WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Credit card fraud detection using machine learning

Kacha Vishw P.[1], Bhatt Mishank J.[2], Mehta Parth N.[3], Trivedi Shlok S.[4], Vyas Dhruv M.[5], PROF. JANKI TEJAS PATEL[6]

[1] Computer Engineering, Sal College of Engineering
[2] Computer Engineering, Sal College of Engineering
[3] Computer Engineering, Sal College of Engineering
[4] Computer Engineering, Sal College of Engineering
[5] Computer Engineering, Sal College of Engineering
[6] Assistant professor, Computer Engineering, Sal College of Engineering

ABSTRACT :

Credit card fraud costs the global financial industry billions of dollars annually. Real-time fraud detection is a difficult problem because of the extremely imbalanced nature of datasets, the ever-changing fraud trends, and the requirement for quick choices. The application of machine learning (ML) techniques to precisely detect fraudulent transactions is investigated in this work. The Kaggle Credit Card Fraud Dataset (2013 European cardholders dataset) is used in the study to analyze a number of algorithms, including K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Logistic Regression. With a 99.93% accuracy rate and an F1-score of 0.87, the Random Forest classifier excelled in the experiment, proving the effectiveness of ensemble methods in fraud detection tasks.

Keywords: Machine Learning, Credit Card Fraud, Classification, Logistic Regression, Random Forest, KNN, Imbalanced Data

Introduction

The Internet has grown exponentially during the past ten years. This has caused services like online bill payment, tap and pay, and e-commerce to proliferate and become more widely used. As a result, scammers have also stepped up their efforts to target credit card transactions. Tokenization and encryption of credit card data are two of the many methods used to safeguard credit card transactions [1]. While these techniques are generally successful, they do not completely guard against fraud in credit card transactions.

A branch of artificial intelligence (AI) called machine learning (ML) enables computers to learn from past experience (data) and enhance their predictive capabilities without being specifically programmed to do so [2]. We use machine learning (ML) techniques to detect credit card fraud in this work. A fraudulent transaction (payment) made by an unauthorized person using a credit or debit card is known as credit card fraud [3]. The Federal Trade Commission (FTC) reports that there were about 1579 data breaches totaling 179 million data points, with credit card theft being the most common type of data breach [4]. Implementing a successful credit card fraud detection technique that can shield users from monetary loss is therefore essential.

Consequently, anonymised attributes are present in the datasets used to create machine learning models for credit card fraud detection. Furthermore, the ever-evolving structure and patterns of fraudulent transactions make it difficult to detect credit card fraud [5]. This paper's reminder is organized as follows. An overview of the classifiers utilized in this study is given in the second part. A survey of related literature is given in Section III. The dataset used in this study is described in detail in Section IV. The GA algorithm is described in Section V. The architecture of the suggested system is described in Section VI. The experiments are carried out in Section VII. Section VIII presents the conclusion.

Research synthesis

The dataset used in this study consists of credit card transactions done by European cardholders over the course of two days in September 2013. There are 284807 transactions in this dataset overall, with 0.172% of those transactions being fraudulent. Time and Amount are among the 30 features (V1,..., V28) in the dataset. Every attribute in the dataset has a numerical value. For purposes of data security and integrity, characteristics V1 through V28 are not identified [19]. One of the main problems we found with this dataset, which has been utilized in references [4, 13, 14, 16], is its poor detection accuracy score.

By carefully adjusting its parameters, the PSO technique was utilized to enhance SSAE's feature learning capabilities. According to the findings, the PSOSSAE performed 97.3% accurately on the Framingham heart disease dataset. Hemavathi et al. [22] used enhanced principal component analysis

(EPCA) to build an efficient FS approach in an integrated setting. The outcomes showed that the EPCA produces the best outcomes in both supervised and unsupervised settings.

In an e-banking setting, Pouramirarsalani et al. [20] used a hybrid FS and GA FS approach to detect fraud. The outcomes of the experiment showed that using an FS approach to financial fraud datasets can improve the models' overall performance. In order to detect credit card fraud, the authors used the GA-based FS technique in combination with the NB, SVM, and RF algorithms in ref. [14]. The experimental results showed that, when compared to the NB and SVM, the RF performed better.

Methodology

1. Data Preprocessing: This process includes feature scaling or normalization methods in order to bring the data on a similar scale. Categorical encoding is applied where required, and imbalanced data handling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and Random Undersampling are used to deal with class-skewed distribution.

2. Feature Selection: Different methods are used for feature selection, such as correlation analysis, Principal Component Analysis (PCA), or information gain techniques. Feature analysis determines which features are most important in relation to the current task and helps improve model performance and interpretability.

3. model assessment metrics are explained in terms of the significance of these metrics for fraud detection. Accuracy is critical since it represents the overall accuracy of the model. Precision shows the accuracy of positive predictions and Recall shows how well the model captures real fraud cases. The F1-Score combines Precision and Recall and is a good measure of model performance in imbalanced classes.

4. Mentions that there is an option to add a diagram to demonstrate the workflow, outlining stages from Data Input to Preprocessing, ML Model, Evaluation, and eventual Output, but the diagram is not provided.

