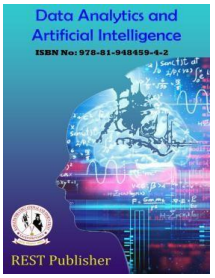
Data Analytics and Artificial Intelligence

Vol: 3(3), 2023

REST Publisher; ISBN: 978-81-948459-4-2

Website: http://restpublisher.com/book-series/daai/



A Novel Approach in Credit Online Fraud Detection System Using Machine LearningTechniques

S. R Sathya Priya, Dhanushya M, Gayathri S, *Jaya keerthana R

Adhiyamaan College ofEngineering (Autonomous), Hosur, Tamil Nadu, India.

*Corresponding Author Email: jayakeerthana279@gmail.com

Abstract: *In the world of finance, as the technology grows, new systems of business making came into picture. Credit card system is one among them. But because of lot of loop holes in this system, lot of problems are aroused in this system in the method of credit card scams. Due to this the industry and customers who are using credit cards are facing a huge loss. There is a deficiency of investigation lessons on examining practical credit card figures in arrears to privacy issues. In the manuscript an attempt has been made for finding the frauds in the credit card business by using the algorithms which adopted machine learning techniques. we get the data set from the European company were the true transaction is numbered as 0 and false transaction is numbered as 1.*

Keywords: *Semi supervised classification, Logistic regression.*

1. INTRODUCTION

Recently, online purchases using credit cards have increased drastically, people are not generally aware of a probable fraudulent transaction that could happen to them. Credit card security is determined by the card's physical characteristics and the privacy of the card number. As a result of globalization and the growth of Internet-based commerce, worldwide credit card purchases have increased. In addition to the rapid increase in credit card purchases, another important factor contributing to theincrease in fraud is credit card fraud. Theft and fraud are committed in a given transaction using a payment card as a fraudulentsource of funds. A vast range of methods to conduct theft are used by Credit Card Fraudsters. To successfully combat credit card fraud, it is important to have a basic understanding of the process of detecting credit card fraud. Due to numerous credit card fraud monitoring and avoidance mechanisms, credit card fraud has stabilized a lot over the years. However, cardholders use fake transactions to scam bank cash. External card fraud, on the other hand, is primarily expressed in the use of stolen fraudulent, stolen credit cards to consume or obtain cash in concealed ways, such as buying valuable, limited amounts of products or items that are easy to sell in cash. This project will specifically explore & analyze the development of a machine learning-based fraud detection system.

2. RELATED WORK

Credit card fraud is a serious and growing problem. While predictive models for credit card fraud detection are in active use in practice, reported studies on the use of data mining approaches for credit card fraud detection are relatively few, possibly due to the lack of available data for research. This paper evaluates two advanced data mining approaches, support vector machines and random forests, together with the well-known logistic regression, as part of an attempt to better detect (and thus control and prosecute) credit card fraud. The study is based on real-life data of transactions from an international credit card operation. The investigation of fraud in the financial domain has been restricted to those who have access to relevant data. However, customer financial records are protected by law and internal policies, therefore they are not available for most of the researchers in the area of fraud detection. This paper aims to present the work of those researchers who have had access to data and present an interesting approach to fraud detection research; which is the generation of a synthetic data set to work on fraud detection research. Some of the domains covered in this review include mobile money payments, e-payments, retail stores, online bank services and credit card payments. We also cover some of the most relevant surveys in the field and point out the impossibility to compare this work due to the lack of common public data set to test

3. EXISTING SYSTEM

He utilization of information mining approaches for detecting credit card fraud is comparatively low, most likely due to a lack of readily available information. The current systems used across the business sector are used to detect fraud in different ways. Each system detects fraud differently. By using authority and mobility to simulate synthetic data, they can remove customer privacy and security restrictions associated with real data when financial fraud is detected.

4. PROPOSED SYSTEM

The project defines the procedure used to hostage the credit card scam. The numerous competent approaches like arrangement orientation, device learning, neural networks, artificial intelligence, fuzzy logic are employed to detect and encounter scams in credit card businesses. Credit card fraud has become progressively widespread in modern years. In Current day, the fraud is one of the key causes of excessive business losses, not only for merchants, distinct clients are also affected. Sothere are some methods to detect such kind of frauds. Initially, clustering model was adopted to categorize the authorized and deceitful operation by means of data cauterization of areas of factor value. Furthermore, Gaussian mixture model is used to model the possibility thickness of credit card. In credit card transactions, various fraudulent activity detections have been implemented so far. We use different techniques. In September 2013, European cardholders made over 2,84,000 credit card transactions using the dataset we are using. In this dataset, frauds make up 0.172% of all transactions, an extremely unbalancednumber.

