BioGecko

Vol 12 Issue 03 2023
ISSN NO: 2230-5807

The Effectiveness of Detecting Credit Card Fraud Using Smote and Ada Boost

J.Deva Paul[1], D.Naga Raju[2], K.Govardhan[3], B.Mahendra[4], Nivetha R [5]*

Department Of Computer Science and Engineering, Bharath Institute of Science & Technology affiliated to Bharath Institute of Higher Education and Research, Chennai, Tamilnadu, India.

*Corresponding authors mail id: nivetha.cse@gmail.com

ABSTRACT:

The most popular payment method in recent years is the credit card. As technology advances, the prevalence of fraud instances rises as well, necessitating the creation of an algorithm for detecting fraud that can accurately locate and terminate fraudulent activities. This study recommends several machine learning-based categorization techniques, such as logistic regression, to deal with the severely unbalanced dataset. The development of technologies like Online card transactions have increased daily as a result of e-commerce and financial technology (FinTech) apps.. As a result, credit card fraud has increased, posing a threat to banks, merchants, and card issuers. Therefore, creating systems to guarantee the safety and accuracy of credit card transactions is crucial. Using skewed real-world datasets compiled. In this study, we develop an ML-based framework for detecting credit card fraud using data from European credit cardholders. We used the Artificial Minority over-sampling technique to resample the dataset in order to addressing the issue of class inequality (SMOTE). The machine learning (ML) methods listed below were employed to rate this framework: Decision Tree (DT) , Data from fake credit card fraud was used to test the suggested method that was extremely skewed, which further confirmed the findings of this analysis. The experiments' findings demonstrated that using Ada Boost increases the efficiency of the advised solutions. Furthermore, the results generated by the improved models were superior to those of prior models techniques. The accuracy, precision, recall, and f1 score will be determined at the end of this study project. Credit card fraud targets are easy to find and contact. Online payment options have multiplied thanks to e-commerce and other websites, increasing the risk of online fraud. Which will allow for the analysis and extraction of behavioral patterns from past consumer transaction data. This classifies cardholders based on how much money they spent on each transaction. The sliding glass door approaches, then aggregate. Then, merge the transactions made by cards from various groups using the sliding window method to get each group's distinctive behaviour. Then, different classifiers are individually trained using the groupings. In order to recover each group's distinct behaviour, aggregate the transactions made by cards from various groups using the sliding window method. Following that, various classifiers are trained using the groups independently.

Keywords: *E-Commerce, Data Phishing, FDS, Area under the Curve (AUC), Matthews Correlation Coefficient (MCC).*

1. INTRODUCTION:

Finding fraudulent credit card transactions is the work [1]. Classifying the fraudulent and legitimate transactions is necessary to accomplish this. The primary goal is to create a fraud detection program that, using machine learning-based classification methods, quickly and accurately identifies fraudulent transactions [2]. As technology develops quickly, less money is paid in cash and more money is sent online, which makes it easier for fraudsters to conduct anonymous transactions [3]. For some purchases, all he needs are the card details, and the user might not be aware if their credit card information has been compromised [4]. Financial fraud has increased recently as a result of the emergence of new technologies and paradigms, especially those in the financial technology (FinTech) and e-commerce sectors. [5]. Due to this, credit card transactions have grown of these technologies'

