PREDICTING FRAUDULENT ACTIVITY IN CREDIT CARD TRANSACTION USING NEURAL NETWORK

A.G.Gayatri, Shaik Nafeesa, P.Anusha Student, Department of Computer Science, Narayana Engineering College Gudur
Dr.V.Sucharita Professor, Department of Computer Science, Narayana Engineering College Gudur
P.Madhavi Assistant Professor, Department of Computer Science, Narayana Engineering College Gudur

Abstract

Advances in communication technology and ecommerce have made the credit card a popular online payment method. Therefore, security in this system is highly anticipated to prevent fraudulent transactions. Credit card transaction fraud transactions are increasing year by year. In this way, researchers also experiment with novels to detect and prevent such fraud. However, there is always a need for some strategy to master this trick. This paper proposes a system to detect fraud in credit card details using an unregulated neural learning process (NN). The proposed method exceeds the existing methods of Auto Encoder (AE), Logistic Regression and K-Means clustering.The proposed NN fraud detection method works with 99.87% accuracy while the existing AE, Logistic Regression and K Means methods provide 92%, 97%, and 99.75% accuracy respectively.

Keywords: Unsupervised Learning, Supervised Learning, Fraud Detection, Auto-Encoders, Credit Card.CV.

Introduction

Credit card fraud can be defined as unauthorized use of customer card data to create purchases or withdraw money from a cardholder record. Misconduct starts with a credit card when a person improperly obtains the number written on the card or important records of the card to be used [1,2]. The cardholder, the card issuer and even the card issuer may not be notified of the fraud until the record is used for the purchase. With online purchases and online payments already in place, there is no longer a requirement for a valid credit card to make purchases. Fraud in online shopping programs is a hot topic these days. Fraud investigators, banking systems, and electronic payment systems such as PayPal must have an effective and sophisticated system to detect fraud to prevent rapidly changing fraud. According to a report by Cyber Source from 2017, the current loss of fraud per order channel, i.e., the percentage of fraud losses on their web store was 74 per cent and 49 per cent on their mobile channels. Based on this information, we can see what is wrong with all the patterns of fraudulent practices that have made changes in the past. The rise of E-commerce business has led to growth in the use of credit cards for online transactions and purchases. With the increase in credit card usage, the number of fraud cases has doubled. Credit card fraud is the practice of fraudulent payments without the knowledge of the cardholder.

Related work

The volume of internet users is increasingly causing e-commerce transactions to rise again[3]. We see the amount of fraud in online sales increasing again. Fraud prevention in e-commerce will be developed using machine learning, this task of analyzing the appropriate machine learning algorithm, algorithm to be used by Decision Tree, Naive Bayes, Random Forest, and Neural Network. The test result using a confusion matrix achieves the highest accuracy of the neural network by 96 percent, the random forest 95 percent, the Naïve Bayes 95 percent, and the resolution tree by 91 percent. The Synthetic Minority Over-sampling Technique (SMOTE) is able to increase the F1-Score average from 67.9% to 94.5% and the G-Mean rating from 73.5 percent to 84.6%. In the existing system, models are built based on Auto Encoder (AE), Logistic Regression and K-Means mergers to measure fraudulent and non-fraudulent transactions[11][12][4]. These methods provide low accuracy scores and remember scores and robustness due to the high calculation time.

Methodology

K-Means Clustering

This algorithm attempts to reduce the range of points in the collection with its centroid. The main purpose of the algorithm is to reduce the total distance between points and their centroid collection. Collection refers to the collection of data points that are grouped together due to certain similarities. We need to define the target number k, which refers to the number of centroids you need in the database. A centroid is an imaginary or real place representing a collection center. Each data point is assigned to each collection by reducing the number of squares within the collection[5][6].

In other words, the K-means algorithm identifies the k number of centroids, and then delivers the entire data point to the nearest collection, while keeping as few centroids as possible. Select the number of clusters k. Select k random points from data as centroids. Now, Assign all points to the nearest cluster centroid. Rewrite centroids for newly created collections. Repeat steps 3 and 4.

