Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

A Research Paper on the Detection of Credit Card Fraud Using Machine Learning

[1]Akhil Sharma

M.Tech Scholar

Jaipur National University Jaipur

[2]Sachin Jain

Assistant Professor

Jaipur National University Jaipur

[3]Dr. Deepak Dembla

Professor

JECRC University Jaipur

Article Info
Page Number: 1635-1644
Publication Issue:
Vol. 72 No. 1 (2023)

Article History
Article Received: 15 October 2022
Revised: 24 November 2022
Accepted: 18 December 2022

ABSTRACT

Financial organisations, businesses, and customers are all very concerned about credit card theft. Both financial losses and reputational harm could be severe. By analysing huge datasets and finding patterns and anomalies that may signal fraudulent actions, machine learning is a potent technique that may be used to detect credit card fraud. Algorithms that use machine learning can spot unexpected spending patterns, such as purchases that diverge from a customer's customary patterns or those that are made in peculiar places. They can also spot suspect patterns of behaviour, such multiple transactions occurring quickly or transactions using credit card data that has been compromised.

Overall, machine learning is an effective method for detecting credit card fraud that can save financial institutions and retailers from loss of money and reputational harm. Machine learning is projected to become an even more crucial technique for credit card fraud detection in the years to come due to the rising accessibility of massive datasets and sophisticated analytics tools.

1. Introduction

Fraud is the abuse of a system by a for-profit organization that may not lead to it immediate legal consequences. Fraud is a critical problem in many industries such as healthcare, banking, insurance and telecommunications. A fraudulent minority creates a great burden on society by financing fraudulent transactions. Any company seeking to correct fraudulent transactions in the above and probably many other companies, is designated as a fraud detection process. Due to the complexity and enormous size of modern business systems, criminals can find security holes and use them to access data or deceive someone Even if fraud is discovered by authorities and security rules are taken care of, criminals will look for and find other ways to commit fraud and thus change their behavior over time. Manual

detection by human experts is very expensive even to correct frauds that have occurred; cannot detect all types of fraudulent transactions; fraud cannot be detected during this test, and changes and trends in fraudulent behavior cannot be detected.

Average fraudsters are people who don't belong to an organised crime group and only occasionally commit fraudulent crimes. Even though these fraudsters pose a detection risk, organised or individual criminal fraudsters are more likely to do more damage to the impacted business system. These fraudsters operate in an organised manner, frequently 3 conduct identity theft, and alter their behaviour over time to evade detection by detection systems and new laws. It is extremely expensive to manually review all transactions and activity, especially for large firms. Application fraud and transaction fraud are other subcategories of fraud. While fraudulent identity information is involved in the application case, the transactional case involves criminals abusing legitimate user and account information.

The difficulties associated with fraud detection can be summed up as follows:

• Class distributions imply that there are varying ratios of lawful to illicit transactions.

• A variety of fraud schemes can harm a company.

• Different types of fraud exhibit various behavioural traits,

• Swindlers adapt their behaviour and fraud techniques to avoid being caught by any new detection system.

Kinds of credit card fraud

- Stolen cards
- Cards that never arrived
- Theft of identity
- Counterfeit cards

Systems for detecting fraud in real time and almost real time

Real-time fraud detection is matched by the blocking time. It seeks to approve or deny transactions. This procedure must be quick and extremely accurate. In fact, it must only reject transactions that are Topping for the Cardblocker PIN reading terminal 8 highly likely to be fraudulent; otherwise, it will interfere with cardholders' normal purchasing patterns. The majority of transactions that real-time fraud detection systems block are fraudulent transactions; nonetheless, many fraudulent transactions go undetected. Precision, not recollection, is the real-time fraud detection system's top priority.

2. Research Techniques

I think that the CRISP-DM approach will make it simpler to get effective and premium results since it involves the project in the entire path, starting with understanding the business

and data and continuing with data preparation, modelling, and evaluation to ensure the model is working properly.

Step 1 - Business Knowledge Since using a credit card to make payments is similar to taking out a loan, many people are experiencing the problem of having their credit lines violated by those who are acting fraudulently, which is having an 12 impact on their daily lives, as was previously mentioned. Many individuals will have significant loans that they are unable to repay if the issue is not resolved, which will make their lives difficult and prevent them from being able to buy required items. In the long term, not being able to pay back the debt might result in them being sent to jail. In essence, the issue raised is how to identify credit card fraudsters' transactions in order to prevent security breaches and protect client safety.