development. As a result, the quantity of credit card-related financial fraud instances has rapidly increased [6]. When a criminal uses a credit card in an unauthorized or undesired that type of behavior is referred to as credit card fraud. This occurs when credit card authentication information is fraudulently obtained, for as via intercepting an online transaction or by making a clone of an already-used card. [7]. In order to execute credit card transactions, Implementing this is essential solutions that can guarantee the security and integrity of all systems by identifying credit card fraud [8]. In this study, we apply using real-world data to evaluate methods for identifying Machine learning (ML) is used to commit credit card fraud [9]. Extra Tree (ET), Extreme Gradient Boosting (XGBoost), Logistic Regression (LR), Decision Tree, and Support Vector Machine (SVM) are further ML techniques that were taken into account in this study (DT). Each of these ML techniques was assessed individually for efficacy and classification quality [10]. Additionally, each technique was combined with the Adaptive Boosting (Ada Boost) algorithm to boost their robustness. As a result of the growth of technologies like e-commerce and financial technology (FinTech) apps, online card transactions are becoming more and more commonplace every day. Banks, retailers, and card issuers have all been impacted by the rise in credit card theft [11]. Therefore, in order for credit card transactions to remain accurate and confidential, it is essential to develop systems that protect them and Extra Tree were the machine learning (ML) methods that were utilized to evaluate this framework (ET). With these ML algorithms, the Adaptive Boosting (Ada Boost) technique was applied precision, the Area Under the Curve and the Matthews Correlation Coefficient (AUC) (MCC) (AdaBoost) technique were combined [12]. The Area under the Curve and Adaptive Boosting (AdaBoost) methods were merged with the Extra Tree (ET), Extreme Gradient Boosting (XGBoost), Logistic Regression (LR), Decision Tree, (AUC), the recall, and the precision of the models were used to assess their performance (AUC).

2. METHODOLOGY:

2.1 Module Description:

- Data Collection
- Data Preparation
- Model Selection
- Fraud Detection using Random Forest Algorithm
- Accuracy on test set
- Saving the Trained Model

2.1.1 Data Collection:

Data collection is required to provide answers to certain research questions, test hypotheses, and collect data on variables of interest in a methodical and defined manner. The systematic and planned gathering and measuring of information on relevant variables is known as data collection. It tries to answer specific research queries, test hypotheses, and assess results.

2.1.2 Data Preparation:

The data will be transformed. By eliminating any missing data and some columns. The titles of columns we wish to maintain or keep will first be listed. After that, we drop or eliminate all columns save for the ones we wish to keep. Finally, we eliminate or remove the rows from the data collection that contain missing values.

2.1.3 Model Selection:

Two datasets are required when building a machine learning model: one for training and the other for testing. But there is only one left now. Let's divide this in two according to an 80:20 ratio. The data-frame will also be split into a feature column and a label column. We imported the sklearn train test split function here. Use it to divide the

dataset after that. Additionally, test size = 0.2 divides the dataset into two parts: 20% for the test and 80% for the train. The dataset is divided using a random number generator that is seeded by the random state option. Four datasets are returned by the function. They were given the labels train x, train y, test x, and test y. If we look at the shape of this dataset, we can see how the data was split and fitted to numerous decision trees using RandomForestClassifier.

2.1.4 Fraud Detection using Random Forest Algorithm:

The system must determine whether fraud has taken place in the transaction in this module. Additionally, it must inform the user of the outcome.

2.1.5 Accuracy on test set:

We got an accuracy of 0.93% on test set.

2.1.6 Saving the Trained Model:

The first thing to do Utilize a library like pickle to save your trained and tested model as an.h5 or.pkl file when you're ready to use it in a production-ready environment. Make sure Pickle is set up in your environment. Now that the module has been imported, let's dump the model into a.pkl file.

2.2 Computer Architecture

The source is used to retrieve the credit card dataset, followed by cleaning and validation operations are carried out on it. These operations include removing duplicate data, filling in blank columns, and transforming required variables into classes or factors. The data is then divided into two categories: the training dataset and the test data set.