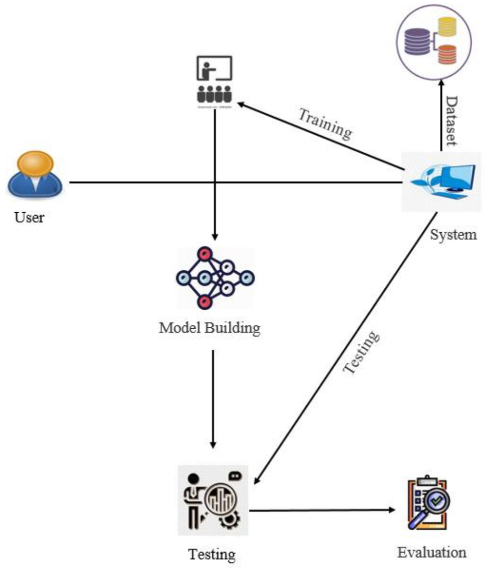


Fig 1.Workflow of credit card fraud detection system

Logistic Regression

Logistic Regression was used in biology in the early twentieth century. It was then used in many social science programs. Logistic Regression is used when dependent (targeted) variables are categorized. For example, Predicting whether an email is spam (1) or (0), whether Tumor is harmful (1) or not (0).

Consider a situation in which we need to distinguish whether the email is spam or not. If we apply line deflection to this problem, there is a need to set a limit depending on what segmentation can be made. State whether the actual category is negative, predicting a continuous value of 0.4 and the limit value is 0.5, the data point will be classified as incorrect which may lead to negative results in real time. In this example, it can be proved that the reversal of the line is not suitable for the problem of separation. The rotation of the line is infinite, and this brings the order of things into the image. Their value is from 0 to 1.

Types of Backbone

a. Binary Logistic Backlash

   The answer by category has only two possible outcomes. Example: Spam or not

b. Imultinomial Logistic Regression

   Three or more categories without order. Example: Predicting which foods are preferred (Veg, Non-Veg, Vegan)

c. Standard Operating Processes

   Three or more categories per order. Example: Movie rating from 1 to 5.

Auto Encoders:

Autoencoders are a form of feeding neural networks where input is similar to output. They press the input into a low-level code and recreate the output from this representation. The code is a compact "summary" or "compression" of the input, also called a hidden space representation. The autoencoder has 3 elements: encoder, code and decoder. The encoder presses the input and generates the code, decoder and rebuild the input using this code only. Autoencoders are mainly an algorithm for reducing (or reducing) that has a few key features:

Data explicit: Autoencoders can only logically compress data similar to the ones in which they are trained. Since they study certain aspects of the training data provided, they are different from the standard compression algorithm such as gzip. So we can't expect an auto-trained autoencoder in handwritten digits to compress landscape images.

Loss: The autoencoder output will not be exactly the same as the input, it will be a near-diminished representation. If you want to get lost, this is not the way to go.

Unsupervised: To train autoencoder we don't need to do anything desirable, just throw green input data into it. Autoencoders are considered an unregulated learning process because they do not need explicit labels to be trained in them. But to put it bluntly they are watching themselves because they are producing their own labels from training data.

Neural Networks:

A neural network is a series of algorithms that attempt to recognize the basic relationships in a set of data through a process that mimics how the human brain works. Neural Networks is used to solve many business problems such as sales forecasts, customer surveys, data verification, and risk management. It is used conceptually in a mathematical model. It contains a large number of interconnected substances called neurons to perform all functions. The information stored in neurons is basically the limited communication of neurons. Neural networks are a set of algorithms, modeledfreely on the back of the human brain, designed to detect patterns. They translate sensory data into machine understanding, label insertion or green input integration. The patterns they see are numerical, contained in vectors, in which all real data must be interpreted, be it images, sound, text or timeline[7][8][9].

Neural networks help us collect and classify. You can think of it as a layer of integration and further separation of the data you store and manage. They help to collect non-labeled data according to the similarities between sample inputs, and split data when they have a labeled database to train. (Neural networks can also extract features that other algorithms of integration and division; so you can think of deep neural networks as objects for large machine learning applications that include algorithms to enhance learning, fragmentation and deceleration.