Step 2 - Analysing Information A high-quality dataset was essential for the information analysing phase because the model is based on it. The dataset was investigated by taking a closer look at it, which provided the knowledge required to confirm the dataset's quality, in addition to reading the description of the entire dataset and each attribute. Additionally, it's crucial to have a dataset with a variety of mixed transaction types. Finally, identifiers to clarify the basis for the classification of the transaction type, including fraudulent and real as well as a class to clarify the type of transaction. During my search for the best dataset, I made sure to adhere to each of those guidelines.

Step 3 - Extraction of Data The preparation phase of the dataset starts after selecting the most appropriate dataset; it involves selecting the desired attributes or variables, cleaning the dataset by excluding Null rows, deleting duplicate variables, treating outliers as necessary, and transforming data 13 types to the desired type. Data merging, where two or more attributes are combined, can also be performed. All of those changes provide the desired outcome, which is to prepare the data for modelling. Given that the dataset used for this project had neither missing nor duplicate variables, nor did it require merging, none of the adjustments previously indicated had to be made. However, the types of the data had to be changed in order to create graphs. This required using the programme Sublime Text in addition to importing the data into Weka and performing analysis.

Step 4 – modeling During the modelling phase, four machine learning models were developed: KNN, SVM, Logistic Regression, and Naive Bayes. Later in the article, a comparison of the results will be provided so that readers may choose which approach is best for identifying credit card fraud. The dataset is divided into two parts: the training set, which will comprise 70% of the dataset, and the testing set, which will comprise the remaining 30%. Weka was used to generate all four models, with only KNN and Naive Bayes being developed in R. Both tools' visualisations will be offered.

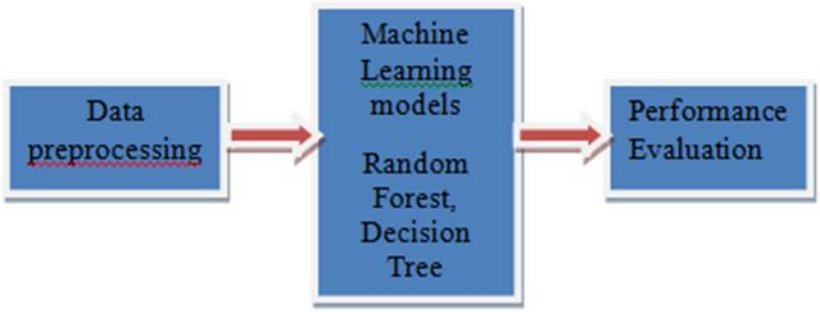
Step 5 - Assessment and Implementation In order to determine which model is the best and most appropriate for identifying fraudulent credit card transactions, the models' accuracy, efficiency, and suitability will all be given in the final step of the assessment process.

Analysis Of Important Literature

There have been several studies conducted on credit card fraud. To locate previously published studies, we shall summarise some of the publications in this review. The use of (supervised techniques) like Logistic Regression, Decision Tree, Random Forest, and XGBoost as well as (unsupervised methods) like K-Means Clustering and Autoencoder in Keras was covered in this part. Researchers like Maniraj(2019), Dornadula (2019), , Azhan (2020), Joshi(2020), Priya & Saradha (2021), Shirgave(2019) and Roy(2021), have identified supervised and unsupervised method of machine learning as the most common methods.

Methodology

The methods used in this study to distinguish between legitimate and fraudulent transactions is covered in this chapter. The steps taken in this investigation are shown in Figure



Sizing Of Functions

The range of distinct variables of a dataset is normalised at that point of the data preparation approach. It is located between 0 and 1, or near 0, depending on the scaling method used. Some machine learning algorithms may be overlooked or tampered with if input variables have enormous values that apply to additional input variables. With the Robust Scaler approach, commonly referred to as robust standardization.

Decision Tree

Modern data analysis methods are employed by machine learning systems. The decision tree model is the most often used method in machine learning applications. Especially when used to collect and analyse massive volumes of data, decision trees function incredibly rapidly and intelligently. Just by controlling a transaction in a particular way based on the features extracted from the data, the decision tree model operates. It starts with a basic root question and branches, using the details to create specific components that ultimately result in endpoints or the tree's leaves. When continuous data partitioning is dependent on a specified parameter, decision trees are nonparameterized supervised learning techniques that may be utilised for categorisation and regression applications. Nodes, Edges, and Leaf nodes make up its structure.

