ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/336639810

Credit Card Fraud Detection Using ANN

Article · March 2019

CITATION
1

READS
1,977

2 authors, including:

Peter Augustin
Christ University, Bangalore
10 PUBLICATIONS **65** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Deep Learning View project

Healthcare using IoT View project

All content following this page was uploaded by Peter Augustin on 18 October 2019.

The user has requested enhancement of the downloaded file.

Credit Card Fraud Detection Using ANN

Abdel Wedoud Oumar, Peter Augustin D

Abstract*:*Fraud on its own was and is devastating a lot of businesses, be them small or large. Particularly in the field of finance where we can see constant attacks on both individuals and enterprises alike. As such, credit cards are the most targeted as they are linked to both personal information and accounts. It is also evident to say that credit card fraud detection research is very much needed to deter and mitigate the impact of fraud on the financial field in general. It is important to identify frauds before it is too late so that the stolen credit card cannot be used for fraudulent transactions. To effectively detect these fraud transactions, we use a data consisting of fraudulent and non-fraudulent transactions to create a model that classifies these transactions with a high accuracy based on a machine learning technique. We used Artificial Neural Network with Logistic Regression to measure and in order to achieve high accuracy, we refined the parameters using the algorithms Back-propagation which has proved to have a high accuracy rate giving the model the ability to distinguish a fraudulent transaction from a normal one.

Index Terms*: *Artificial Neural Networks, Logistic Regression, Backpropagation, Credit Card Fraud.

I. INTRODUCTION

There's no lack of ways criminals commit fraud, there's insurance claim fraud, payroll fraud only to name a few. Fraud is something no business can claim that they are free of as every business has to deal with or at least has a financial unit where their main aim is to be fraud free and that concept itself is almost nonexistent as the more customers the company have the more the chances of some of them are actually fraudsters who aim to defraud the company.

There's no lack of ways criminals commit fraud. There are several types of fraud that are known to us, there's insurance claim fraud, payroll fraud only to name a few as well as certain types of frauds the certain industries suffer from. Such fraud types that occur in the retail business would include but not limited to employee discount, return product, loyalty points and gift cards. Due to fraud, some companies are hesitant to expand, therefore losing on potential profit. Fraud Detection can be viewed as the act of monitoring all transactions and the behavior of the cardholders in order to detect any sort of suspicious and out of the ordinary behaviors for the purpose of deciding that if it was the cardholder who made the transaction or a third-party fraudster.

In order to mitigate fraud and prevent it, a kind of fraud detection method has been popular with some financial institutes which involves checking the behaviors of the actual cardholder in order to set a pattern that way if a transaction is made using the cardholder's information and it does not fall in the same behavior pattern. We can therefore say it might be a fraudulent transaction. Most of the recently published work when it comes to the detection of credit card fraud is mainly to do with the close monitoring of behavior patterns of the cardholder in order to distinguish a normal transaction from a fraudulent one and therefore label them according to the patterns. However, this paper will be focusing on how detecting fraud through online purchase where fraud occurs the most when it comes to Card-not-Present (CNP) transactions (i.e. online transactions where the card doesn't need to be swiped). This paper proposes the use of an Artificial Neural Networks model that proved itself capable of achieving a high accuracy, the larger the data is. Assuming the company chose to implement the model within their platform as their customers' sensitive information exposed to potential fraudsters to defraud but with model implemented we can observe that the company will get a highly accurate probability of detecting fraud in their customers' transactions which effectively protects their customers' sensitive information as well as their money which in return builds trust between the company and the customers as it provides a safe platform for them to trust and for their daily needed transactions without any worries of being defrauded of their money.

II. RELATED WORK

This section discusses a few relevant fraud detection models proposed by various researchers for the fraud detection environment.

Vinicius Almendra et al., [1] have built a system based on the supervised learning approach to take in text comments and recognize in the textual comments written by buyers which states some common statements regarding seller behavior towards them and by combining all the data collected using the said system with publicly available data in order to detect fraudulent sellers by using supervised learning methods namely Bayes Classifier and Support Vector Machines, they have also proposed to classify suspended sellers as fraudsters or not based on the results of feedback comment classification.