Related Work

The authors of ref. [13] used a number of machine learning (ML) algorithms, such as logistic regression (LR), decision trees (DT), support vector machines (SVM), and random forests (RF), to create a system for detecting credit card fraud. A credit card fraud detection dataset created in 2013 from European cardholders was used to assess these classifiers. This dataset is extremely uneven since the ratio of fraudulent to non-fraudulent transactions is substantially skewed. Despite the excellent results, the authors proposed that the performance of the classifiers may be improved by using sophisticated pre-processing methods. Using machine learning, Varmedja et al. [14] suggested a technique for detecting credit card fraud. Kaggle provided the authors with a dataset on credit card fraud [19].

Transactions made by European credit cardholders within two days are included in this dataset. To evaluate the effectiveness of the suggested approach, the following machine learning techniques were used: RF, NB, and multilayer perceptron (MLP). With a 99.96% fraud detection accuracy, the experimental findings showed that the RF algorithm operated at peak efficiency. The accuracy scores for the MLP and NB approaches were 99.93% and 99.23%, respectively.

The authors acknowledge that further study is necessary to develop a feature selection technique that could increase the precision of other machine learning techniques. A performance analysis of machine learning algorithms for detecting credit card fraud was carried out by Khatri et al. [15]. The following machine learning techniques were taken into consideration by the authors of this study: DT, k-Nearest Neighbor (KNN), LR, RF, and NB. The authors created a very unbalanced dataset from European cardholders in order to evaluate the effectiveness of each machine learning technique.

Conclusion

Together with the RF, DT, ANN, NB, and LR, a GA-based feature selection technique was suggested in this study. The RF was incorporated into the fitness function of the GA. Five ideal feature vectors were produced after the GA was further applied to the dataset of credit card transactions made by European cardholders. The GA-RF (using v5) achieved an overall ideal accuracy of 99.98%, according to the experimental findings obtained using the GA chosen qualities. Additionally, other classifiers, like the GA-DT, used v1 to reach an impressive 99.92% accuracy rate. The GA-ANN, which has a 100% accuracy rate and an AUC of 0.94, came in second. We want to validate our approach using additional datasets in further research.

REFERENCES

1. Iwasokun GB, Omomule TG, Akinyede RO. Encryption and tokenization-based system for credit card information security. Int J Cyber Sec Digital Forensics. 2018;7(3):283–93.
2. Burkov A. The hundred-page machine learning book. 2019;1:3–5.
3. Maniraj SP, Saini A, Ahmed S, Sarkar D. Credit card fraud detection using machine learning and data science. Int J Eng Res 2019; 8(09).

4. Dornadula VN, Geetha S. Credit card fraud detection using machine learning algorithms. Proc Comput Sci. 2019;165:631–41.
5. Thennakoon, Anuruddha, et al. Real-time credit card fraud detection using machine learning. In: 2019 9th international conference on cloud computing, data science & engineering (Confluence). IEEE; 2019.
6. Robles-Velasco A, Cortés P, Muñuzuri J, Onieva L. Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliab Eng Syst Saf. 2020;196:106754.
7. Liang J, Qin Z, Xiao S, Ou L, Lin X. Efficient and secure decision tree classification for cloud-assisted online diagnosis services. IEEE Trans Dependable Secure Comput. 2019;18(4):1632–44.
8. Ghiasi MM, Zendehboudi S. Application of decision tree-based ensemble learning in the classification of breast cancer. Comput in Biology and Medicine. 2021;128:104089.
9. Lingjun H, Levine RA, Fan J, Beemer J, Stronach J. Random forest as a predictive analytics alternative to regression in institutional research. Pract Assess Res Eval. 2020;23(1):1.
10. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
11. Ning B, Junwei W, Feng H. Spam message classification based on the Naive Bayes classification algorithm. IAENG Int J Comput Sci. 2019;46(1):46–53.
12. Katare D, El-Sharkawy M. Embedded system enabled vehicle collision detection: an ANN classifier. In: 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC); 2019. p. 0284–0289.
13. Campus K. Credit card fraud detection using machine learning models and collating machine learning models. Int J Pure Appl Math. 2018;118(20):825–38.
14. Varmedja D, Karanovic M, Sladojevic S, Arsenovic M, Anderla A. Credit card fraud detection-machine learning methods. In: 18th international symposium INFOTEH-JAHORINA (INFOTEH); 2019. p. 1-5.
15. Khatri S, Arora A, Agrawal AP. Supervised machine learning algorithms for credit card fraud detection: a comparison. In: 10th international conference on cloud computing, data science & engineering (Confluence); 2020. p. 680-683.
16. Awoyemi JO, Adetunmbi AO, Oluwadare SA. Credit card fraud detection using machine learning techniques: a comparative analysis. In: International conference on computer networks and Information (ICCNI); 2017. p. 1-9.
17. Seera M, Lim CP, Kumar A, Dhamotharan L, Tan KH. An intelligent payment card fraud detection system. Ann Oper Res 2021;1–23.
18. Guo S, Liu Y, Chen R, Sun X, Wang X. X, Improved SMOTE algorithm to deal with imbalanced activity classes in smart homes. Neural Process Lett. 2019;50(2):1503–26.
19. The Credit card fraud [Online]. https:// www. kaggle. com/ mlg- ulb/ credi tcard fraud
20. Kasongo SM. An advanced intrusion detection system for IIoT based on GA and tree based algorithms. IEEE Access. 2021;9:113199–212.