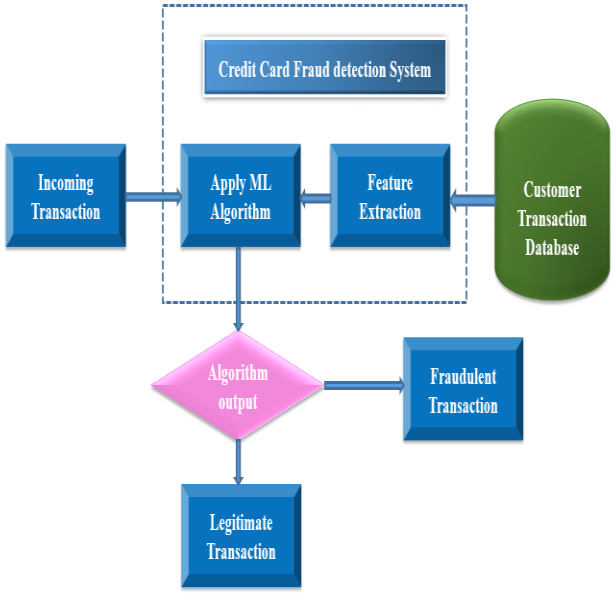

Figure 1: Credit Card Transaction System

The initial sample has now been randomly divided into the train and teat datasets. Finally, the Classifier algorithm is used to train our models. On the Python Natural Language Toolkit library, we use the categorize module

as shown in above fig.1. We make use of the acquired labelled dataset. The models will be assessed using the remaining labelled data we have. Pre-processed data was categorized using a few machine learning methods. Random forest classifiers were selected. These algorithms are widely used in jobs involving text classification.

3. RESULTS & DISCUSSIONS

3.1 Existing System:

The current detecting credit card fraud technique only identifies fraud after it has already occurred. When a consumer notices an inconsistency in a transaction, the existing system maintains a vast quantity of data, and the customer files a complaint. At that point, the fraud detection system begins to function. It seeks to determine whether fraud has actually occurred before using transactions to use the fraud detection system created by the master and visa cards. A classification approach using machine learning, with Credit Card Fraud Detection as the foundation. In the event of the current system, there is no confirmation of the recovery of fraud and customer satisfaction. Intrusion detection to track fraud location, etc. Secure electronic system that examines the actions of authorized users. Mechanisms for data mining to sort and prepare user data. If your card is detected as having suspicious activity, online alerts are provided. If any odd charges appear on the card, the issuer will alert you. Credit scores serve several purposes, and they grow when a person responsibly utilizes a credit card and makes on-time payments without going over their credit limit or paying late. One can reduce the amount of interest charged by the card issuer by making timely payments and maintaining a low card balance. Eco-friendly, requires no physical presence, uses advancing technology, and saves time. Each payment system has a cap on the maximum balance that can be held in an account, the daily transaction limit, and the output volume. You are unable to access your online account if the Internet connection drops. The threat is rather low if you abide by the security guidelines. The worst case scenario is when a processing company's system is compromised and personal information about cards and their owners is exposed. The payment system's database contains details about every transaction, including the sum, the date, and the beneficiary. Additionally, it implies that the intelligence service has access to this data. This can occasionally lead to fraudulent behaviour.

3.2 Proposed System:

Building models to determine whether a credit card transaction is fraudulent is the major objective of this research. We'll try supervised learning as a strategy. The suggested system uses the Random Forest Algorithm for dataset classification and regression. First, a dataset for credit cards will be gathered, and only then will it be analysed. After dataset analysis, dataset cleaning is required. Cleaning is required to remove all of the duplicate and null values, which are present in every dataset by nature. The dataset must then be split into two groups for comparison and analysis: the trained dataset and the testing dataset. The Random Forest Method must be applied after the dataset has been divided; this algorithm will give us more accurate information on credit card fraud transactions. The Random Forest Algorithm will be used to partition the dataset into four groups, and the classification outcomes will be displayed as a confusion matrix. The classification of the aforementioned data will be used to conduct performance analysis. The accuracy of transactions involving credit card fraud can be obtained from this study, and the results will then be graphically displayed. If your card is detected as having suspicious activity, online alerts are provided. If any odd charges appear on the card, the issuer will alert you. Credit scores serve several purposes, and they grow when a person responsibly utilizes a credit card and makes on-time payments without going over their credit limit or paying late. One can reduce the amount of interest charged by the card issuer by making timely payments and maintaining a low card balance. Eco-friendly, time-saving, and requires no physical presence while using advancing technology.

3.3 Algorithms:

3.3.1 Random Forest Algorithm

- The most potent and widely used machine learning algorithm is called the random forest classifier.
- It makes decisions via a variety of separate decision trees.

- It can perform both classification and regression tasks because of its simplicity and diversity.