In 2014, EvandroCaldeiraet al., [2] have proposed that due to different types of frauds, multiple ways can be used and they went a different way to try and predict fraud in online transactions, specifically in credit cards, as it is the most common form of fraud and unlike most, they applied four different techniques rather than one which is true in most cases, by applying the following techniques Bayesian Network, Logistic Regression, Neural Network and Random Forest on Brazil's most popular electronic payment service PagSeguro. In the best case they have achieved

Revised Manuscript Received on May 10, 2019

Abdel Wedoud Oumar, Computer Science,CHRIST (Deemed to be University), Bengaluru, India.

Dr.Peter Augustin D, Computer Science, CHRIST (Deemed to be University), Bengaluru, India.

a gain of 43.66%, which is no easy feat. Although, the datasets were only ranging over months, the results that they achieved present some significant gains when compared to the existing scenarios the company was adopting at the time.

AnusornCharleonnan [3] discussed the proposal of an ensemble model based on linear mapping, non-linear mapping and probability with RUS which is a data sampling technique which removes imbalances or removes examples in majority the majority class which is also called under-sampling by adjusting the class distribution of the training data set and for classifying class imbalance problems they used RUSMRN which is based on AdaBoost.M1 algorithm in order to build the ensemble of classifiers which meant to solve the class imbalance problems in the data and in this case it's used to detect fraud in credit cards datasets. The model's performance is measured by sensitivity, specificity and accuracy in the four techniques, RUSMRN, RUSBoost, AdaBoost and Naïve Bayes and they observed that RUSMRN model came out with the best result at 79.73% accuracy. RUSMRN which is based on the RUS data sampling technique and MRN algorithm which resulted in the RUSMRN which combines both data sampling and boosting to improve classification accuracy of unbalance characteristic data.

KosemaniTemitayo Hafiz et al. [4], in this paper, the authors after realizing that fraud increased although several fraud prevention technologies are already available, decided to focus on the creation of a scorecard from relevant evaluation criteria, features, and capabilities of predictive mainly used in fraud detection, the said scoreboard is meant to give a comparative look on five credit card predictive analytics vendor solutions already adopted in Canada. In this paper they studied the use of predictive analytics technologies in the context of fraud detection for the purpose of monitoring, detecting and distinguishing fraudulent transactions from normal ones.

Germ´an E. Melo-Acosta et al. [5], have proposed to use BRF (Balanced Random Forest) which can be used in both supervised and semi-supervised scenarios through a co-training approach and they proposed this methodology in order to detect credit card fraud automatically in the context of Big Data using Machine Learning techniques. The proposed system is meant to deal with three major challenges related to fraud detection datasets beside the face that they are rarely found there's also the problem of class imbalance, the inclusion of labelled and unlabeled samples. The proposed system resulted in a success and has overcome the said challenges.

John O. Awoyemi et al. [6], have done a comparative analysis on several majorly known and used algorithms namely Naïve Bayes, k-Nearest Neighbor and Logistic Regression as to ascertain which one would have a higher accuracy result and said algorithms are measured by Matthews Correlation Coefficient and Balanced Classification Rate and the fact that they are performing on a skewed dataset which means it's highly imbalanced as credit card fraud dataset are rarely available and the results were 97.92%, 97.69% and 54.86% for Naïve Bayes, k-Nearest Neighbor and Logistic Regression respectively, which in this case we can see that k-Nearest Neighbor has the highest accuracy.

David Kenyon et al. [7], Have investigated the use of Big Data and Data Science in order to predict insurance claim fraud as they used Big Data, Data Science and Predictive Analytics in a use-case to showcase that it's also possible to apply in the developing world. The proposed model was built using 34 distinct fields and a random sample of 50,000 short-term insurance claims all with Privacy Preserving Data-Mining technique and tested R, Hadoop and Hive which resulted in showing that we can Big Data and Data Science to predict and prevent insurance claim fraud.

