JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING ALGORITHMS

Vaibhav Takle
Computer Engineering
Modern Education Society's College of Engineering.
Pune, India.
vaibhavdtakle@gmail.com

Vallabh Nandal
Computer Engineering
Modern Education Society's College of Engineering.
Pune, India.
vallabhnandal11@gmail.com

Chetan Ujade
Computer Engineering
Modern Education Society's College of Engineering.
Pune, India.
ujadechetan@gmail.com

Priyanka Rupanawar
Computer Engineering
Modern Education Society's College of Engineering.
Pune, India.
priyarupanawar10@gmail.com

Prof. Shraddha Khonde
Computer Engineering
Modern Education Society's College of Engineering.
Pune, India.
Shraddha.khonde@mescoepune.org

Abstract :

Financial fraud is a growing problem with long term consequences in the financial industry and while many techniques have been discovered to solve this problem faced by various companies, data mining has been successfully applied to finance databases to automate analysis of huge volumes of complex data. In the identification of credit card fraud in online transactions, data mining has also played a significant role. Fraudulent transactions detection is a data mining problem. It becomes difficult for two reasons: first, normal and fraudulent behavioural patterns vary often, and second, credit card fraud data sets are extremely biassed. This paper investigates and checks the performance of Decision Tree, Random Forest, DNN , DNN with SMOTE, Logistic Regression and Logistic Regression with SMOTE on highly skewed credit card fraud data. European cardholders provided a credit card transaction dataset with 284,786 transactions. Both raw and pre-processed data are used in these procedures. The accuracy, sensitivity, and precision of the methodologies are used to assess their effectiveness. The results show that Classifiers such Decision Tree, Random Forest, Logistic Regression, and DNN, DNN with SMOTE, have the best accuracy.

I.INTRODUCTION

A credit card is a compact, thin plastic or fibre card that carries information about the individual, such as a photograph or signature, and allows the person identified on the card to charge products and services to his connected account, which is deducted on a regular basis.

ATMs, swiping machines, retail readers, banks, and online transactions all read card information these days. Each card has a unique card number, which is extremely significant; the card's security is based primarily on the card's physical security as well as the privacy of the credit card number.

The tremendous expansion in credit card transactions has resulted in a significant surge in fraudulent incidents. To detect fraud, a variety of data mining and statistical tools are employed. Artificial intelligence and pattern matching are used in several fraud detection strategies. It is critical to detect fraud using effective and secure ways.

Credit card fraud is on the rise, and financial losses as a result of fraud are skyrocketing. As new technology emerges, the Internet or online transactions are expanding in popularity. Credit cards account for the majority of these transactions. In 2018, London's credit card fraud losses were predicted to be at US$844.8 million. To decrease these losses, fraud protection or detection must be implemented. As technology develops at a quick pace, various forms of frauds emerge. So far, various machine algorithms have been developed to identify fraud, including a deep neural network and two machine learning models, which will be created to solve the problem and compare model results. Data sampling strategies will be utilized to improve the model.

II. PROBLEM STATEMENT

Card fraud is on the rise as a result of fraud. People who have defaulted on their credit cards are becoming more common. Billions of dollars are lost each year as a result of fraud. There is a scarcity of research to evaluate the fraud. A variety of machine learning approaches are employed to detect real-world credit card fraud. The use of ANN and hybrid algorithms is used.

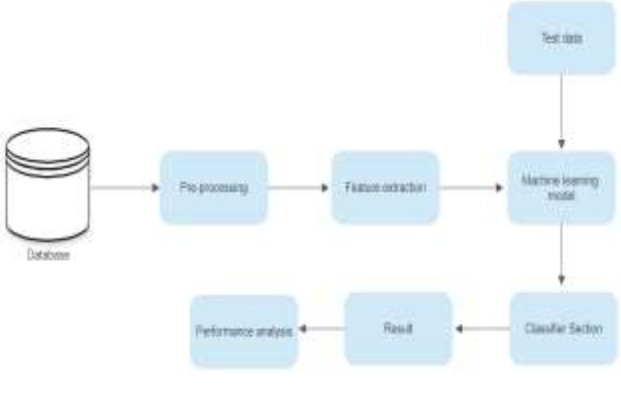
III. OBJECTIVES

The study's purpose is to employ machine learning algorithms to detect credit card fraud based on transaction time and quantity.