Random Forest

Random Forest is a widely used machine learning method. It is a method used to address classification and regression issues. The "forest" is a collection of a very large number of different decision trees. Every different tree predicts a class. Any class with the most votes is

taken into consideration for prediction. As a result, the method uses a bagging strategy to create a collection of decision trees which will eventually form a forest. This method's advantage is that it runs the model rapidly and intelligently balances the mistakes without the requirement for feature selection. The drawback of this method is that it is readily able to identify fraud since it is cognizant of data with a wide range of values and characteristics with more values. The forest which the approach creates is referred to as decision tree outfit, which consisted which is typically trained using a technique called the bagging process, a form of Bootstrap approach to a significant variance approach utilised in machine learning. Algorithms that aggregate numerous models into a single package include bagging and random forests. Both algorithms work well for various kinds of predictive modelling issues. It is one of the top algorithms for fraud detection in the financial system. The benefit of using random forest is the way it can be applied to problems involving regression as well as classification. Finding the best feature among all characteristics for modelling becomes more important when using the Random method since it always adds unpredictability, especially when dividing the node. The random forest higher parameter increased the model's capacity for prediction or accelerated model execution. One of the problems with machine learning modelling is the overfitting problem. Even so, a classifier based on random forests is helpful since it can generate many of forest trees and won't overfit the model.

Data Assessment

Preprocessing of Data

The dataset preparation was easy due to the fact that are no NAs or duplicate variables. The first change that took place to enable the dataset to be opened in the Weka programme was to change the kind of the attribute known as class compared to numbers to Class and to recognise the class as 1,0 using the programme Sublime Text. To be able to generate both the structure and the visualisation, another change was made to the R program's type.

Modeling of the Data

Each of the models were built using both Weka and R after making sure the data was prepared for modelling. Weka was the only tool used to generate the SVM model; R and Weka were used to create the KNN, Logistic Regression, and Naive Bayes models.

KNN

The K-Nearest Neighbour technique (KNN) is a supervised machine learning method that may be used in both classification and regression scenarios (Mahesh, 2020).Two Ks, K=3 and K=7, including data obtained from Weka and R, were utilised to determine the optimal KNN model. K=3 I made the decision to build two models with K=3 and K=7 while building the KNN model. Figure displays the R-created model that had a 99% accuracy rate, correctly identified 90,708 transactions, and missed 145. In terms of the Weka programme, the model had an accuracy score of 99.94% but misclassified 52 transactions. The average accuracy is 99.89% since there are various degrees of accuracy.

Confusion matrix and statistics

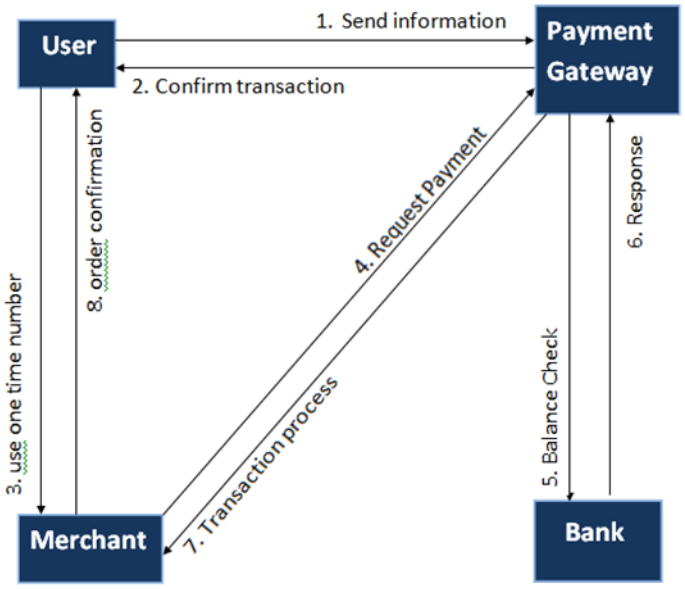
For k=3

Prediction	Reference: genuine	Reference: fraud
genuine	90708	145
fraud	0	15

Naive Bayes

The major applications of Naive Bayes include clustering and classification, based on a conditional likelihood of occurring. It is a classification technique that believes that the existence of a specific characteristic inside a class is unconnected to the existence of any other feature (Mahesh, 2020). The second model produced by R is called Naive Bayes, and following figure displays how it performed. It misclassified 2,051 transactions, classifying 33 fraudulent transactions as nonfraudulent and 2018 fraudulent transactions as nonfraudulent, for an accuracy score of 97.77%. The Weka-created Naive Bayes model's accuracy is slightly different; it is 97.73%, and there have been 1,938 occurrences of misclassification.