In 2017, Richard A. Bauder et al. [8], have proposed a way that we can reduce the price of healthcare as fraud increases it therefore they made a compared several Machine Learning methods in order to detect fraud in the health industry. They used two different sampling methods and four performance metrics. Each of the ten models studied are based on supervised, unsupervised, and hybrid approaches and the supervised models use several algorithms including Gradient Boosted Machine (GBM), Random Forest (RF), Deep Neural Network (DNN), and Naive Bayes (NB). While the unsupervised methods include Autoencoder, Mahalanobis distance, k-Nearest Neighbors (kNN), and Local Outlier Factor (LOF) and the hybrid which includes a neural network model that is pre-trained using the unsupervised autoencoder and another method using a combination of multivariate regression and Bayesian probability. and are trained and tested on the 2015 Medicare PUF data with fraud labels obtained from the LEIE database. The results showed a successful detection of fraud while supervised methods performed the best compared to unsupervised and hybrid methods.

Balasupramanian.N et al. [9], have proposed that since every user at a transaction time have a specific pattern behavior as such they created patterns based on past purchases and as every user has a unique pattern, the proposed model will check every time the user requests a transaction which has to go through the bank, the trained model will check if patterns match and if so the transaction gets approved and if it doesn't match the user's purchasing pattern the model will suspect that it came from someone else therefore consider it a fraudulent transaction and it will be rejected and the card which the transaction used will be blocked. The framework proposed which used both Big Data and Machine Learning concepts and will take every user's past purchasing history and process it to optimally prevent online fraudulent transactions even before they happen.

In 2018, Shiyang Xuan et al. [10], have proposed the usage of two kinds of Random Forest algorithm for the purpose of training the model by feeding it both fraudulent transactions and normal ones. Different in their base classifiers, the two Random Forest algorithms are compared for the purpose of analyzing their performance on credit card detection. Although, they experimented with other known techniques such as support vector machine, naive Bayes, and neural networks, it was Random Forest they bested them in this case and the comparison between the said Random Forests resulted in RF2 has higher accuracy than its counterpart with 96.77% and 91.96% respectively as the only difference between the two Random Forest algorithms is

that the way of how nodes were split, in RF1 the data is distributed by comparing the distances between records and two centers while In RF2 the data is distributed according to the attribute which has minimum Gini impurity.

III. INFERENCE OF RELATED WORK

This section, includes the brief summary of the proposed models by other authors who are referenced in this paper.

In 2011, Vinicius Almendra, Denis Enchescu's paper aimed to deter fraud in online auction websites and achived a result that is able to detect fraud sellers based on the buyers' comments. EvandroCaldeira, Gabriel Brandao, Adriano C. M. Pereira's paper which analyzed an existing system and to predict fraud they went in and they have achieved a gain over the already adopted system. AnusornCharleonnan's paper proposed an ensemble model for fraud detection using RUS, MRN algorithms. KosemaniTemitayo Hafiz, Dr. Shaun Aghili, Dr. PavolZavarsky's paper focused on creation of a scorecard from relevant evaluation criteria, features, and capabilities of predictive analytics vendor solutions. David Kenyon, J.H.P Eloff's paper aimed to predict insurance claims fraud and looked at the usage of Big data to deter insurance claim fraud. John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare's paper compared well-known algorithms and made a wide study on major fraud detection to see which one performs better. German E. Melo-Acosta, Freddy Duitma-Munoz, Julian D. Arias-Londono's paper addressed credit card fraud detection problem in the context of Big Data, based on machine learning techniques. Richard A. Bauder ,Taghi M. Khoshgoftaar's paper aimed to deterred Medicare fraud using many known algorithms and compared several machine learning methods to detect Medicare fraud. Balasupramanian N., Ben George Ephrem, Imad Salim Al-Barwani's paper suggested a machine learning and big data analytics technique to detect and prevent any fraudulent online transaction. Shiyang Xuan, Guanjun Liu, Zhenchuan Li's paper aimed to detect fraud using two kinds of Random Forest as two kinds of Random Forest algorithms were implemented, and they achieved high accuracy on both kinds.

