ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940www.ijsem.org

Vol 8, Issuse.4 Oct2020

CREDIT CARD FRAUD DETECTION USING RANDOM FOREST

[1]M Arya Bhanu,
[2]P V Sarath Chand,
[3]Erugu krishna,
[4]B Venkateswarlu,

Abstract

Real-world credit card fraud detection is the primary emphasis of the project. Credit card fraud has lately increased dramatically as a result of the amazing surge in the number of transactions. The goal is to get something without paying for it or to get money out of a bank account without authorization. All credit card issuers must now have effective fraud detection systems in order to reduce their losses. Making the business is a major difficulty since no cardholder or card must be present for a transaction to be completed.. Merchants are unable to determine if a consumer presenting their card is in fact the legitimate owner. It is possible to increase the accuracy of fraud detection by using the suggested technique, which makes use of a random forest. The random forest technique is used to analyse a data set and the current dataset of the user. Finally, improve the precision of the output data. The accuracy, sensitivity, specificity, and precision of the procedures are assessed. Processed characteristics are used to identify fraud, and a graphical model depiction is presented. The accuracy, sensitivity, specificity, and precision of the procedures are assessed.

Keywords:Fraud Detection, Random Forest, Credit Card.

INTRODUCTION

In credit card transactions, researchers have kept in mind a variety of methodologies to construct models based on artificial intelligence, data mining, fuzzy logic, and machine learning in order to identify fraudulent activity. Detecting credit card fraud is a challenging, but common, challenge. Machine learning was used to develop the credit card fraud detection in our suggested system. Machine learning is becoming more sophisticated. Machine learning has been shown to be effective at detecting fraud. During online transactions, a significant quantity of data is exchanged, resulting in a binary outcome: legitimate or fraudulent. Features are built into the bogus datasets. Customers' age and credit card balance are only two examples of these data pieces. There are hundreds of characteristics, and each one has a different impact on the likelihood of a transaction being fraudulent. Note that the machine's artificial intelligence, which is guided by the training set, generates the level at which each characteristic contributes to the fraud score. This level is not . chosen by a fraud analyst. Accordingly, a transaction that is made using a credit card will have the same fraud weighting as one that is made using a debit card. But if this decreased, the contribution level would also decrease at the same rate. These models are self-taught and do not need any further programming, such as a manual review. Classification and regression methods are used for credit card fraud detection in Machine learning. For online or offline fraud card transactions, we employ a supervised learning system such as the Random Forest method. An evolved variant of Decision Trees is Random Forest. Random forest outperforms all other machine learning algorithms in terms of performance and accuracy. A key goal of random forest is to alleviate the problem of feature space oversampling, which was discussed before. De-correlated trees are pruned by setting a stopping point for node splits, which I'll explore in more depth later on in this article

Assistant Professor, Mail ID:mabhanuu@gmail.com

Assistant Professor,HOD , Mail ID:chandsarath70@gmail.com

Assistant Professor, Mail ID:krishna.cseit@gmail.com

Assistant Professor, Mail ID:bvenkat1109@gmail.com
Department of CSE Engineering,
Pallavi Engineering College Hyderabad, Telangana 501505.

PROBLEM DEFINITION

Fraudulent credit card transactions result in billions of dollars in losses each year. As ancient as mankind, deception may take on almost any shape. According to the 2017 PwC global economic crime study, around 48% of firms have been the victim of economic crime. As a result, there is a pressing need to find a solution to the credit card fraud detection issue. As new technologies emerge, fraudsters now have more options than ever before. Using a credit card is common in today's culture, and the number of people falling victim to credit card fraud is on the rise. Hugh In addition to merchants and banks, those who use credit cards are affected by financial losses that have been fraudulently perpetrated. For a business, non-financial costs from fraud might be difficult to measure in the near term, but they may become apparent over time. It's very uncommon for customers to switch credit card providers after becoming a victim of fraud with a particular organisation.

SCOPE OF THE PROJECT

Keeping track of fraudulent transactions is becoming more challenging as new technology is introduced. It is now possible to automate and save part of the effective work that goes into identifying credit card fraud thanks to the growth of machine learning, artificial intelligence, and other key information technology sectors.

