SJIF Impact Factor 2021: 8.013| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 7 | July 2021 - Peer Reviewed Journal

SUPERVISED MACHINE LEARNING ALGORITHMS FOR DETECTING CREDIT CARD FRAUD

G.Bhargav Chowdari
UG Student, Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences,Chennai

Article DOI: https://doi.org/10.36713/epra7636
DOI No: 10.36713/epra7636

ABSTRACT

One of the most serious ethical challenges in the credit card industry is fraud. Our paper's major goal is to identify credit card theft and offer a reasonable solution to the problem. Credit card fraud has cost customers and banks billions of dollars around the world. Fraudsters are constantly attempting to come up with new ways and tricks to commit fraud, despite the fact that there are several measures in place to prevent it. Fraud detection is extremely important in the banking and finance industries. For detection purposes, we will use an artificial neural network. As a result, in order to prevent it, we will develop a system that will not only detect fraud, but will also detect it before it occurs. In order to detect new scams, our system will learn from previous frauds. Mining algorithms were used to detect fraud, but they failed miserably. We use machine learning methods to detect fraud in credit card transactions in our paper. The research employs supervised learning methods that are applied to a kaggle dataset that is severely skewed and imbalanced. We used robust scalar to balance the set, resulting in 51 percent non-fraud cases and 49 percent fraud ones. Logistic regression, random forest, decision tree, and KNN have all been implemented, with additional learning curves displaying which algorithm performs best.

Accuracy, specificity, precision, and sensitivity are the evaluation criteria, and a comparative chart is created to show the comparative analysis of various supervised learning algorithms.

KEYWORDS: *KNN,Neural network,Logistic regression,Random forest,Decision tree*

INTRODUCTION

In today's environment, using a credit card is a routine occurrence. It is frequently used for online payments and transactions.Credit cards can be used in a variety of ways. The use of credit cards has expanded tenfold, increasing the chances of fraud in such purchases. Credit card fraud costs the global economy billions of dollars. Fraud is defined as deception for the purpose of making illegal gains with someone else's money. Credit card fraud can be done in a number of different ways. By using lost or stolen cards, making fake or counterfeit cards, duplicating the original website, removing or altering the magnetic strip on the card that carries the user's information, by phishing by skimming or stealing data from a merchant's end. One of the techniques of purchasing products or services is by using a credit card. Fraud detection is the process of distinguishing between fraudulent and non-fraudulent transactions so that customers can enjoy their shopping or other transactions without delay. Many detections, such as the evolutionary algorithm, item set mining, and migrating birds' algorithm, have been used to solve this problem. The dataset for credit card fraud detection is extremely rare, and even if it exists, it is highly skewed and imbalanced, making it difficult to deploy algorithms appropriately. As a result, just a few changes to the dataset are required before the algorithms can be executed.

SJIF Impact Factor 2021: 8.013| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 7 | July 2021 - Peer Reviewed Journal

ALGORITHMS USED

1.) Decision Tree:

This is an algorithm that uses a tree-like graph and all possible outcomes to predict the final decision. It employs conditional control statements

2.) Logistic regression:

Linear regression and logistic regression are quite similar, but there is one difference: in logistic regression, a curve is obtained, whereas in linear regression, a straight line is obtained.

3.) Random forest :

This is an algorithm for classifying and regressing data. Random Forest is mostly comprised of decision tree classifiers. Random Forest is chosen over decision tree because it eliminates the problem of overfitting in the training set. To train each tree, we can randomly choose a subset of the training set and then construct a decision tree.

4.) K-nearest neighbor classifier:

KNN is used for classification and regression, and it does classification in the same way as Euclidean, Manhattan, and Minkowski distance functions do. Continuous variables are preferred by the Euclidean and Manhattan models, although the Minkowski model works well with categorical variables.

DETECTION OF CREDIT CARD FRAUD

The convenience of using a credit card to order anything from the comfort of your own home has also brought scammers closer to this technology. Credit cards are an easy target because they allow you to earn a large sum of money in a short period of time. The most vulnerable to fraud are transaction products, such as credit cards. Other items, such as personal loans and retail, are, on the other hand, at significant risk.

Techniques of Credit card fraud detection:

1) Credit Card Imprints, either electronic or manual: When a fraudster skims information from the card's magnetic strip.
2) Skimming is the most common method used to do counterfeit card fraud. A false magnetic swipe card is created, which has all of the information from the genuine card.
3) Card ID Theft: This is a type of application fraud.
4) Account Takeover: One of the most popular types of fraud is account takeover. The account information may be accessed by the fraudster.
5) False Merchant Sites: This is similar to a phishing attempt in which the customer is duped into visiting a fraudulent website that looks extremely similar to a legitimate one.
6) The User is charged an additional fee by the vendor.
7) Bankruptcy deception. This section discusses bankruptcy fraud and suggests that credit bureau reports be used as a source of information about the applicants' public histories, as well as the usage of a bankruptcy model.

Credit Card Detection Issues

There is a scarcity of study into real-world fraud detection issues. Credit card fraud has existed in this age of modern technologies due to a low rate of experimental analysis. The fundamental issue is that finance departments do not offer sensitive information to researchers in order for them to come up with a solution. Because only a small percentage of transactions are fraudulent, an effective classifier must be able to handle complex data. Because many transactions are identical, the classifier must be able to distinguish between correct and fraudulent transactions. To detect new types of frauds, overall accuracy must be good.

RELATED WORK

With credit card fraud being perpetrated on a massive scale around the world, the financial system is taking a major impact. In one study, the researchers are using a genetic algorithm to allow only legitimate clients to receive credit cards, and before purchasing anything from the online market, a classification is performed to detect fraudulent or genuine transactions. The transactions are also detected using the user's username and password. The second paper goes through all of the different sorts of fraud that can occur in the credit card sector, as well as the strategies that can be used to eliminate fraud from the banking industry. To reduce credit card theft, researchers employed a neural network.