Steps

1. Select N records at random from the dataset.
2. Create a decision tree using the n records provided.
3. Repeat steps 1 and 2 until you have the amount of trees you want in your algorithm.
4. In case of a regression problem, for a new record, Each and every tree in the forest forecasts a value for Y (output).

3.4 Data Flow Diagram:

A system's data processing and transmission are shown in a two-dimensional diagram called a data flow diagram (DFD). The graphical representation shows how each data source engages with different data sources to produce a common outcome. One must; in order to create a data flow diagram.

- Determine the external inputs and outputs
- Find the relationships between the inputs and outputs.
- Explain how these linkages relate to one another and what they lead to using the graphics.

3.3.2 Role of DFD:

- Both programmers and non-programmers can understand the documentation help. DFD only takes into what processes accomplish.

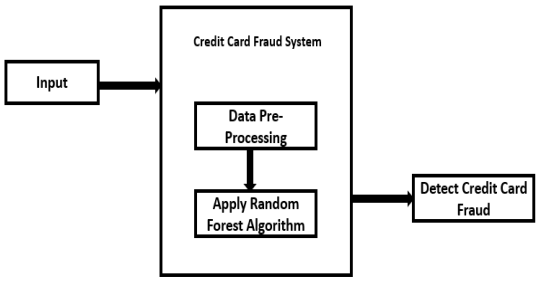

Figure 2: Data Flow Diagram to detect fraud system

- A physical DFD assumes who processes the data and where it flows.
- By monitoring the process's input data and tracking how they change when they leave, it enables analysts to identify as shown in fig.2 areas of interest inside the company and examine them.

3.5 Functional Requirements:

It is a prerequisite for the software products' technical specifications. It is the initial phase in the procedure of requirement analysis and includes the functional, performance, and security requirements for specific software systems. The system's performance is mostly dependent on the high-quality hardware utilized to run the program with the required capabilities.

Usability

It details how simple a system must be to use. Short or long questions can be asked with ease because the Porter stemming algorithm prompts the user's chosen response.

Robustness

It describes a program that operates effectively both in typical and unexpected circumstances. It is the user's capacity to handle execution faults for pointless requests.

Security

Security is the condition of allowing restricted access to resources. Unauthorized users cannot access the system because of the system's strong security measures.

Reliability

It is the likelihood of how frequently the software malfunctions. MTBF is a common unit of measurement (Mean Time Between Failures). The prerequisite is necessary to make sure that processes are carried out entirely and without interruption. It is capable of supporting any weight, enduring indefinitely, and even overcoming failures.

Compatibility

The version above all web browsers supports it. Any web server, including local host, can be used to make the system real-time.

Flexibility

The project's adaptability is set up in a way that allows it to function in many surroundings while being used by various users.

Safety

Safety is a precaution done to avoid problems. Each enquiry is handled securely without revealing any personal information to third parties.

3.6 Non- Functional Requirements:

Portability

The usability of the same software in various settings. Any operating system can be used to run the project.

Performance

The resources needed, the time interval, the throughput, and every other aspect of the system performance are determined by these criteria.

Accuracy

The information is retrieved quickly and with great accuracy thanks to the requesting query. The technology provides a high level of effective security.

Maintainability

The project is straightforward since it is simple to make updates without compromising its stability. In essence, maintainability refers to how simple it is to maintain the system. It refers to how simple it is to analyses, modify, and test an application as well as to manage a system. The project's capacity to be maintained is simple as further updates can be easily done without affecting its stability.

3.7 Input Design:

A connection between the user and the information system is created through the input design. Developing requirements and techniques for data preparation is part of the process of converting processing-friendly format for transaction data. Users can enter the data directly into the system or a computer can read it from a written or printed document to accomplish this. The input process is designed with the goal of minimizing input requirements, errors, delays, and unnecessary steps, while maintaining a clear workflow. The entry of the data protects its privacy, usability, and security. Input Design kept the following things in mind: Which information should be supplied as input

- What order or coding should be used for the data
- The dialogue serves as a direction for operating staff involvement.

- Methods for creating input validations and what to do in the event of an error.