RELATED WORK

[1] "KosemaniTemitayo Hafiz, Dr. Shaun Aghili, Dr. PavolZavarsky" use predictive analytics technology to detect credit card fraud in Canada.

Predictive analytics vendor solutions presently utilised to identify credit card fraud are the subject of this paper's assessment criteria, features, and capabilities. The scorecard compares five credit card predictive analytics vendor solutions that have been implemented in Canada against each other. A list of credit card fraud PAT vendor solution problems, hazards, and limits was compiled as a result of the study results.

Card fraud may be detected with the use of hybrid methods like [2] BLAST-SSAHA. The authors of the paper are "Amlan Kundu, SuvasiniPanigrahi, Shamik Sural, Senior Member of the IEEE, and Arun K. Majumdar" Paper proposes a two-stage sequence alignment in which the profile Analyzer (PA) first compares incoming transaction sequences to the real cardholder's prior spending sequences, and then compares these two sets of transactions. The profile analyzer then sends the strange transactions it found to a deviation analyzer (DA) to see whether they match up with previous fraudulent behaviour. It is up to these two analysts to make the ultimate determination on the nature of a transaction. Combining two sequence alignment methods such as BLAST and SSAHA in order to attain online response times of both PA and DA is suggested.Researchers Investigate a Distance Sum Detection Model for Credit Card Fraud"Wen-Fang YU, Na Wang," says the voiceover. Due to the rapid expansion of China's credit card market and increased amount of international commerce, there has been an uptick in credit card fraud. The emphasis of bank risk management is now on how to better identify and prevent credit card fraud. Outlier detection based on distance sum is proposed as a fraud detection model for credit card transactions that takes into account both the rarity and unusual nature of credit card fraud. Using this approach to identify credit card fraud is practical and accurate, according to the results of several experiments.Detection of Fraudulent Credit Card Transactions Using SVM and Decision Trees. Nipane, Poonam, Kalinge, Vidhate, Vidhate, Kunal, and Deshpande are the members of this group.Fraud is rising over the globe as e-commerce advances, resulting in massive losses for businesses. Currently, credit card fraud is a major source of financial loss for both businesses and consumers. They include decision trees, genetic algorithms, meta-learning strategies and neural networks, as well as HMMs. Support Vector Machine (SVM) and decision trees are being employed to tackle the challenge in a contemplated system for fraud detection. Financial losses may be minimised to a larger extent by the use of a hybrid method..Credit Card Fraud Detection using Supervised Machine Learning (SVM). SITARAM PATEL AND SUNITA GOND.There are numerous kernels in the SVM (Support Vector Machine) based approach proposed in this thesis, instead of only the spending profile. To summarise, there has been significant progress made in the simulation in terms of the rates of both "true positive" and "true negative," as well as reductions in the rates of "false positives" and "false negatives."Support Vector Machines and Decision Trees for Credit Card Fraud Detection, by Y. Sahin and E. DumanIn this work, decision trees and support vector machines (SVM) classification models are constructed and used to the challenge of detecting credit card fraud. SVM and decision tree algorithms were compared in this work for the first time using an actual data set for credit card fraud detection.

SYSTEM ANALYSIS

EXISTING SYSTEM

Cluster Analysis and Artificial Neural Networks have been used to identify credit card fraud in an existing system, where data normalisation is conducted before to Cluster Analysis, and findings demonstrate that by grouping characteristics, neuronal inputs may be decreased. Using normalised data with MLP-trained data may provide good results. Unsupervised learning was used in this study. Finding novel approaches for fraud detection and increasing the accuracy of outcomes were the main goals of this article. Real-world transactions from a big European corporation are used in this study, and personal information is kept private. An algorithm's accuracy hovers around 50%. The purpose of this work was to identify an algorithm and lower the cost metric in some way. By 23%, they were able to come up with Bayes' lowest risk method.

Disadvantages

One of the novel collative comparison measures provided in this study appropriately depicts the profits and losses owing to fraud detection. The suggested cost measure is used to demonstrate a Bayes minimal risk technique that is cost sensitive.

