International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST52310236

Credit Card Fraud Detection Using Random Forest and Cart Algorithm

Dr. Ch. Anusha [1], B. Maruthi Lalitha [2], B. Sravya [3] ,V. Sri Bindu[4] ,T. Omkanth [5]

[1]Associate Professor, Department of Information Technology, Kallam Haranadha Reddy Institute of Technology, Chowdavaram, Guntur (Dt), Andhra Pradesh, India

[2, 3, 4, 5]B. Tech Students, Department of Information Technology, Kallam Haranadha Reddy Institute of Technology, Chowdavaram, Guntur(Dt), Andhra Pradesh, India

ARTICLEINFO

Article History:

Accepted: 01 March 2023
Published: 15 March 2023

Publication Issue

Volume 10, Issue 2
March-April-2023

Page Number

226-231

ABSTRACT

The project is primarily concerned with detecting credit card fraud in the real world. The phenomenal increase in the number of credit card transactions has recently resulted in a significant increase in fraudulent activities. The goal is to obtain goods without paying for them or to withdraw unauthorized funds from an account. In order to minimize losses, all credit card issuing banks must implement effective fraud detection systems. One of the most difficult challenges in starting the business is that neither the card nor the cardholder must be present at the time of purchase. This makes it impossible for the merchant to determine whether or not the customer making a purchase is the legitimate cardholder. Using the proposed scheme and the random forest algorithm, the accuracy of detecting fraud can be improved. Random forest algorithm classification process to analyze data set and user current dataset. Finally, improve the accuracy of the output data. The techniques' performance is measured using accuracy, sensitivity, specificity, and precision. Then, by processing some of the provided attributes, the fraud detection is identified and the graphical model visualization is provided. The techniques' performance is measured using accuracy, sensitivity, specificity, and precision.

Keywords: Credit Card, Fraud Detection, Random Forest.

I. INTRODUCTION

Various fraudulent activity detection techniques have been implemented in credit card transactions, and researchers have kept methods to develop models based on artificial intelligence, data mining, fuzzy logic, and machine learning in mind. Credit card fraud detection is a difficult, but common, problem to solve. We used Machine learning to detect credit card fraud in our proposed system. Machine learning techniques are improving. Machine learning has been identified as an effective measure for detecting fraud.

During online transaction processes, a large amount of data is transferred, yielding a binary result: genuine or fraudulent. Features are built within the sample fraudulent datasets. These are data points such as the customer account's age and value, as well as the origin of the credit card. There are hundreds of features, each of which contributes to the likelihood of fraud to varying degrees. It should be noted that the level to which each feature contributes to the fraud score is generated by the machine's artificial intelligence, which is driven by the training set, but is not determined by a fraud detector analyst. So, in terms of card fraud, if the use of cards to commit fraud is proven to be high, the fraud weighting of a credit card transaction will be equally so. However, if this were to decrease, the contribution level would decrease as well. Simply put, these models should self-learn without any explicit programming, such as manual review. Machine learning is used to detect credit card fraud by deploying classification and regression algorithms. To classify fraudulent card transactions, we use supervised learning algorithms such as the Random forest algorithm, which can be used online or offline. Random forest is a more sophisticated version of Decision tree. Random forest outperforms the other machine learning algorithms in terms of efficiency and accuracy. By selecting only a subsample of the feature space at each split, random forest aims to reduce the previously mentioned correlation issue. Essentially, it aims to de-correlate the trees and prune them by defining a stopping criteria for node splits, which I will go over in more detail later.

1.1 PROBLEM DEFINITION

Every year, fraudulent credit card transactions cause billions of dollars in losses. Fraud is as old as humanity and can take on an infinite number of different forms. According to the 2017 PwC global economic crime survey, approximately 48% of organizations experienced economic crime. As a result, there is a strong desire to solve the problem of credit card fraud detection. Furthermore, the advancement of new technologies opens up new avenues for criminals to commit fraud. Credit card use is widespread in modern society, and credit card fraud has increased in recent years. Hugh Financial losses caused by fraud affect not only merchants and banks, but also individuals who use credit. Fraud may also harm a merchant's reputation and image, resulting in non-monetary losses that, while difficult to quantify in the short term, may become visible over time. For example, if a cardholder is a victim of fraud with one company, he may no longer trust them and opt for a competitor.