Python Web Frameworks

A web framework is a code library that simplifies the life of a developer while building a reliable, scary and sustainable web system. Web frameworks include what engineers have learned over the past two decades while designing web sites and applications. The frameworks make it easy to reuse the standard HTTP operating code and build projects so that other engineers with framework knowledge can quickly build and maintain the application. Frames provide functionality in their code or extensions to perform the normal functions required to use web applications. These general functions include:

a. The URL line
b. HTML, XML, JSON, and other embedded output formats
c. Data management
d. Protection against cross-site forgery (CSRF) application and other attacks
e. Session and recovery session

Not all web frameworks include code for all of the above functionality. The framework falls into the scope of making a single use case to provide all the web framework features known to all developers. One frame takes the "battery-based" approach where everything that is possible comes with a bunch of frames while others have a small core package that can work on extensions provided by other packages.

When learning to use one or more web frameworks, it helps to have an idea of what the code under the covers does.

• Frame is a very well-crafted short video that explains how to choose between web frameworks. The author has some ideas about what should be the outline. For the most part I agree though I have found sessions and ORMs in the database to be a useful part of a framework where it is well done.

Results and analysis

From the below picture which is given as the input dataset, we can see how our proposed theory on credit card fraud detection has satisfied on use of different algorithms. The below figure represents the web application which has four options that includes Home, Load dataset, preprocess data, view data, select model and graphs.

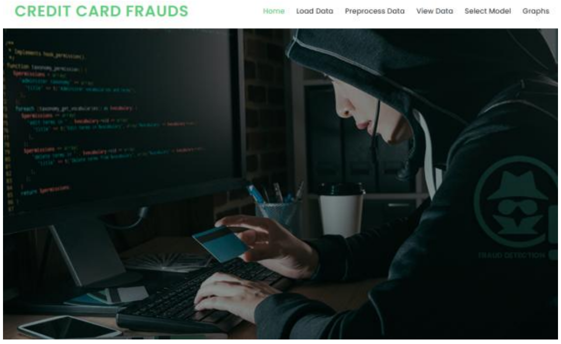


Fig 9. Web application home page

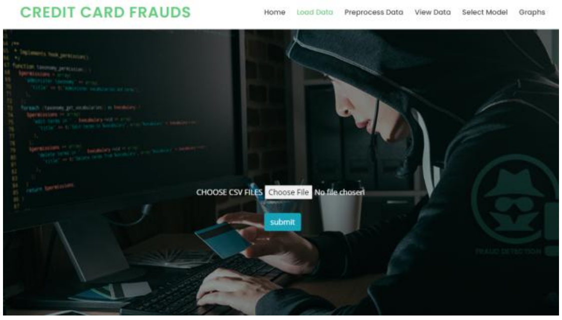
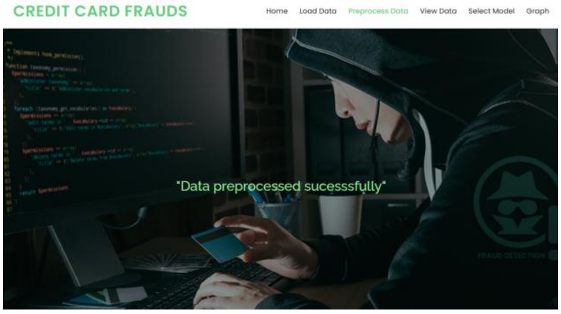
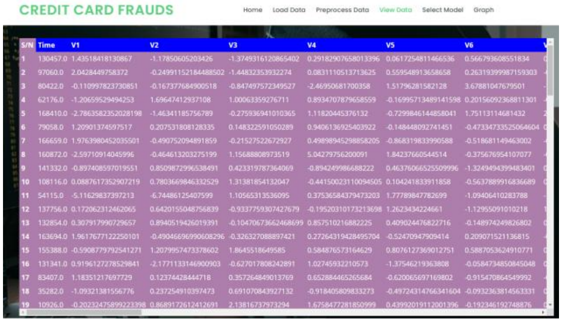


