ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/05.2024-29348967/IJAEAST.2024.02.0005, February 2024, Volume-2, Issue-2, pp.75-81
International Journal of Advances in Engineering Architecture Science and Technology

Fraud Detection in Online Transactions Using Machine Learning

U. Yerraji Pavan[1], B. Bhanu Prakash[2], P. Sasidhar[3], K.V. Sai Charan[4], P. Mounika[5]

[1,2,3,4]Student, Department of AIML & DS, Dadi Institute of Engineering and Technology (Autonomous), Andhra Pradesh, India

[5]Assistant Professor, Department of AIML & DS, Dadi Institute of Engineering and Technology (Autonomous), Andhra Pradesh, India

pavanyadav85997@gmail.com[1], bourothubhanuprakash2002@gmail.com[2], pallasasidhar7@gmail.com[3], kotagirivenkatasaicharan@gmail.com[4], polinatimounika010593@gmail.com[5]

Received: 23-01-2024 **Accepted:** 27-02-2024 **Published:** 28-02-2024

Abstract

Background: As we approach the era of modernity, the prevalence of online payments is markedly increasing.
Objectives: Opting to pay online proves significantly advantageous for the consumer in emergency situations and eliminating the inconvenience of carrying physical currency.
Statistical Analysis: Moreover, the inclination to abstain from holding cash is becoming more pronounced. Nevertheless, it is imperative to acknowledge that “Positive developments often coexist with challenges”.
Findings: The adoption of online transaction methods has given rise to the potential for fraudulent activities, which can manifest in the utilization of various payment applications. Consequently, the implementation of robust online transaction fraud detection mechanisms becomes paramount.
Applications and Improvements: The primary objective revolves around the identification of such fraudulent activities, encompassing factors such as the scrutiny of publicly available data and the evolving nature of fraudulent patterns.

Keywords: Logistic Regression, Support Vector Classifier, KNN Classifier, Decision Tree, Random Forest.

1. Introduction

We are really close to having a cashless society in the modern world. Numerous studies and surveys indicate that the number of people conducting transactions online has increased significantly, and it is anticipated that this trend will continue in the next years. Although this may be welcome news, there is also an increase in fraudulent transactions. We still have a significant amount of money lost to fraudulent transactions even with the use of numerous security technologies. When someone uses someone else's debit or credit card for personal purposes without the owner's knowledge or the authorities who issue the card, it's known as an online fraud transaction. Fraud detection is the process of keeping an eye on user populations' behaviors in order to gauge, identify, or steer clear of undesirable activity, which includes fraud, intrusion, and defaulting. The majority of the time, a victim of this kind of scam is unaware of it until the very end.

The behavior of such fraudulent acts may be researched to limit it and guard against similar occurrences in the future. Necessary preventative actions can be made to halt this misuse. Put another way, this is a highly pertinent issue that has to be addressed by fields like data science and machine learning, since these fields can automate the answer. From the standpoint of learning, this issue is especially difficult since it is characterized by a number of variables, including class imbalance. The quantity of legitimate transactions

greatly exceeds the quantity of fraudulent ones. Furthermore, the statistical characteristics of the transaction patterns frequently alter over time.

However, these are not the only difficulties in putting a real-world fraud detection system into practice. In real-world scenarios, automated systems swiftly sort through the enormous volume of payment requests to decide which ones to approve. Algorithms for machine learning are used to analyze all approved transactions and flag those that seem suspect. Experts look into these allegations and get in touch with the cardholders to verify whether or not the transaction was fraudulent. In order to train and update the algorithm and finally enhance the fraud-detection effectiveness over time, the investigators supply input to the automated system. Therefore, the goal of this project is to develop a system that uses machine learning to identify these kinds of scams.

2. Literature Review

The issue of detecting payment fraud has been tackled using a number of ML and non-ML based techniques.

The paper [1] offers a useful examination of the training speed, generalization performance, and parameter configuration of this innovative approach. Furthermore, a thorough comparison of gradient boosting, random forests, and XGBoost has been carried out using both the default parameters and carefully adjusted models. The comparison's findings could suggest that XGBoost isn't always the greatest option, but it does have certain benefits over other algorithms.