5. MATERIAL AND METHODS

Classification model: XGB BOOST Gradient boosted trees can also be implemented with XG Boost. It's an open-source program that's popular and efficient. Gradient boosting Combine estimates from a set of simpler and weaker models in an attempt to accurately predict target variables. The gradient boosted trees algorithm is implemented by the open-source softwareXG Boost. The supervised learning algorithm combines the Controlled learning algorithms include estimating arrays of weaker and simpler models to predict target variables with higher accuracy. A regression tree is usually used as a weak learner in gradient boosting, and a regression tree maps its input data to the leaf containing the continuous score. By combining a convex loss function (predicting the target output in terms of the predicted output) with a penalty term for model complexity, XG Boost minimizes a regularized (L1 and L2) objective function. In each iteration of training, additional trees are added that predict the remnants of the previous tree. These new trees are then combined with the previous tree to make the final prediction. Gradient boosting reduces losses when adding a new model by using a gradient descent algorithm. **Logistic *regression:*** Additionally, we used a classification algorithm, Logistic Regression, which is an algorithm for predicting binary values (1 / 0, Yes / No, False / True) from a set of independent variables. Predicts the probability of an event occurring as a function of a dependent variable when the resulting variable is categorical. Your logistic regression then becomes a linear regression.

6. METHODOLOGY

***Data collection*:** Data used in this paper is a set of product reviews collected from credit card transactions records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called labelled data.

Data pre-processing: Data pre-processing consist of three steps formatting, cleaning and sampling. Formatting is a process to format the data. Cleaning is removing the unwanted data and sampling is process to sample the data.

Distribution: The dataset consists of 2,84,807 credit card transactions, out of which only 492 transactions are fraudulent. Our dataset the data

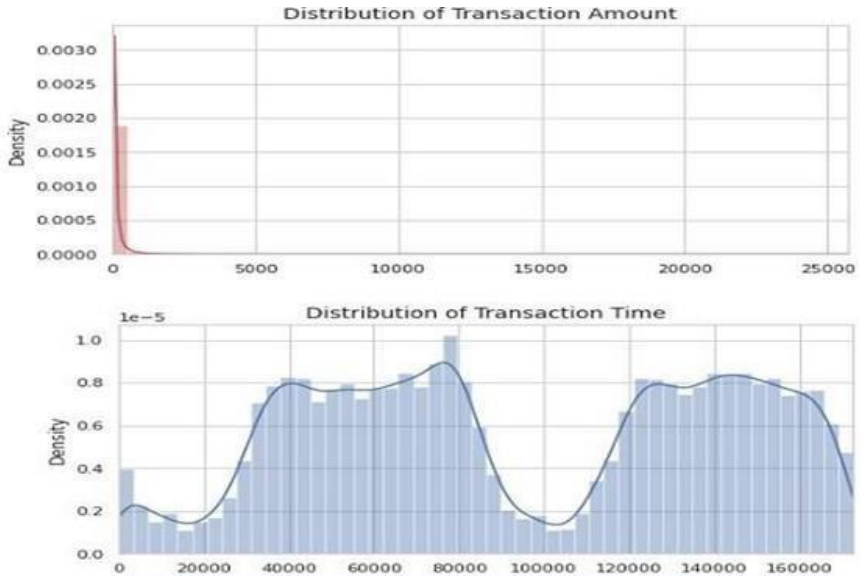


FIGURE1.

***Anomaly detection*:** In order to avoid over fitting we need to remove all the outliers in our dataset. We will use anomaly detection techniques for removing outliers. Once we have all the classes that are highly correlated with the dependent variable, we will remove the extreme outliers. Here are some of the features before removing outliers and after removing outliers. We will draw the positively correlated classes using Boxplot.