IV. PROPOSED SYSTEM

These are the proposed strategies for identifying credit card system frauds in this study. Different machine learning algorithms are evaluated to see which one is better for identifying fraud transactions and can be utilized by credit card merchants, such as Logistic Regression, Decision Trees, and Random Forest. Figure is an architecture diagram for illustrating the overall system framework.

Algorithm steps:
Step 1: Read the Dataset.
Step 2: Random Sampling is done on the data set to make it balanced.
Step 3: Separate the dataset into two parts: one for training and the other for testing.
Step 4: Accuracy and performance metrics has been calculated to know the efficiency for different algorithms.
Step 5: Then, based on efficiency for the provided dataset, find the optimum algorithm.

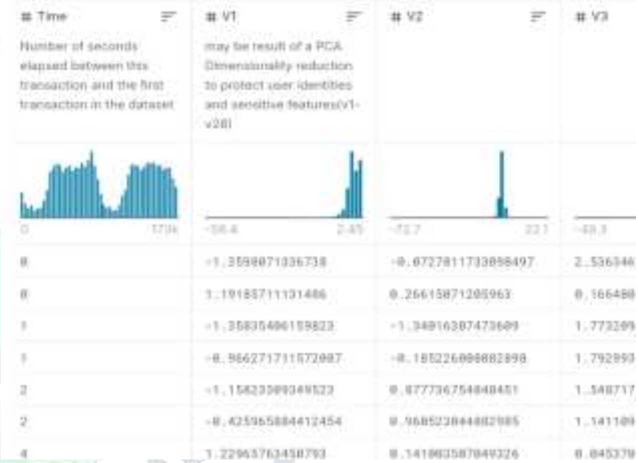


System Architecture

V. TRANSACTION DATABASE

This information includes transactions from Europe cardholders in September 2013. There are 492 fraudulent transactions out of a total of 2,84,807 transactions. Because there are fewer fraud cases than there are transactions, the data is unbalanced. The data set has been converted to a PCA transformation and solely comprises numeric values. Due to privacy and confidentiality concerns, numerous background details are withheld, leaving simply PCA converted data. Only time and amount are not PCA converted; all other supplied values (v1, v2, v3, v4, v5, v6, v7, v8, etc.) are PCA transformed numeric values. The fraud feature class has a value of 1 and the normal transaction has a value of 0.

Fig.



# Time	# V1	# V2	# V3
Number of seconds elapsed between this transaction and the first transaction in the dataset	may be result of a PCA Dimensionality reduction to protect user identities and sensitive features(v1-v28)		
0 172k	-56.4 2.45	-72.7 22.1	-48.3
0	-1.3598071336738	-0.0727811733098497	2.536346
0	1.19185711131486	0.26615071205963	0.166480
1	-1.35835406159823	-1.34016307473609	1.773209
1	-0.966271711572087	-0.185226008082898	1.792993
2	-1.15823309349523	0.877736754848451	1.548717
2	-0.425965884412454	0.960523044882985	1.141109
4	1.22965763450793	0.141003587049326	0.045370

credit card transaction dataset

VI. DESIGN AND IMPLEMENTATION OF ALGORITHMS

Understanding the issue statement and data, performing statistical analysis and visualization, and then assessing whether the data is balanced were the steps we used to forecast the outcome. The data in this collection is unbalanced, therefore it's balanced by oversampling, then scaled with standardization and normalization before being tested with several machine learning techniques. For any data science project, some packages are essential, such as Numpy, which is a numeric Python package, and pandas, and for data visualization, matplotlib and seaborn, which are both based on matplotlib, are utilized. In this project, Jupyter notebook is used to analyse the entire code, which allows the code to be displayed as a block of codes, making it easy to execute each portion and discover issues. The python sklearn package is used to create a user interface for training and testing the algorithms.

The data can be trained or tested using the test and train buttons.

A. Machine learning algorithms

1.Random forest

A random forest RF (conceptually) is a collection of decision trees. Each tree votes for a class, and the projected class is the one with the most votes. A decision tree is built on the whole dataset, while a random forest randomly selects features to build multiple decision trees and average the result.