The proposed system in this research paper [2] classifies fraud activities by combining the SMOTE technique with the Xgboost classification algorithm. Synthetic Minority Oversampling Technique is referred to as SMOTE. This method will help you add more cases to your dataset in a balanced manner. SMOTE increases the percentage of only the minority cases, using the entire dataset as input. They used only publicly available datasets for credit card frauds using XGBoost to measure and validate their performance.

An algorithm for detecting fraud that uses Xgboost is presented in this paper [3]. In order to eliminate certain anomalies, they first cleaned the data. Furthermore, the minority class was oversampled using the SMOTE (Synthetic Minority Oversampling Technique) in order to address the imbalanced distribution of labels. On the other hand, categorical data is encoded using the label encoding algorithm for categorical features. Lastly, Xgboost, a very popular and successful algorithm, was used for the classification.

This study [4] demonstrated how the XGBoost model's optimization can effectively address class imbalance in the dataset on its own. The RandomizedSearchCV technique, which can be found in the Python scikit-learn package, is used to determine the ideal parameters. Two real-world imbalanced datasets were used in the experiment, and various sampling techniques were integrated to conduct the mathematical derivative of XGBoost. Our results demonstrated that, in the absence of sampling, the suggested XGBoost can attain greater accuracy for highly unbalanced data.

The study [5] examines and contrasts a variety of cutting-edge methods, datasets, and assessment standards used to this issue. It covers supervised and unsupervised machine learning techniques that use clustering, ANNs (Artificial Neural Networks), SVMs (Support Vector Machines), and HMMs (Hidden Markov Models), among other techniques. The authors of [6] suggest an SVM-based method for identifying metamorphic malware. The issue of unbalanced data sets—fewer malware samples than benign files—and effective, highly accurate malware detection techniques are also covered in this paper.

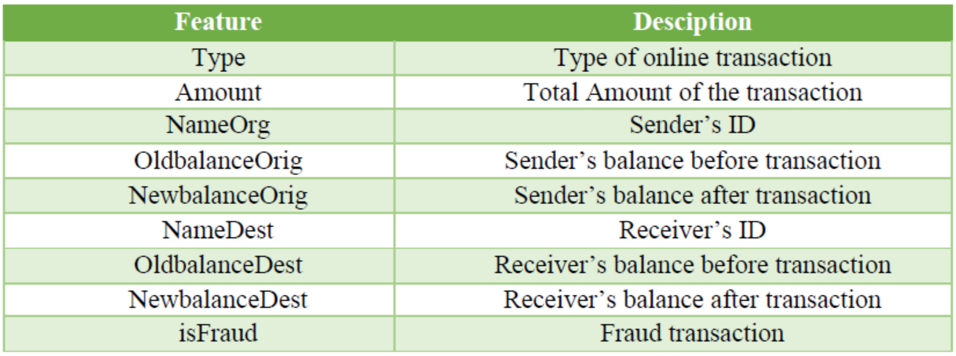
3. Methodology

By obtaining a decision boundary in the feature space defined by input transactions, we hope to distinguish between transactions that are fraudulent and those that are not. A vector of a transaction's feature values can be used to represent it. We have constructed binary classifiers using Logistic Regression, Support Vector Classifier, KNN Classifier, Decision Tree, Random Forest.

Dataset

The dataset is crucial to the model's classification. The dataset was obtained from the knowledgeable data science organization Kaggle's official website. Millions of transactions are detailed in this dataset, some of which are fraudulent transactions. As a result, the system is developing more smoothly and consistently. This dataset highlights the challenges in acquiring data on the increasing risk of digital financial fraud.

Table 1. Digital Financial Fraud



Feature	Desciption
Type	Type of online transaction
Amount	Total Amount of the transaction
NameOrg	Sender's ID
OldbalanceOrig	Sender's balance before transaction
NewbalanceOrig	Sender's balance after transaction
NameDest	Receiver's ID
OldbalanceDest	Receiver's balance before transaction
NewbalanceDest	Receiver's balance after transaction
isFraud	Fraud transaction

Six million rows of data contain a highly imbalanced distribution of positive and negative classes, which presents the primary technical challenge for fraud prediction. Transaction type, amount, nameOrig, oldbalanceOrg, newbalanceOrig, nameDest, oldbalanceDest, and newbalanceDest are the parameters of this dataset.