Credit Card Detection Issues

There is a scarcity of study into real-world fraud detection issues. Credit card fraud has existed in this age of modern technologies due to a low rate of experimental analysis. The fundamental issue is that finance departments do not offer sensitive information to researchers in order for them to come up with a solution. Because only a small percentage of transactions are fraudulent, an effective classifier must be able to handle complex data. Because many transactions are identical, the classifier must be able to distinguish between correct and fraudulent transactions.

SJIF Impact Factor 2021: 8.013| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 7 | July 2021 - Peer Reviewed Journal

To detect new types of frauds, overall accuracy must be good.

EXPERIMENT AND ANALYSIS

The dataset was taken from kaggle and contains 284,807 transactions, 492 of which are fraud transactions with categories of 0 and non-fraud transactions with categories of 1. Because the dataset is skewed and unbalanced, the initial step is to scale and sample it to equalise fraud and non-frauds. Data is non-fraudulent in 99.8% of cases. We were unable to supply the original features due to PCA transformation, and the attributes are represented by the letter V from V1 to V28. The only attributes accessible are time and quantity. The amount is the transaction amount, and the time is the average time between two transactions. Class 1 will be for fraud cases, whereas class 0 will be for non-fraud cases.

The original dataset's class distribution is very unbalanced. Non-fraudulent transactions account for 99.83 percent of all transactions. Only 0.172 percent of transactions are fraudulent. We will receive a lot of errors if we run the models on this data, and the most of the True negatives will be missed. Because the classifier will treat fraud as non-fraud in this circumstance, accurate results will be displayed. Without detecting fraud cases, the forecasts will have a high accuracy rate.

RESULTS

The outcomes are surprising accurate. The contradictory findings will not precisely anticipate the outcome. These results are extremely accurate, yet they are useless in the actual world. As a result, we'll take a sample of minority classes. The data imbalance is the root of the problem.

First, we'll scale the Time and Amount column to the same size as the other column. We'll also subsample the data to ensure that we acquire an equal number of fraud and non-fraud cases because the original data frame was significantly imbalanced, which could lead to issues like overfitting and incorrect correlation.

Scaling the Data Set

The characteristics time and amount will be scaled in the same way as the other columns. A subsample of the data frame is also constructed to ensure that fraud and non-frauds are equally represented. There will be a 51 percent non-fraud and 49 percent fraud distribution in the subsample. Scaling is used to get rid of the overfitting. Because it is robust to outliers, the scaling is done with robust scalar. There will be 492 fraud cases and 492 non-fraud instances after completing rigorous scalar. We combined the 492 fraud cases with non-fraud cases to produce a new dataset. Because resilient scalar always gives the same approximation, the addition of any outlier has no influence on it. We'll perform random under sampling in the second step, but first we'll split the data into test and training sets.

Under Sampling

Under sampling is used to obtain more precise and balanced data, and it also aids in reducing overfitting. We'll start by determining how unbalanced our data is, then we'll determine how many transactions are considered non-fraud, and finally we'll balance the data by reducing the fraud ratio to a 50/50 ratio.

Learning Output

As the difference between the training and cross validation scores widens, the likelihood of our model being overfit grows, implying that there will be more volatility. Similarly, if both the training and cross validation scores are on the low side, it indicates that our model is underfit. In both sets, KNN had the highest score.

CONCLUSION

Many algorithms have been used to detect fraud, but none can detect 100% fraud. There are still issues, which we attempted to address in our work. We employed supervised machine learning algorithms to achieve credit card fraud detection using a dataset accessible on Kaggle in this research. Because the dataset was significantly skewed, our first objective was to sample it. On the majority class, which was non-fraud, we used random under sampling. We ran our supervised learning algorithms after attaining a 50/50 ratio of fraud and non-fraud. A subset of the dataset was constructed with an equal number of fraud and non-fraud cases.

The accuracy of logistic regression was 94.9 percent, decision tree accuracy was 91.9 percent, and random forest accuracy was 92.9 percent. KNN has a 93.9 percent success rate. Although logistic regression was more accurate, plotting the learning curves revealed that the majority of the algorithm was underfit, whereas KNN can only learn. As a result, KNN is a stronger classifier at detecting credit cards.

SJIF Impact Factor 2021: 8.013| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 6 | Issue: 7 | July 2021 - Peer Reviewed Journal

REFERENCES

1. *S P Maniraj,AdityaSaini (2019), "Credit card fraud detection using machine learning and data sciences", International journal of engineering research and technology.*
2. *Shiv Shankar Singh, (2019) "Electronic credit card fraud detection system by collaboration of machine learning models", International journal of innovative technology and Exploring Engineering.*
3. *Lakshmi S V S S,Selvani Deepthi Kavilla (2018), "Machine Learning for credit card fraud detection system", International Journal of Applied Engineering Research*
4. *Manirajsp,(2019)," Credit Card Fraud Detection using Machine Learning and Data Science"International journal of training and research.*
5. *vaishnavi nath(2019)," Credit Card Fraud Detection using Machine Learning Algorithms", International journal of engineering research and technology.*
6. *Suresh K Shirgave, Chetan J. Awati, Rashmi More, Sonam S. Patil (2019), "A review on credit card fraud detection using machine Learning, International journal of Scientific and technology research.*
7. *Daniyal Baig," Credit Card Fraud Detection Using Supervised Learning Algorithms"grin.*
8. *Samanesh Sorournejad, Zahra Zojaji, Reza Ebrahimi Atani, Amir Hassan Monadjemi, "A survey of credit card fraud detection techniques: Data and technique oriented perspective".*