3.8 Output Design:

A quality output is one that displays the information clearly and satisfies the end user's needs. Any system's processing outcomes are transmitted via outputs to users and other systems. How information will be displaced for both immediate demand and the hard copy output? It is decided during output design. It serves as the user's main and most direct information source. A system's ability to engage with tools that support user decision-making is enhanced by effective and clever output design.

An information system's output format should achieve one or more of the aforementioned goals. Share specifics on recent events, the current situation, or predictions for the future. Signal significant occurrences, chances, issues, or cautions.

- Start an action.
- Verify an activity.

4. CONCLUSION

In this essay, we have discussed how Random Forest can be used to detect unauthorized online credit card transactions. Additionally scalable for processing huge volumes of transaction data, the suggested Fraud Detection System. The Random Forest-based credit card fraud detection system doesn't have a difficult process for doing a fraud check, just like the current system. In comparison to the current method, the suggested fraud detection system generates results more rapidly and honestly. The Random Forest aims to simplify detection processing while attempting to reduce complexity.

FUNDING: No funding sources

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The encouragement and support from Bharath Institute of Higher Education and Research, Chennai, Tamilnadu, India is gratefully acknowledged for providing the laboratory facilities to carry out the research work.

REFERENCES:

1] R. R. Subramanian, R. Ramar, “Design of Offline and Online Writer Inference Technique”, International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 2S2, Dec. 2019, ISSN: 2278-3075

[2] Subramanian R.R., Seshadri K. (2019) Design and Evaluation of a Hybrid Hierarchical Feature Tree Based Authorship Inference Technique. In: Kolhe M., Trivedi M., Tiwari S., Singh V. (eds) Advances in Data and Information Sciences. Lecture Notes in Networks and Systems, vol 39. Springer, Singapore

[3] Joshva Devadas T., Raja Subramanian R. (2020) Paradigms for Intelligent IOT Architecture. In: Peng SL., Pal S., Huang L. (eds) Principles of Internet of Things (IoT) Ecosystem: Insight Paradigm. Intelligent Systems Reference Library, vol 174. Springer, Cham

[4] R. R. Subramanian, B. R. Babu, K. Mamta and K. Manogna, "Design and Evaluation of a Hybrid Feature Descriptor based Handwritten Character Inference Technique," 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Tamilnadu, India, 2019, pp. 1-5.

[5] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.-E. Portier, L. He-Guelton, and O. Caelen, ‘‘Sequence classification for credit card fraud detection,’’ Expert Syst. Appl., vol. 100, pp. 234–245, Jun. 2018.

[6] U. Fiore, A. D. Santis, F. Perla, P. Zanetti, and F. Palmieri, ‘‘Using generative adversarial networks for improving classification effectiveness in credit card fraud detection,’’ Inf. Sci., vol. 479, pp. 448–455, Apr. 2019.

[7] Credit Card Fraud Dataset. Accessed: Sep. 4, 2019. [Online].Available: https://www.kaggle.com/mlg-ulb/creditcardfraud/data.

[8] I. Mekterović, L. Brkić, and M. Baranovi, ‘‘A systematic review of data mining approaches to credit card fraud detection,’’ WSEAS Trans. Bus. Econ., vol. 15, p. 437, Jan. 2018.
[9] M. A. Al-Shabi, ‘‘Credit card fraud detection using auto encoder model in unbalanced datasets,’’ J. Adv. Math. Comput. Sci., vol. 33, no. 5, pp. 1–16, 2019.
[10] S. V. S. S. Lakshmi and S. D. Kavilla, ‘‘Machine learning for credit card fraud detection system,’’ Int. J. Appl. Eng. Res., vol. 13, no. 24, pp. 16819–16824, 2018.
[11] X. Zhang, Y. Han, W. Xu, and Q. Wang, ‘‘HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture,’’ Inf. Sci., May 2019. Accessed: Jan. 8, 2019.
[12] N. Carneiro, G. Figueira, and M. Costa, ‘‘A data mining based system for credit-card fraud detection in e-tail,’’ Decis. Support Syst., vol. 95, pp. 91–101, Mar. 2019.