Data Pre-Processing

This step includes the following:

- Data Set Cleaning
- Normalizing Data Set
- Balancing Data

Data Visualization

Data visualization is the graphical representation of data, crucial for exploratory data analysis (EDA) to understand data distribution, patterns, and relationships. It aids in identifying outliers and understanding variable relationships. In imbalance detection, visualization helps assess class distribution, crucial in addressing class imbalance in machine learning models.

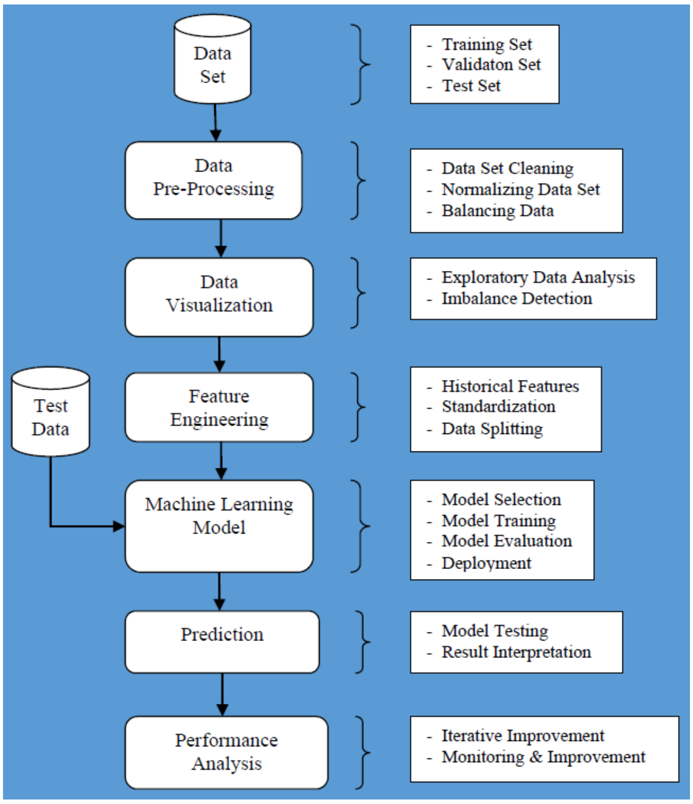


Figure 1. Data Preprocessing

Feature Engineering

Feature engineering is pivotal in optimizing machine learning model performance. It leverages historical data to capture trends and standardizes features for balanced contribution. Data splitting ensures robust model evaluation across training, validation, and testing sets. These steps collectively enhance model accuracy and generalization, crucial for effective fraud detection in online transaction systems.

Machine Learning Model

Logistic Regression: It differs from linear regression by predicting probabilities rather than continuous values. This approach offers simplicity, interpretability, and efficiency. By transforming outputs using the logistic function, it ensures predictions fall between 0 and 1, representing the likelihood of fraud. The model's coefficients provide valuable insights into the importance of each feature for fraud detection. However, logistic regression assumes linear relationships between features and outcomes, which may limit its effectiveness in complex datasets.

Support Vector Classifier (SVC): The way SVC operates is by identifying the hyperplane in the feature space that best divides various classes. SVC is advantageous for its effectiveness in high-dimensional spaces and its ability to handle non-linear relationships between features through the use of kernel functions. However, SVC can be sensitive to the choice of kernel parameters and may not provide probability estimates by default. Despite these limitations, SVC is a powerful tool for fraud detection, particularly when combined with appropriate tuning and optimization techniques.

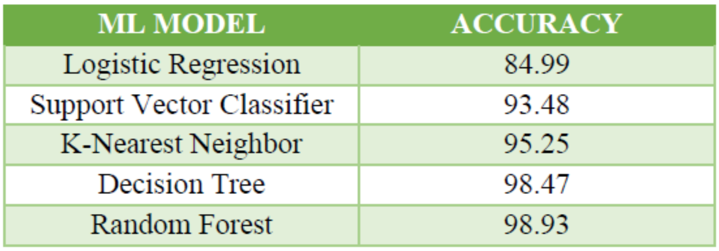
K-Nearest Neighbor (KNN): KNN works by classifying a data point based on the class of its nearest neighbors in the feature space. KNN's simplicity and adaptability to non-linear boundaries make it a popular choice, although it can be computationally intensive with large datasets. Proper tuning of the K parameter is crucial for optimal performance. Overall, KNN is a powerful tool for fraud detection, particularly in scenarios where data distribution is complex or the dataset is not too large.