Assessment and Implementation



The assessment and implementation phase of the CRISP-DM method is the last stage. As shown in table below, all models are evaluated to determine which one is most effective in detecting unauthorised transactions with credit cards. The total number of instances that are accurately predicted is known as accuracy. Accuracy is represented by a confusion matrix that includes True Positive, False Positive, True Negative and False Negative values. The fraudulent transactions that True Positive reflects are those that the model properly identified as fraudulent. The not-fraudulent transactions that the model accurately identified as such are represented by True Negative. False positive, the third rating, denotes transactions that are malicious but were mistakenly classed as legitimate. The confusion matrix is shown in table

below, and the last category is False Negative, which are transactions that were not fraudulent but were still recognised as such.

predicted	+	-
+	TP	FN
-	FP	TN

Confusion Matrix

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Model		Accuracy
KNN	K=3	99%
	K=7	99.89%
Naïve Bayes	Naïve Bayes	97.77%

Proposed System

The suggested system's fundamental prerequisites for implementation include a web connection and a smartphone or computer application. Each user must register with the system, providing the necessary information and answering three security questions in case they forget their username or password. Each user must also create a password that has an uppercase letter, a special character, and a number. The user's password is also used by our system as an encrypted link among the user and the company that issued the card. To manage online purchases without using a physical credit card, the cardholder should check in to the system using his or her credentials to oversee any payment operations.

Deployment

1. Mobile device or personal computer. To obtain the only once credit card number and give the server with transaction information, like the amount of the transaction and the merchant's website, the user must either use a smart device or a desktop computer. In order to complete any transaction, the user must also confirm his or her identity by inputting the login and password.

2. The majority of system actions are carried out on the server, such as authenticating the user, saving transaction information in a database, using fraud detection algorithms to assess the risk of the transaction, and responding appropriately.

3. A database. All user data and credit card transactions should be stored in the system's database.

4. Access to the internet. To complete the intended transaction, both the client and the host computer must both be online.

the suggested system makes advantage of the present network infrastructure and security measures, like SSL (secure socket layer). The system requires that each user register an account, and each account's data should be saved in the database for use in future login processes and verification needs.

Component for Detecting Fraud using machine learning In order to increase the security of our system with a number of characteristics, the suggested system incorporates the machine learning identification of fraud method : 1. Place (longitude and latitude) 2. I.P. address (if a PC was used for the transaction). 3. IMEI number (if a smartphone was used for the transaction) 4. Time (the number of daylight hours) 5. The time difference between each transaction (the number of days between each transaction) 6. The sum being exchanged As a second security layer, the suggested system will implement the L.R. algorithm depending on the characteristics. Every transaction is kept on the server in a database. The L.R. algorithm may therefore create a pattern of spending based on the history of each cardholder's transactions. Based on a judgement made by the L.R. algorithm, our system is built to reject suspicious transactions.

Conclusion

Due to the rise in financial crime and problems with database breaches, credit card security is crucial for both users and credit card issuers. To enhance the security of the present credit card systems, this study covers this issue. By suggesting a secure way utilising one-time credit card numbers, we have created a functional plan to circumvent the security difficulties with permanent credit card numbers. To establish a reliable online payment system, this study integrates the only once credit card strategy with machine learning algorithms. Our method prevents consumers' money from being taken away by fraudsters, even if it adds a little more work for the client by asking for an only once credit card number each time before moving on to the standard procedure of making a purchase via the internet.