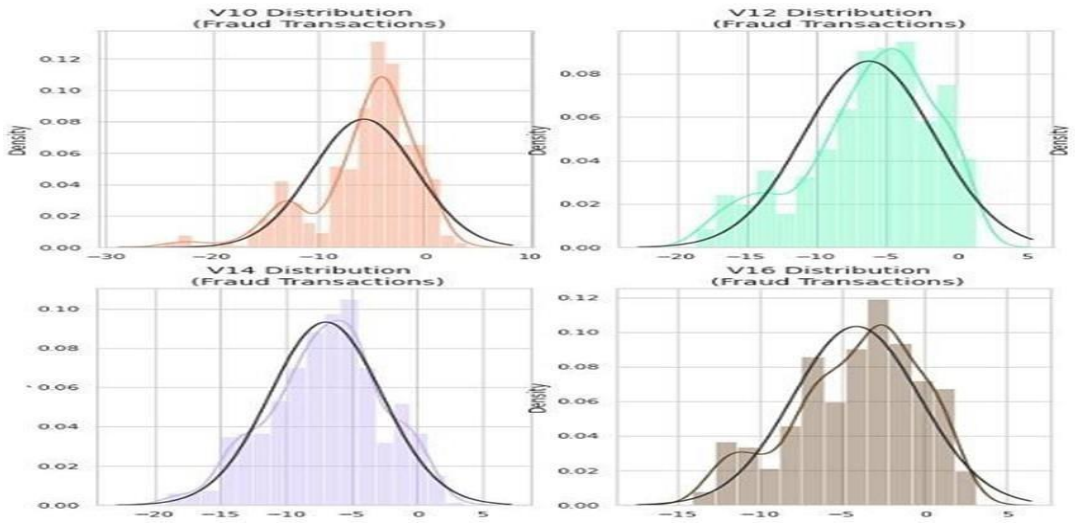


FIGURE 2.

7. RESULT AND DISCUSSION

Metrics for performance: A confusion matrix serves as a basic measure for performance. The confusion matrix consists of two by two matrix tables with four outcomes produced by the binary classifier. The confusion matrix provides a variety of measures including sensitivity, specificity, accuracy, and error rate. The accuracy of the prediction is calculated as the sum of two correct predictions (P+Q) divided by the total number of datasets (R+S).Essentially, it is (1-error rate).

A=P+Q/R+S

Where,

A=Accuracy

P=True Positives
Q=True
Negatives
R=False Positives
S=false Negatives
We will find the ROC score and cross-validation score forboth models in order to validate the model
Accuracy of train prediction=0.93456
Accuracy of train prediction=0.93452

8. CONCLUSION

If the train value and the test value are same then it is a true transaction. If the train and the test values are different then it is said to be fraudulent transcation. From this we can find whether the transaction is fraudulent or not. On the whole dataset, one can try implementing one method. Another drawback is that we cannot determine the names of fraud and non- fraud transactions for the given dataset using machine learning. The project can be further developed by finding a way to address this issue using various methods.

REFERENCES

[1] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, “Data mining for credit card fraud: A comparative study," Decision Support Syst., vol. 50, no. 3, pp. 602613, 2011.

[2] E. A. Lopez-Rojas and S. Axelsson, “A review of computer simulation for fraud detection research in financial datasets,” in 2016Future Technologies Conference (FTC), 2016, pp. 932–935.

[3] D. Al-Jumeily, A. Hussain, A. MacDermott, G. Seeckts, and J. Lunn,“Methods and techniques to support the development of frauddetection system,” in 2015 International Conference on Systems,Signals and Image Processing (IWSSIP), 2015, pp. 224–227.

[4] Ayushi Agrawal, Shiv Kumar, and Amit Kumar Mishra,“Implementation of Novel Approach for Credit Card Fraud Detection,” in 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 1–4.

[5] A. Shen, R. Tong, Y. Deng, "Application of classification models on credit card fraud detection", Service Systems and Service Management 2007 International Conference, pp. 1-4, 2007.

[6] Y. Sahin, E. Duman, "Detecting credit card fraud by ANN and logistic regression", Innovations in Intelligent Systems and Applications (INISTA) 2011 International Symposium, pp. 315-319,2011.

[7] F. T. Liu, K. M. Ting, and Z. H. Zhou, “Isolation forest,” in Proc. The Eighth IEEE International Conference on Data Mining, 2008, pp.413-422.

[8] C. S. Hemalatha, V. Vaidehi, and R. Lakshmi, “Minimal infrequent pattern-based approach for mining outliersin data streams,” Expert Systems with Applications, vol. 42, no. 4, pp. 1998-2012, March 2015.