Decision Tree: Decision Tree is a robust algorithm for fraud detection in online payments, dividing the feature space into segments for classification. Known for its simplicity and interpretability, it can be easily visualized. Despite a tendency to overfit, Decision Tree remains a valuable tool due to its effectiveness in handling various data types. Its ability to capture non-linear relationships and interactions between features makes it suitable for complex fraud detection tasks. Additionally, Decision Tree models can be easily explained to non-technical stakeholders, making them valuable in decision-making processes.

Random Forest: Random Forest, a robust machine learning algorithm for fraud detection in online payments, builds on decision trees' foundation. During training, it builds many decision trees and outputs the mode of classes for categorization. Known for reducing overfitting compared to individual trees, Random Forest is effective for complex fraud detection. Like Decision Trees, it handles various data types and provides interpretable results. Its ensemble approach enhances accuracy and robustness by combining multiple trees to capture intricate patterns and interactions in data. Despite its computational demands, the ensemble technique justifies the trade-off with significantly improved performance.

Accuracy Comparison

Table 2. Accuracy Comparison



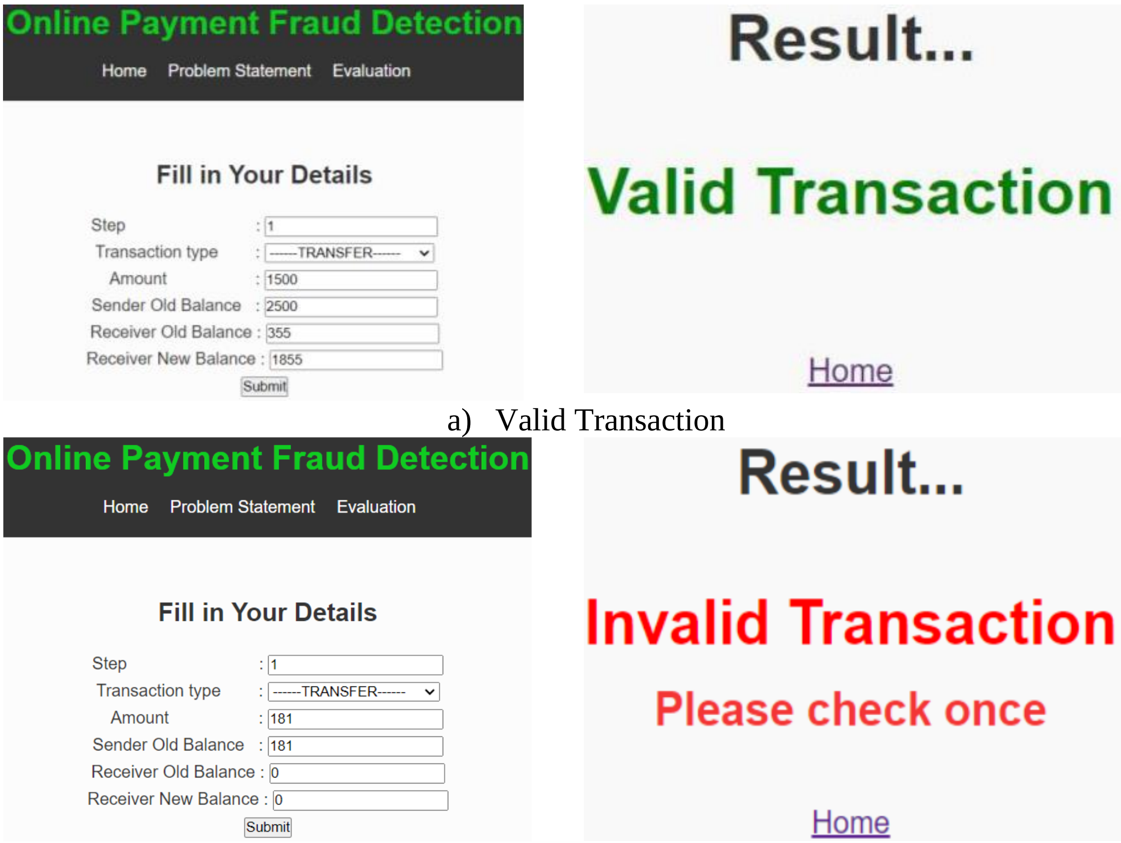
ML MODEL	ACCURACY
Logistic Regression	84.99
Support Vector Classifier	93.48
K-Nearest Neighbor	95.25
Decision Tree	98.47
Random Forest	98.93

Our ultimate model is ‘Random Forest’, which we have chosen based on Accuracy and Performance of the Algorithms.