1.1 SCOPE OF THE PROJECT

In this proposed project, we created a protocol or model to detect fraudulent credit card transactions. This system is capable of providing the majority of the necessary features for detecting fraudulent and legitimate transactions. As technology evolves, it becomes more difficult to track fraudulent transaction behaviour and patterns. With the advancement of machine learning, artificial intelligence, and other fields of information technology, it is now possible to automate the process and save some of the effective amount of labour that is put into detecting credit card fraudulent activity.

II. RELATED WORK

[1] The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada. “Kosemani Temitayo Hafiz, Dr. Shaun Aghili, Dr. Pavol Zavarsky.”

This study focuses on the development of a scorecard based on relevant evaluation criteria, features, and capabilities of predictive analytics vendor solutions currently used to detect credit card fraud. The scorecard compares five credit card predictive analytics vendor solutions used in Canada side by side. A list of credit card fraud PAT vendor solution challenges, risks, and limitations was derived from the research findings.

[2] BLAST-SSAHA Hybridization for Credit Card Fraud Detection. “Amlan Kundu, Suvasini Panigrahi, Shamik Sural, Senior Member, IEEE, and Arun K. Majumdar”

The authors propose a two-stage sequence alignment method in which a profile Analyzer (PA) first determines the similarity of an incoming sequence of transactions on a given credit card with the genuine cardholder's past spending sequences. The profile analyzer then forwards the unusual transactions to a deviation analyzer (DA) for possible alignment with previous fraudulent behaviour. The observations of these two analysts are used to make the final decision about the nature of a transaction. We propose a new approach for combining two sequence alignment algorithms, BLAST and SSAHA, to achieve online response time for both PA and DA.

[3] Research on Credit Card Fraud Detection Model Based on Distance Sum.“Wen-Fang YU, Na Wang”.

Credit card fraud is on the rise in China, alongside the growth of credit cards and trade volume. How to improve the detection and prevention of credit card fraud has become the focus of bank risk management. It proposes a credit card fraud detection model based on distance sum based on the infrequency and irregularity of fraud in credit card transaction data, incorporating outlier mining into credit card fraud detection. Experiments show that this model can detect credit card fraud and is accurate.

[4] Fraudulent Detection in Credit Card System Using SVM & Decision Tree. “Vijayshree B. Nipane, Poonam S. Kalinge, Dipali Vidhate, Kunal War, Bhagyashree P. Deshpande”.

With the advancement of electronic commerce, fraud is spreading throughout the world, causing significant financial losses. Credit card fraud is a major cause of financial losses in the current scenario; it affects both traders and individual clients. The methods presented to detect credit card fraud include decision trees, genetic algorithms, meta learning strategies, neural networks, and HMM. The artificial intelligence concepts of Support Vector Machine (SVM) and decision tree are being used to solve the problem in the contemplate system for fraudulent detection. As a result of implementing this hybrid approach, financial losses can be reduced to a greater extent.

5] Supervised Machine (SVM) Learning for Credit Card Fraud Detection. “Sitaram patel, Sunita Gond”.

This thesis proposes an SVM (Support Vector Machine)-based method with multiple kernel involvement that includes several fields of user profile rather than just spending profile. The simulation results show an increase in the TP (true positive) and TN (true negative) rates, as well as a decrease in the FP (false positive) and FN (false negative) rates.

[6] Detecting Credit Card Fraud by Decision Trees and Support Vector Machines. “Y. Sahin and E. Duman”.

In this study, classification models based on decision trees and support vector machines (SVM) are developed and applied to the problem of detecting credit card fraud. This is one of the first studies to compare the performance of SVM and decision tree methods in detecting credit card fraud using real data.