IV. PROPOSED MODEL

Credit Card Data

The data used in this study was collected from European cardholders which on September 2013. The data contains transactions made by these cardholders. The data is highly unbalanced as it contains very few fraudulent transactions with 0.172% of all the transactions being fraudulent. The data consists of total of 284807 transactions. The numerical inputs of the data are results of PCA transformation. The class column has values of 0 and 1 indicating whether the transaction is fraudulent or not respectively.

ANN and Logistic regression

There are various studies done on credit card fraud. machine learning classifiers such as random forest, support vector machines, et al. Each of these approaches has been looked into to create a model that classifies the fraudulent transactions from those non-fraudulent.

In our study, we use Logistic regression technique. It is a model that gives a probability for each of the classes under study, which is fraud or non-fraud classes in the data under study. Logistic regression takes uses the feature inputs linearly combined using weights as an equation very much like a linear regression except it is uses a sigmoid function to return a value between 0 and 1 which are the probabilities for each class. where the more the probability is close to 1 for a class, the more likely that the record belongs to that particular class. Linear regression is often used in binary classification like our case where we need to classify a transaction as a fraud or not but it can still be used for multi-classification problem and produce good results as well.

Artificial neural networks (ANN) are a group of nonlinear, statistical modelling techniques inspired from the human brain. In some cases, the classification of the records of the dataset cannot be done with a straight line, but it needs to be classified in a nonlinear manner. One of the things the artificial neural networks excel at is this particular problem. The deeper the nets, the more nonlinearity occurs. Though there are times when making the network deeper does not aid the performance or the effectiveness of the model, rather it just increases the training time.

ANN can be extremely helpful in modelling a complex transactional pattern. Therefore, they are very suited for our study on credit card fraud detection. A neural network is a structure of neurons where a neuron will take inputs, sum them and pass it to an activation function where the connections between neurons are associated with weights or coefficients. This network is structured in a such a way that there are layers connected in a systematic way. One layer being the input layer and one being the output layer and the others are hidden layers. Each layer is connected to the preceding layer and the following layer given there is one.

The connections associated with weights or coefficients are the keys to a better model. By adjusting them, while minimizing an error function the network keeps getting better for each iteration referred to as epoch.

In order to avoid critical problems as overfitting which is the case where the model starts to learn the data given to it too well that it becomes useless for real life applications, we make sure the number of epochs is neither too small nor too big.

Model Training and Evaluation

In the training phase, we started by defining the input and output layers where the input layer consists of the 27 features that will be fed to the model in order to learn what a fraudulent transaction is like. Giving two hidden layers with 20 and 15 nodes to the first and the second hidden layer respectively with RELU activation function which is helpful because it reduces the likelihood of the gradient to vanish which results in faster learning. Splitting the data into training set and test set and later into validation set to make sure our model is generalized and ready to use for any real-life application of it. The training phase is an important phase wherewe minimize the error function and get the best coefficients for our logistic regression using the backpropagation algorithm. We use different metrics in

our study, accuracy, precision and recall to understand the effectiveness of our model at predicting the fraudulent transactions.

Accuracy is the fraction of predictions that the classifier got correctly. It's found by dividing the number of correct predictions with the total number of predictions. It can also be expressed as

$$\left(\frac{True\ positives\ +\ True\ negatives}{Total\ number\ of\ examples}\right)$$

Precision identifies the frequency with which the model was correct when predicting the positive class. It can be expressed as