Fig 10. Load dataset(CSV file)



"Data preprocessed sucesssfully"

Fig 11.Preprocessing on data



S/N	Time	V1	V2	V3	V4	V5	V6
1	130457.0	1.43518418130867	-1.17850605203426	-1.3749316120865402	0.29182907658013396	0.0617254811466536	0.566793608551834
2	97060.0	2.0428449758372	-0.24991152184488502	-1.44832353932274	0.0831110513713625	0.559548913658658	0.26319399987159303
3	80422.0	-0.110997823730851	-0.167377684900518	-0.847497572349527	-2.46950681700358	1.51796281582128	3.67881047679501
4	62176.0	-1.20659529494253	1.69647412937108	1.00063359276711	0.8934707879658559	-0.16995713489141598	0.20156092368811301
5	168410.0	-2.7863582352028198	-1.46341185756789	-0.275936941010365	1.11820445376132	-0.7299846144858041	1.75113114681432
6	79058.0	1.20901374597517	0.207531808128335	0.148322591050289	0.9406136925403922	-0.148448092741451	-0.47334733525064604
7	166659.0	1.9763980452035501	-0.490752094891859	-0.21527522672927	0.49898945298858205	-0.868319833990588	-0.518681149463002
8	160872.0	-2.59710914045996	-0.464613203275199	1.15688808973519	5.04279756200091	1.84237660544514	-0.375676954107077
9	141332.0	-0.897408597019551	0.8509872996538491	0.423319787364069	-0.894249986688222	0.46376066525509996	-1.3249494399483401
10	108116.0	0.0887617352907219	0.7803669846332529	1.31381854132047	-0.44150023110094505	0.104241833911858	-0.5637889916836689
11	54115.0	-5.11629837397213	-6.74486125407599	1.10565313536095	0.37536584379473203	1.77789847782699	-1.09406410283788
12	137756.0	0.172062312462065	0.6420155048756839	-0.9337759307427679	-0.19520310173213698	1.2623434224661	-1.12955091010218
13	132854.0	0.307917990729657	0.8940519426019391	-0.10470673662468699	0.857510216882225	0.409024476822716	-0.148974249826802
14	163694.0	1.9617677122250101	-0.49046696990608296	-0.326327088897421	0.27264319428495704	-0.52470947909414	0.209071521136815
15	155388.0	-0.5908779792541271	1.2079957473378602	1.8645518649585	0.584876573164629	0.8076127369012751	0.5887053624910771
16	131341.0	0.9196127278529841	-2.1771133146900903	-0.627017808242891	1.02745932210573	-1.37546219363808	-0.0584734850845048
17	83407.0	1.18351217697729	0.12374428444718	0.357264849013769	0.652884465265684	-0.620065697169802	-0.915470864549992
18	35282.0	-1.09321381556776	0.237254910397473	0.691070843927132	-0.918405809833273	-0.49724314766341604	-0.0932363814563331
19	10926.0	-0.20232475899223398	0.8689172612412691	2.13816737973294	1.6758477281850999	0.43992019112001396	-0.192346192748876