PROPOSED SCHEME

To classify credit card data, we are using a random forest algorithm in our suggested system. Classification and regression are both possible with Random Forest. Basically, it's a set of decision tree classifiers. Over-fitting is a problem for decision trees, although random forests solve this problem. Randomly sampling a portion of the training set is used to train each tree in the decision tree, which then divides on a feature taken from the whole feature set. For big data sets with numerous characteristics and instances, training in a random forest is incredibly rapid since each tree is trained independently of the others. For example, the Random Forest approach provides a decent estimate of generalisation error and is tolerant to over fitting.

ADVANTAGES OFPROPOSED SYSTEM

In a regression or classification task, Random Forest can naturally rank the relevance of variables. • The 'amount' feature refers to the total transaction value. Features 'class' are used in the binary classification and take values of 1 (fraud) or 0 (no fraud), depending on the outcome of the classification (not fraud).

REQUIREMENT SPECIFICATIONS

Software products must meet the criteria specified in the requirements specification. It is the first phase in the requirements analysis process, and it outlines the functional, performance, and security needs of a specific software system, among other things. Requirements specifications are used to offer an in-depth look into the characteristics and objectives of a software project.

4.1HARDWARE REQUIREMENTS

It has an Intel processor, 4GB of RAM, and a 260GB hard drive.Keyboard - Windows standard keyboardA two- or three-button mouse is required for this task.

SOFTWARE REQUIREMENTS

Python Anaconda OS - Windows 7, 8 and 10 (32 and 64 bit)

5. FEASIBILITY STUDY

5.1TECHNICAL FEASIBILITY

Using Anaconda, it is clear that the hardware and software required for the planned system are available.

5.2 ECONOMICAL FEASIBILITY

The suggested system's cost is lower than that of competing software. 5.3 SUITABILITY FOR USE IN ACTIVITIES In order to work on the programme, it is required to set up the appropriate software.

6. SYSTEM ARCHITECTURE

Cleansing and validation are conducted on credit card data before it is separated into two parts: one for training and the other for testing. The training dataset is used to train the model, while testing datasets are used to evaluate the model's performance. Teat and train datasets have now been generated from the original sample at random.

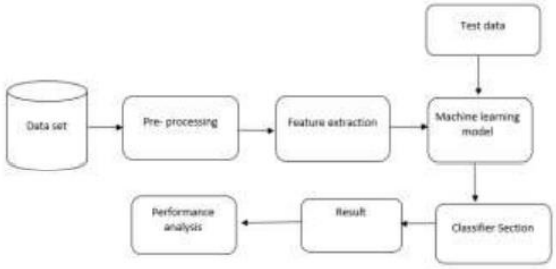

Figure 6.1 shows the proposed system's architecture.

7. MODULES IN THE SYSTEM

7.1 SUMMARY OF THE MODULE

1. DATA COLLECTION

Credit card transaction records were utilised to gather product evaluation data for this study. This stage is all about narrowing down the data set that you'll be using. In order to solve ML issues, you need a lot of data (examples or observations) that you already know how to analyse.

7.1.2 MODULE 2: DATA PRE-PROCESSING

You may do this by formatting, cleaning, and sampling from the data you've picked. The most typical data pre-processing stages are as follows: There is a possibility that the data you have chosen is not in a format that allows you to work with it. A flat file or a relational database may be all that is required, but if the data is in a proprietary file format, you may want the data to be in a relational database. Data cleaning is the process of removing or replacing incorrect or incomplete information. There may be data instances that lack the information you assume you need to solve the situation. There may be a need to eliminate these occurrences. Additionally, some of the characteristics may include confidential or sensitive information that must be omitted from the data as a whole. It's possible that you'll have access to more data than you need for your project. Increasing the amount of data in an algorithm might cause it to take longer to perform and need more computing and memory resources. It is possible to choose a smaller representative sample of the data that may be quicker for exploring and developing ideas before evaluating the whole dataset, which may save time and effort.

7.1.3 MODULE 3: FEATURE EXTRATION

Go ahead and do it now. Feature extraction is the process of reducing the attributes of a feature. Feature extraction is different from feature selection in that it actively alters the existing qualities rather than ranking them. The original qualities are linearly combined to get the changed attributes or features. Finally, we use the Classifier method to train our models. Python's Natural Language Toolkit library's categorise module is what we're using. We make use of the acquired tagged data. All of our labelled data will be utilised to test the models. Pre-processed data was classified using machine learning methods. Random forest was used as a classifier. Text categorization is a frequent use case for these algorithms.

