In the experimental study, the default parameters are set for each classifier implemented and feature selection method since these parameters give promising experimental results. The evaluation results of each machine learning method are obtained by dividing the data set into 10 pieces by cross-validation (Table 2). To perform 10-fold cross-validation on this data, the data is divided into ten equal-sized folds, each with 362 samples. The model is trained ten times, using a different fold as the validation set and the other nine folds as the training set, to better assess the performance of the model for the entire data set.

Table 2 shows the performance comparison of classifiers in terms of precision, recall, and F1 score on the first version of the dataset. As it can be seen from Table 2, the performance of the all classifiers are over 80%.

Table 2. Performance of classifiers on the dataset containing TotalAmount, OrderItemCount, SuccessOrder, FailedOrder, Last24HoursReturnOrder, LastWeekReturnOrder, PaymentMethodCode features

Classifier	Accuracy	Precision	Recall	F-measure
Decision Tree	0.83	0.82	0.78	0.80
Logistic Regression	0.89	0.91	0.86	0.88
Extreme Gradient Boosting	0.90	0.85	0.92	0.88
Random Forest	0.81	0.80	0.87	0.83

In Table 3, the second version of the dataset is inputted for classifiers. It is seen that when the IsGuestOrder feature is included the performance of classifiers increases. For instance, the performance of logistic regression is increased 3% in terms of F1 score on the dataset including IsGuestOrder feature.

Table 3. Performance of classifiers on the dataset containing IsGuestOrder in addition to dataset version 1.

Classifier	Accuracy	Precision	Recall	F-measure
Decision Tree	0.83	0.82	0.79	0.80
Logistic Regression	0.93	0.95	0.89	0.92
Extreme Gradient Boosting	0.90	0.87	0.91	0.89
Random Forest	0.86	0.81	0.90	0.85

5. Conclusion

Detecting fraud in e-commerce is a challenging task that requires the use of sophisticated techniques to detect fraudulent transactions. The current state of the art shows that machine learning techniques are promising for detecting fraudulent activities in e-commerce transactions. In this study, four different machine learning methods (decision tree, logistic regression, extreme gradient boosting, and random forest) are performed considering different features in the collected data. The performance of the classifiers is compared against two versions of the dataset to find the most relevant attribute in detecting fraudulent activity. In the first version of the dataset, the features TotalAmount, OrderItemCount, SuccessOrder, FailedOrder, Last24HoursReturnOrder, LastWeekReturnOrder and PaymentMethodCode are used as input to the classification algorithms. In the second version of the dataset, the feature IsGuestOrder is added as an additional feature. It can be seen that the performance of the classifiers increases when the feature IsGuestOrder is included. Since the performance of the logistic regression was calculated to be over 92%, we can say that the results of the study are motivating for future work.

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