2. Decision Tree

According to Wikipedia, a decision tree is a flowchart-like structure in which each internal node represents a feature test (e.g., weather if sunny or rainy) and each leaf node represents a class label created after all tests have been completed. A decision tree's purpose is to figure out how to divide datasets into groups based on conditions. As a result, it is non-parametric.

3. DNN with SMOTE

The simplest way of over-sampling is to duplicate data in the minority class, but no new information will be added to the model. Alternatively, we can synthesize data from existing ones, referred as Synthetic Minority Over-sampling, or SMOTE for short.

SMOTE, initially presented by Nitesh Chawla in 2002, works by selecting data that are close or similar in the feature space and drawing a line between data and make new data at a point on the line. It is effective because new data are close to the minority class in the feature space.

4. DNN with Under-sampling

The most basic technique is to choose the majority class at random to balance with the minority class. But the limitation is that data randomly removed from the majority class may be useful to create a robust model.

5. Logistic Regression

Logistic regression is a supervised classification method that uses the dataset's independent variable to compute the probability of a binary dependent variable. The chance of an outcome with two values, zero or one, no or yes, false or true, is predicted using logistic regression. The difference between logistic regression and linear regression is that logistic regression creates a straight line, whereas linear regression produces a curve. Based on the usage of one or more predictors or independent variables, logistic regression generates logistic curves that plot the values between zero and one.

A regression model with a categorical dependent variable that analyses the link between numerous independent factors is known as logistic regression. There are many different forms of logistic regression models, including binary, multiple, and binomial logistic models. Based on one or more factors, the binary logistic regression model is used to estimate the likelihood of a binary response.

VII. CONCLUSION

Credit card fraud is a prevalent problem that causes people to lose a lot of money as well as banks and credit card companies to lose money. This research aims to assist people in recovering their income, as well as for banked companies, by developing a model that can utilizing the timing and amount features in the Kaggel data set, you may more effectively discriminate between fraudulent and non-fraudulent transactions. To begin, we use supervised machine learning algorithms like logistic regression, decision trees, and support vector machines to build the model.

In order to solve this issue statement, we applied another aspect of artificial intelligence called time series analysis. In our current project, we used both time and quantity features to forecast if a transaction is fraudulent or not.

VIII. REFERENCES

[1] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and G. N. Surname, Random forest for credit card fraud detection, IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),2018.

[2] Satvik Vats, Surya Kant Dubey, Naveen Kumar Pandey, A Tool for Effective Detection of Fraud in Credit Card System, published in International Journal of Communication Network Security ISSN: 2231 1882, Volume-2, Issue-1, 2013.

[3] Rinky D. Patel and Dheeraj Kumar Singh, Credit Card Fraud Detection & Prevention of Fraud Using Genetic Algorithm, published by International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[4] Wen-Fang YU, Na Wang, Research on Credit Card Fraud Detection Model Based on Distance Sum, published by IEEE International Joint Conference on Artificial Intelligence, 2009.

[5] Salvatore J. Stolfo, Wei Fan, Wenke Lee and Andreas L. Prodromidis; "Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project"; 0-7695-0490-6/99, 1999 IEEE.

[6] S. Ghosh and D. L. Reilly, Credit card fraud detection with a neural- network, Proceedings of the 27th Annual Conference on System Science, Volume 3: Information Systems: DSS/ Knowledge Based Systems, pages 621-630, 1994. IEEE Computer Society Press.

[7]MasoumehZareapoor,Seeja.K.R, M.Afshar.Alam, Analysis of Credit Card Fraud Detection Techniques: based on Certain Design Criteria, International Journal of Computer Applications (0975 8887) Volume 52 No.3, 2012

[8] 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.

[9] G. Singh, R. Gupta, A. Rastogi, M. D. S. Chandel, A. Riyaz, "A Machine Learning Approach for Detection of Fraud based on SVM", International Journal of Scientific Engineering and Technology, vol. 1, no. 3, pp. 194-198, 2012, ISSN ISSN: 2277-1581.

[10] S. Patil, H. Somavanshi, J. Gaikwad, A. Deshmane, R. Badgujar, "Credit Card Fraud Detection Using Decision Tree Induction Algorithm", International Journal of Computer Science and Mobile Computing (IJCSMC), vol. 4, no. 4, pp. 92-95, 2015, ISSN ISSN: 2320-088X.

[11] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and G. N. Surname, “Random forest for credit card fraud detection”, IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),2018