$$\left(\frac{True\ positives}{True\ positives\ +\ False\ positives}\right)$$

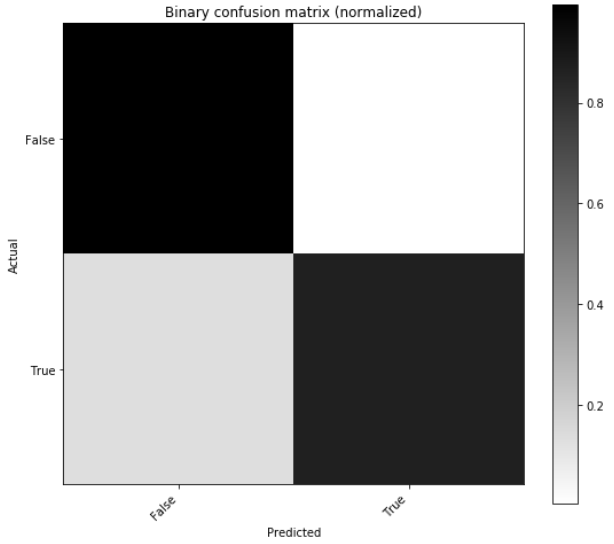
Recall is metric that answers the question: Out of all positive labels, how many did the model correctly identify? It can be expressed as

$$\left(\frac{True\ positives}{True\ positives\ +\ False\ negatives}\right)$$

V. RESULTS

The model was trained on 380950 samples of transactions and validated on 93987 samples per epoch and resulted in a very good accuracy of 0.9948, recall is 0.8639 and precision of 0.2134. The precision is quite low due to the percentage of fraudulent transactions in the dataset being very low of only 0.172%.

The following is the confusion matrix which is 3x3 matrix which summarizes how successful the model predictions were. One axis represents the actual label for the transaction and the other represents the predicted label of the transaction by the model in terms of true positive, true negative, false positive, and false negative.



VI. CONCLUSION

Credit card fraud is a serious issue when it comes to payment nowadays seeing how we are in the consumer era and to mitigate that particular problem, we built a model that can distinguish between a normal transaction from a fraudulent one. In this study, the model built using Artificial Neural Networks with Logistic Regression. We minimized the error function using the Back-propagation algorithm which resulted in a model that correctly predicts 99.48% of the time. We can conclude that this model is a generalized model and very reliable which can be used to detect any type of credit card fraudulent transactions.

REFERENCES

1. Vinicius Almendra, Denis Enchescu, "A supervised learning process to elicit fraud cases in online auction site", 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computin, 2011.
2. EvandroCaldeira, Gabriel Brandao, Adriano C. M. Pereira "Fraud Analysis and Prevention in e-Commerce Transaction", 9th Latin American Web Congress, 2014
3. AnusornCharleonnan, "Credit Card Fraud Detection Using RUS and MRN Algorithms", The 2016 Management and Innovation Technology International Conference (MITiCON-2016)
4. KosemaniTemitayo Hafiz, Dr. Shaun Aghili, Dr. PavolZavarsky, "The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada", 2016.
5. David Kenyon, J.H.P Eloff, "Big Data Science for Predicting Insurance Claims Fraud", 2017.
6. John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare , "Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis", 2017.
7. German E. Melo-Acosta, Freddy Duitma-Munoz, Julian D. Arias-Londono, "Fraud Detection in Big Data using Supervised and Semi-supervised Learning Techniques", 2017
8. Richard A. Bauder, Taghi M. Khoshgoftaar, "Medicare Fraud Detection using Machine Learning Methods", 16th IEEE International Conference on Machine Learning and Application, 2017
9. Balasupramanian N., Ben George Ephrem, Imad Salim Al-Barwani, "User Pattern Based Online Fraud Detection and Prevention using Big Data Analytics and Self Organizing Maps", International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2017
10. Shiyang Xuan, Guanjun Liu, Zhenchuan Li, "Random Forest for Credit Card Fraud Detection", 2018

AUTHORS PROFILE

Abdel Wedoud Oumaris doing his Masters of Science in computer science at CHRIST (Deemed to be University).He has obtained his degree of Bachelor of Computer Applications at CHRIST (Deemed to be University).His research interests include Deep Learning.

Dr.Peter Augustin Dis working as an associate professor in the department of computer science at CHRIST (Deemed to be University). He has obtained degrees of BSc in computer science and Masters of Computer Applications from ManonmaniamSundaranar University in 1997 and 2000. He obtained Ph. D. degree in 2016 from CHRIST (Deemed to be University). His research interestsinclude cloud computing and deep learning.

View publication stats