References

1. Bhanusri, A., Valli, K. R. S., Jyothi, P., Sai, G. V., & Rohith, R. (2020)
2. Credit card statistics. Shift Credit Card Processing. (2021, August 30) https://shiftprocessing.com/credit-card/
3. Adepoju, O., Wosowei, J., lawte, S., & Jaiman, H. (2019) https://ieeexplore.ieee.org/document/8978372/
4. Alenzi, H. Z., & Aljehane, N. O. (2020) Fraud Detection in Credit Cards using Logistic Regression (thesai.org)
5. Gupta, A., Lohani, M. C., & Manchanda, M. (2021) Financial fraud detection using naive bayes algorithm in highly imbalance data set: Journal of Discrete Mathematical Sciences and Cryptography: Vol 24, No 5 (tandfonline.com)
6. Itoo, F., Meenakshi, & Singh, S. (2020) Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection | SpringerLink
7. Kiran, S., Guru, J., Kumar, R., Kumar, N., Katariya, D., & Sharma, M. (2018)

8. Jain, Y., NamrataTiwari, S., & Jain, S. (2019)
9. Mahesh, B. (2020)
1. 10.Malini, N., & Pushpa, M. (2017)
2. A Practical Implementation of Face Detection by using Viola Jones Algorithm in MATLAB GUIDE R Yagnik, A Jangid, S Jain
3. International Journal of Engineering Research & Technology (IJERT) IJERT … 6 2014
4. Review of some recent findings for productivity improvement using line balancing heuristic algorithms P Shukla, S Malviya, S Jain
5. International Journal of Innovative Research in Technology 5 (6), 83-90 1 2018
6. Experimental investigation on the process parameter of rapid prototyping technique for the improvement of disposal glass material replace by PLA
7. S Choudhary, S Malviya, S Jain
8. IJARIIT 4 (3) 1 2018
9. Automatic Railway Track Inspection for early warning using Real time image processing with GPS X Verma, A Jeewan, S Jain, M Vats
10. International Journal on Recent and Innovation Trends in Computing and … 1 2015
11. Development of Top K-Association Rule Mining for Discovering pattern in Medical Dataset A Sharma, A Sangwan, B Thankchan, S Jain, V Singh, S Saurabh
12. European Journal of Molecular & Clinical Medicine 7 (4), 2020 1
13. Summarizing Text Using Lexical Chains P Jain, S Jain International Journal on Recent and Innovation Trends in Computing and … 1
14. A Neoteric Strategy of Hill Cipher for Analysis of Degenerate Matrices KeyS Jain, MK Arya
15. Journal of Algebraic Statistics 13 (1), 194-198 2022
16. ARTIFICIAL INTELLIGENCE IN EDUCATION S Jain Purakala Vol 31 Issue 1 31 (1), 16-25 2022
17. Detection in Wireless Network S Jain International Journal of Research and Analytical Reviews (IJRAR) 9 (2), 643-646,2022
18. AREVIEW PAPER CLOUD COMPUTINGSJ Amita Kashyap Purakala Vol 31 Issue 1 31 (1), 13-15 2022
19. THE ROLE OF CRYPTOGRAPHY TOWARDS NETWORK SECURITY SKD, Mr. Sachin Jain Purakala Vol 31 Issue 1 31 (1), 1-5 2022
20. Study of Existing IOT Frameworks for Building Future Automation System and Connected Objects, AK Sachin Jain,Vikram Singh, The Ciência & Engenharia - Science & Engineering Journal 10 (Issue: 1), 19-23 2022
21. Improving the performance of heterogeneous hadoop clusters using the map reduce big data algorithm, SG Sachin Jain, Amita Kashyap, Sheetal Kumar Dixit
22. International Journal of Advanced Science and Technology 28 (2019), 740-748 2019
23. Comparative Analysis of E-Governance Models SJ, D Bandil

24. International Journal on Future Revolution in Computer Science … 2019
25. A SURVEY OF CLOUD COMPUTING'S LIMITATIONS AND POTENTIAL SOLUTION KS Mr Sachin Jain Purakala Vol 31 Issue 1, 2022 31 (1), 26-30
26. Framework for ATM Card & Safety: Preventing ATM Transaction from Hackers and Frauds D Malviya, S Jain
27. International Journal on Recent and Innovation Trends in Computing and …
28. An Integrated Approach For Energy Efficient Routing Over Ad-Hoc Network Using Soft Computing A Kumar, S Jain, M Vats International Journal on Recent and Innovation Trends in Computing and …
29. Reliability Assessment and Feasibility Study of Software Metrics in Object Oriented Environment S Jain, SK Dixit, S Gautam, S Sharma, J Qureshi, V Singh
30. Benchmark Classification of Handwritten Dataset by New Operator R Ranjan, M Vats, S Jain International Journal on Recent and Innovation Trends in Computing and …