III. SYSTEM ANALYSIS

In the existing system, a case study involving credit card fraud detection was investigated, and the results obtained from the use of Cluster Analysis and Artificial Neural Networks on fraud detection demonstrated that by clustering attributes neuronal inputs could be minimized. Furthermore, promising results can be obtained by using normalized data that has been MLP trained. Unsupervised learning was used in this study. The significance of this paper was to discover new methods for detecting fraud and to improve the accuracy of results. The data set for this paper is based on real-world transactional data from a large European company, and personal information in the data is kept private. An algorithm's accuracy is around 50%.

Disadvantages:

1. This paper proposes a new collative comparison measure that accurately represents the gains and losses due to fraud detection.
2. Using the proposed cost measure, a cost-sensitive method based on Bayes minimum risk is presented.

3.1 PROPOSED SCHEME

In the proposed system, we use the random forest algorithm to classify the credit card dataset. Random Forest is a classification and regression algorithm. In a nutshell, it is a set of decision tree classifiers. Random forest outperforms decision trees because it corrects the habit of overfitting to their training set. A random subset of the training set is sampled to train each individual tree, and then a decision tree is built, with each node splitting on a feature chosen at random from the full feature set. Even for large data sets with many features and data instances, random forest training is extremely fast because each tree is trained independently of the others. Random Forest is an algorithm.

3.2 ADVANTAGES OF PROPOSED SYSTEM

- Random Forest can rank the importance of variables in a regression or classification problem in a natural way.
- The transaction amount is represented by the 'amount' feature. The binary classification's target class is represented by the feature 'class,' which has a value of 1 for positive cases (fraud) and 0 for negative cases (not fraud).

IV. SYSTEM ARCHITECTURE

First the credit card dataset is taken from the source and cleaning and validation is performed on the dataset which includes removal of redundancy, filling empty

spaces in columns, converting necessary variable into factors or classes then data is divided into 2 part, one is training dataset and another one is test data set. Now the original sample is randomly partitioned into teat and train dataset.

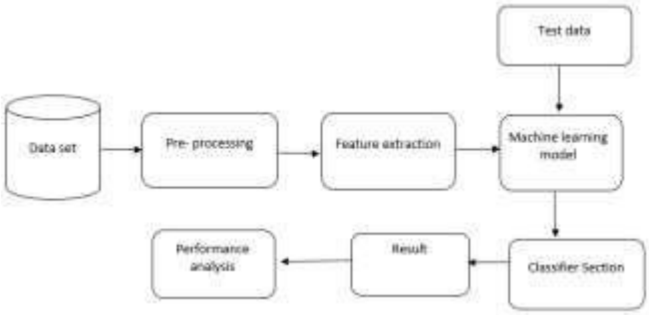

Figure 4.1- ARCHITECTURE OF THE PROPOSED SYSTEM

V. SYSTEM MODULES

MODULE 1: DATA COLLECTION

The data in this paper is a collection of product reviews gleaned from credit card transaction records. This step is concerned with selecting a subset of all available data with which to work. ML problems begin with data, preferably a large amount of data (examples or observations) for which you already have an answer. Labeled data is data for which you already know the target answer.

MODULE 2: DATA PRE-PROCESSING

Organize your selected data by formatting, cleaning and sampling from it.

Three common data pre-processing steps are:

Formatting: The data you have chosen may not be in a format that you can work with. You may have data in a relational database and want it in a flat file, or you may have data in a proprietary file format and want it in a relational database or a text file.

Cleaning: Cleaning data is the process of removing or replacing missing data. There may be data instances that are incomplete and do not contain the information you believe you require to solve the problem. These instances may require removal. Furthermore, some of the attributes may contain sensitive information, and these attributes may need to be removed entirely from the data.

Sampling: There may be far more selected data available than you require. More data can result in much longer algorithm running times and higher computational and memory requirements. Before considering the entire dataset, you can take a smaller representative sample of the selected data, which may be much faster for exploring and prototyping solutions.