Fig 12. View of dataset

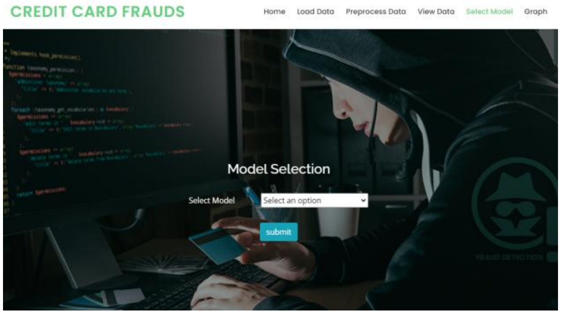


Fig 13. Selection of model

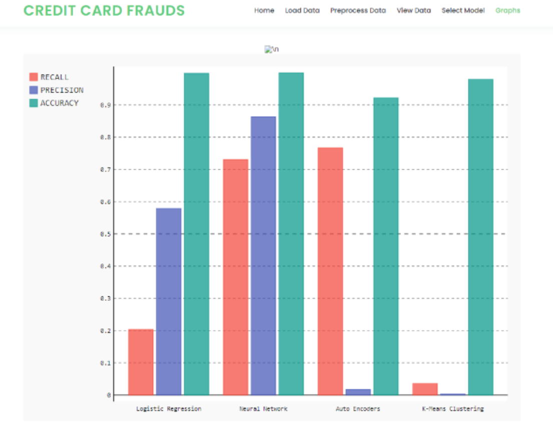


Fig 13. Performance on different algorithms applied on data based on precision, recall and accuracy.

The Dataset: Different algorithms were checked against the credit card dataset which has 31 labels. Here we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.It contains only numerical input variables which are the result of a PCA transformation.The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.It contains only numerical input variables which are the result of a PCA transformation. **Precision** (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while **recall** (also known as sensitivity) is the fraction of relevant instances that were retrieved.

Conclusion

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase**.** The accuracy of supervised algorithms that includes logistic regression and neural networks is 99.8010 and 99.9414 respectively. The unsupervised algorithms that includes Autoencoders and k-means clustering has accuracy of 92.1704 and 97.9506. Neural networks have efficiently worked on the dataset taken as it has highest accuracy.

REFERENCES

[1]. Thaha, Altyeb and Malebary, Sharaf. (2020). A Wise Way to Receive Credit Card Fraud Using a Light Gradient Boosting Machine. IEEE access. 8. 25579-25587.

[2]. Assaghir, Zainab & Taher, Yehia & Haque, Rafiqul &Hacid, Mohand-Said &Zeineddine, Hassan. (2019). An Experimental Study on Uncomplicated Approaches to Credit Card Fraud Detection Detection. IEEE access.

[3]. L. Meneghetti, M. Terzi, S. Del Favero, G. A Susto, C. Cobelli, "DataDriven Anomaly Recognition for Unsupervised Model-Free Fault Detection in Artificial Pancreas", Ieee Transactions On Control Systems Technology, (2018)) pages 1-15

[4]. V. Sucharita, S. Jyothi ,D.M. MamathaA Comparative Study on Various Edge Detection Techniques used for the Identification of Penaeid Prawn Species ,International Journal of Computer Applications (0975 – 887) Volume 78 – No.6, September 2013

[5].F. Carcillo, Y.-A. Le Borgne and O. Caelen et al., "Combining unsupervised and surveyed learning to detect credit card fraud", Information Sciences, Elsevier (2019), pages 1-15.

[6].Ashphak, Mr. & Singh, Tejpal &Sinhal, Dr. Amit. (2012). Examination of the Fraud Detection System Study using the Markov Hidden Credit Card Application Model Prof. Amit Sinhal. 1.

[7]. Renjith, Shini. (2018). Finding Fraudulent Merchants in Online Markets using the Support Vector Machine Approach. International Journal of Engineering and Technology Methods. 57. 48-53. 10.14445 / 22315381 / IJETT-V57P210.

[8]. V. Sucharita,S. Jyoti An Identification of Penaeid Prawn Species Based on Histogram Values, ,Volume 3, Issue 7, July 2013 ISSN: 2277 128X

[9].Saputra, Adi and Suharjito, Suharjito. (2019). Fraud Detection using Machine Learning in e-Commerce 10.14569 / IJACSA.2019.0100943.

[10] AK Rai and RK Dwivedi, "Credit Card Fraud Detection using the Unsupervised Machine Learning Based Scheme," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 421-426, doi: 10.1109 / ICESC48915.2020.9155615.

[11]S. Jyothi, V. Sucharita, D.M. Mamatha-A Survey on Computer Vision and Image Analysis based Techniques in Aquaculture CIIT International Journal of Digital Image Processing, 2013

[12].Venkateswara Rao, P., Ramamohan Reddy, A., Sucharita, V.An approach of detecting white spot syndrome of peaneid SHRIMP using improved FCM with hybrid back propagation neural network, International Journal of Pharmacy and Technology, 2016, 8(4), pp. 22351–22363