4. Result

The Random Forest model achieved an impressive accuracy of 98.93% in classifying online transactions as fraudulent or genuine. This high accuracy underscores the model's effectiveness in accurately identifying fraudulent transactions, which is paramount for ensuring the security of online payment systems. The model's capability to handle intricate patterns and interactions in the data has significantly contributed to its exceptional performance. These results underscore the potential of Random Forest as a robust tool for fraud detection in online payments.

Test Cases



Fill in Your Details

Step	: 1
Transaction type	: ------TRANSFER------
Amount	: 1500
Sender Old Balance	: 2500
Receiver Old Balance	: 355
Receiver New Balance	: 1855

Result...

Valid Transaction

a) Valid Transaction

Fill in Your Details

Step	: 1
Transaction type	: ------TRANSFER------
Amount	: 181
Sender Old Balance	: 181
Receiver Old Balance	: 0
Receiver New Balance	: 0

Result...

Invalid Transaction

Please check once

b) Invalid Transaction

Figure 2. Test Cases

5. Conclusion

This study describes the creation of a machine learning model that detects online fraud transactions using the Random Forest method. The main characteristic of this model is the ability to categorize transactions in a given dataset as fraudulent or real. With the provided dataset, this model demonstrated a higher AUC score, accuracy score, and efficient output. The dataset is pre-processed along with feature choices, and the data is then classified into several variables before being fed into the random_forest algorithm model. The end result is to determine if the transactions are real or fraudulent. This model may then be evaluated and trained with bigger data volumes in the future, yielding more precise and accurate answers.

The model's ability to handle complex patterns and interactions contributed to its exceptional performance. Random Forest's ensemble learning approach and decision tree algorithms proved robust in fraud detection. Leveraging these technologies can improve the security and reliability of online transactions. Random Forest's success highlights its potential as a powerful tool for fraud detection. Continued research and development in this area are crucial for staying ahead of online transactions fraud.

6. Future Scope

By utilizing different classification algorithms to take advantage of the categorical features connected to users' accounts in the Paysim dataset, we can further enhance our methods. Time series can also be used to interpret the Paysim dataset. This feature can be used to create time series-based models with CNN-style algorithms.

To train our models, we currently handle the complete set of transactions as a whole. To further enhance our decision-making process, we can develop user-specific models based on the user's past transactional behavior. We think that all of these can significantly raise the quality of our classification on this dataset.

Acknowledgments

The authors express their deep sense of gratitude to the Chairman, Dadi institute of Engineering & Technology Sri Dadi Ratanakar for the facilities provide in carrying out this work successfully.

References

1. K. Chaudhary, J. Yadav, “A review of fraud: A comparative study.” decis. Support syst, vol 50, no3, pp.602-613,2011.
2. Katherine J. Barker, Jackie D’Amato, Paul Sheridon,2008 “Credit card fraud: awareness and prevention”, Journal+- of financial Crime, Vol. 15issue:4, pp.398-410.
3. Dipti Thakur, salamis Bhatia “distribution data Mining approach to credit card fraud detection” SPIT IEEE colloquium and international conference, volume4, 48, issue2002.
4. “Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea” published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017.
5. Customer Transaction Fraud Detection Using Xgboost Model -by Yixuan Zhang, Ziyi Wang, Jialiang Tong, Fengqiang Gao June, 2020.
6. Jerome H. Friedman. Greedy function approximation: a Gradient Boosting machine. The Annals of Statistics, 29(5):1189 – 1232, 2001.
7. Wang, M., Yu, J., & Ji, Z. (2018). Credit Fraud Risk Detection Based on XGBoost-LR Hybrid Model.
8. A. Mishra, C. Ghorpade, “Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques” 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) pp. 1-5. IEEE.
9. Lei, Shimin, et al: An Xgboost based system for financial fraud detection. E3S Web of Conferences, 2020, 214(2).
10. Bergstra J, et al: Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. Computational Science & Discovery, 2015, 8(1).