7.1.4 MODULE 4: Evaluation Model Model

Model development wouldn't be complete without evaluation. Using this method, we can identify the best model to reflect our data, as well as how well the model will perform in future experiments. Using the same data to evaluate model performance as the training set is not acceptable in data science since it may easily lead to models that are both optimistic and too tightly fitting. Hold-Out and Cross-Validation are two strategies used in data science to evaluate models. Both techniques employ a test set that is not viewed by the model to ensure that the model is not overfitting. It is expected that each categorization model's performance is based on the average of all of its individual results. The final product will be shown in this manner. Graphs are used to represent data that has been categorised. Percentage of accurate predictions for the test data is referred to as accuracy. Divide the number of right guesses by the total number of predictions to arrive at the answer.

8. Algorithm Utilized

8.1 Random Forest

Machine learning algorithm Random Forest relies on the concept of ensemble learning. Combining several algorithms or using the same technique numerous times to build a more effective prediction model is known as ensemble learning. Multiple decision trees are combined in the Random Forest method, resulting in a forest of trees, which is why the term "Random Forest" was coined. For both regression and classification, the random forest technique is a good choice.

8.2 WORKING OF RANDOM FOREST

In order to run the random forest algorithm, you must follow these fundamental steps: To begin, choose N records at random from the dataset. Create a decision tree based on the N data that you've collected so far. Repeat steps 1 and 2 for the number of trees you wish to use in your algorithm. Each tree in the forest has the ability to forecast the category to which a new record will belong in the classification issue. Finally, the new record is awarded to the category that receives the most votes, and therefore the new record.

8.3 ADVANTAGES OF USING RANDOM FOREST

The advantages of classifying and predicting using random forests. Since there are several trees and each tree is trained on a portion of data, the random forest approach is not biassed. The random forest method depends on the strength of "the crowd," which reduces the system's overall slant. There is a high degree of predictability in this method. However, the overall method is not much impacted by new data since even if one tree is impacted by the new data, it is very unlikely to influence all trees. With a combination of category and numerical variables, the random forest method performs well. Data with missing values or that has not been scaled adequately might also benefit from the random forest approach because of its robustness.

10. CONCLUSION

More training data improves Random Forest's performance, but testing and application speed decrease. More pre-processing procedures might also be beneficial. To get better results from SVM, additional preprocessing is needed on the data, which is still an issue for the SVM method, despite the fact that the results are excellent.

REFERENCES

[1] Sudhamathy G: Credit Risk Analysis and Prediction Modelling of Bank Loans Using R, vol. 8, no-5, pp. 1954-1966.

[2] LI Changjian, HU Peng: Credit Risk Assessment for ural Credit Cooperatives based on Improved Neural Network, International Conference on Smart Grid and Electrical Automation vol. 60, no. - 3, pp 227-230, 2017.

[3] Wei Sun, Chen-Guang Yang, Jian-Xun Qi: Credit Risk Assessment in Commercial Banks Based On Support Vector Machines, vol.6, pp 2430-2433, 2006.

[4] Amlan Kundu, SuvasiniPanigrahi, Shamik Sural, Senior Member, IEEE,"BLAST-SSAHA Hybridization for Credit Card Fraud Detection", vol. 6, no. 4 pp. 309-315, 2009.

[5] Y. Sahin and E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines, Proceedings of International Multi Conference of Engineers and Computer Scientists, vol. I, 2011

[6] Sitaram patel, Sunita Gond , "Supervised Machine (SVM) Learning for Credit Card Fraud Detection, International of engineering trends and technology, vol. 8, no. -3, pp. 137- 140, 2014.

[7] Snehal Patil, HarshadaSomavanshi, Jyoti Gaikwad, Amruta Deshmane, Rinku Badgujar," Credit Card Fraud Detection Using Decision Tree Induction Algorithm, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.4, April- 2015, pg. 92-95 [

8] Dahee Choi and Kyungho Lee, "Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System", vol. 5, no. - 4, December 2017, pp. 12-24