MODULE 3: FEATURE EXTRATION

The following step is to Feature extraction is a method of reducing the number of attributes. In contrast to feature selection, which ranks existing attributes based on their predictive significance, feature extraction transforms the attributes. The transformed attributes are linear combinations of the original attributes. Finally, we train our models with the Classifier algorithm. On Python, we use the classify module from the Natural Language Toolkit library. We make use of the labelled dataset that was gathered. The remaining labelled data will be used to assess the models. To classify pre-processed data, some machine learning algorithms were used. Random forest classifiers were chosen. These algorithms are widely used for text classification tasks.

MODULE 4: Evaluation Model
Model evaluation is an essential step in the modelling process. It aids in determining the best model to represent our data and how well the chosen model will perform in the future. In data science, evaluating model performance with the data used for training is not acceptable because it can easily produce overoptimistic and overfit models. In data science, there are two methods for evaluating models: hold-out and cross-validation. To avoid overfitting, both methods evaluate model performance using a test set (unseen by the model). The averaged performance of each classification model is used to estimate its performance. The end result will be visualized. Graphs are used to represent classified data. The percentage of precision is defined as accuracy..

VI. Algorithm Utilized Random Forest

Random forest is an ensemble learning-based supervised machine learning algorithm. Ensemble learning is a type of learning in which multiple instances of the same or different algorithms are combined to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm is applicable to both regression and classification problems.

ADVANTAGES OF USING RANDOM FOREST
Pros of using random forest for classification and regression.

1. The random forest algorithm is not biased, since, there are multiple trees and each tree is trained on a subset of data. Basically, the random forest algorithm relies on the power of "the crowd"; therefore, the overall biasedness of the algorithm is reduced.
2. This algorithm is very stable. Even if a new data point is introduced in the dataset the overall algorithm is not affected much since new data may impact one tree, but it is very hard for it to impact all the trees.

3. The random forest algorithm works well when you have both categorical and numerical features.

The random forest algorithm also works well when data has missing values or it has not been scaled well.

VII. APPENDICES

7.1 SAMPLE SCREENSHOTS FROM THE PROJECT

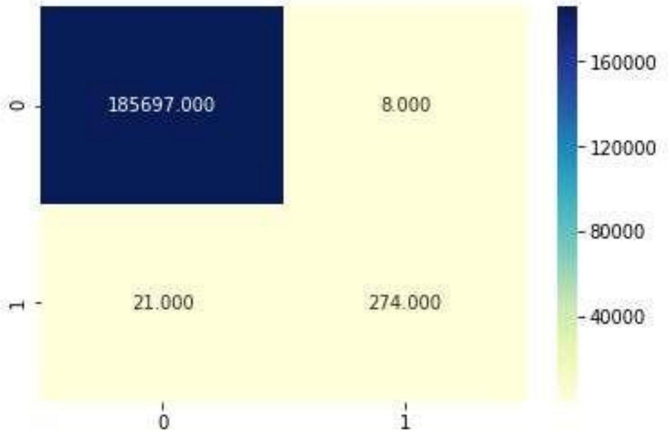

Fig- 7.1: Exact figures of fake and original credit card

VIII. CONCLUSION

The Random forest algorithm performs better with more training data, but its speed during testing and application suffers. More pre-processing techniques would also be beneficial. The SVM algorithm still suffers from the imbalanced dataset problem and requires more preprocessing to produce better results. The results produced by SVM are excellent, but they could have been better if the data had been preprocessed more thoroughly.

IX. 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, Suvasini Panigrahi, 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.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.

[6]. Snehal Patil, Harshada Somavanshi, 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

[7]. 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.

Cite this article as :

Dr. Ch. Anusha, B. Maruthi Lalitha, B. Sravya, V. Sri Bindu, T. Omkanth, "Credit Card Fraud Detection Using Random Forest and Cart Algorithm", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 2, pp. 226-231, March-April 2023. Available at doi : https://doi.org/10.32628/IJSRST52310236
Journal URL : https://ijsrst.com/IJSRST52310236