INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-2083)
Available online at: https://www.ijariit.com

Credit card fraud identification using machine learning

Aishwarya Shirole
aishwaryashirole66@gmail.com
JSPM Narhe Technical Campus, Pune, Maharashtra

Riya Thakur
riyat721@gmail.com
JSPM Narhe Technical Campus, Pune, Maharashtra

Yash Singh
syash4069@gmail.com
JSPM Narhe Technical Campus, Pune, Maharashtra

Harsh Singh
singhharsh942@gmail.com
JSPM Narhe Technical Campus, Pune, Maharashtra

ABSTRACT

Today technology is increasing at very rapid pace, which can be used for good as well as for bad purposes. So with this growing technology e-commerce and online transactions also grown up which mostly contain transactions through credit cards. Credit cards help People to enjoy buy now and pay later for both online and offline purchases. It provides cashless shopping at every shop in all countries. As the usage of credit cards is increasing more, the chances of credit card frauds are also increasing dramatically. Credit card system is most vulnerable for frauds. These credit card frauds costs financial companies and consumers a very huge amount of money annually, fraudsters always try to find new methods and tricks to commit these illegal and outlaw actions. Online transaction fraud detection is most challenging issue for banks and financial companies. So it is much essential for banks and financial companies to have efficient fraud detection systems to reduce their losses due to these credit card fraud transactions. Various approaches have been found by many researchers till date to detect these frauds and to reduce them. Comparison of Local Outlier Factor and Isolation Factor algorithms using python and their detailed experimental results are proposed in this paper. After the analysis of the dataset we got the accuracy of 97% by Local Outlier Factor and 76% by Isolation Forest.

Keywords: *Fraud Detection, Local Outlier Factor, Credit Card, Dataset*

1. INTRODUCTION

Credit cards have been used in people's daily lives to buy. Buying can be done offline or offline. Millions and billions of transactions took place per minute everywhere on earth by using credit card. It offers online and offline electronic payment shopping with the option of ordering now and paying later. Credit card fraud is also on the rise(mostly) for this common use of credit cards or other banking card. Card theft is one of the biggest modern threats to companies.

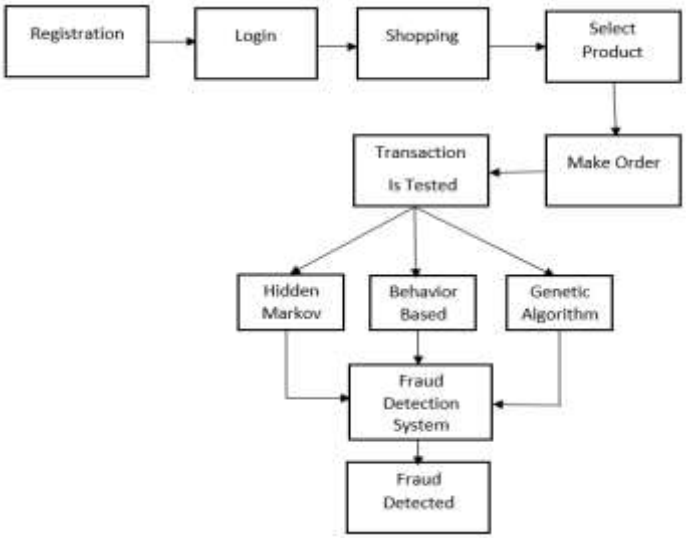


Fig. 1: Credit card fraud identification case diagram

Credit card fraud occurs either with a real card theft or sensitive account-related information, such as payment card number or other details is automatically accessible to the seller in the process of a legal purchase. Fraudsters use a whole range of methods to perform fraud. Damage resulting from this fraud it affects not only financial institutions but also consumers. The ID fraud rate remained strong until the mid-2000s, however during 2008 it increased by 21 points. Such fraud not only effect financial institutions but also consumers. This article analyzes the data set which is taken from Kaggle . The data set includes Credit Card purchases made by consumers in Europe during September 2013.A purchases done by Credit card are defined by tracking the conduct of purchases into two classifications: Fraudulent and Non-fraudulent by this two group correlations are formed and Machine learning algorithm is used to find fraudulent. The major problem is imbalanced data or skewed data i.e. Most purchases are not fraudulent, which renders it extremely challenging to identify fraudulent ones. Another big problem could be mislabeled records, because not every suspicious activity is detected and recorded. The fraudster used modern tactics against the system so that they can easily fraud. To overcome all this problem we used isolation forest and LOF (local outlier factor). Local Outlier factor is an algorithm used to find anomalies where as Isolation Forest algorithm is a supervised method for the classification

2. RELATED WORK

As the use of credit cards are increasing exponentially for both online and offline shopping, the frauds related with it are also increasing. Every day large number of people complaint against fraud transactions of their cards, there are many modern techniques such as genetic programming, data mining and many other techniques which are used for detecting fraud transactions. This paper [3] uses algorithms which consist of techniques for detecting optimal solution for the problem and implicitly generating result of fraudulent transactions. This paper has mainly focus on detecting fraudulent transactions and developing a method of generating test. Genetic algorithm is well-suited in detecting fraud. This method proves accurate in detecting the fraudulent transactions in very less time and reduce the number of false alert.

As data science is prominent as means of identifying fraudulent behavior, present-day methods depend on applying data mining techniques to the asymmetric datasets which contain sensitive variables. In paper [4] authors determine optimal algorithm for analysis as well as best performing combination of factors to detect credit card fraud. It has examine various classification model which were trained on a public dataset to analyse interrelation of certain factors with fraudulence. This paper has proposed effective metrics for searching out false negative rate and measured effectiveness of random sampling to reduce imbalance of the dataset. This paper has also determined best algorithms to utilize with high class imbalances and it was found that better performance rate for finding out credit card fraud under practical conditions can be achieved by Support Vector Machine because this algorithm analyses the purchase time in order to detect weather a credit card transaction is genuine or not more precisely. One of the statistical tools for scientists as well as for engineers to solve various types of problems is Hidden Markov Model. Paper [5] tells Credit card frauds during transactions can be detected using Hidden Markov Model. This model helps to get high fraud transaction coverage at very low false alarm rate and handling large volumes of transactions, hence providing a better and convenient method to detect credit card frauds and giving better and faster results in less time. Using this model customers transaction pattern is analyse and any changes or deviation from regular pattern is considered as non genuine transaction or fraud transaction . It makes detection handling very effective and try to eliminate the complexity. This paper has also described how theycan detect whether an inbound transaction is fraudulent or not and stated that many additional security features like MAC address detection and also shipping address verification are provided for enhanced security and better detection of fraud transaction. In Paper [6] Local Outlier Factor using MATLAB is used in the model to solve the credit card fraud detection for both online and offline transactions and used purchasing amount as the examination of frauds is proposed. They have performed analysis on two datasets and accuracy obtained for dataset 1 is 60-69%, for dataset 2 it is 96% with variation in neighbors. Paper [7] has used standard models wiz NB, SVM, and DL as well such as hybrid machine learning models as Ada Boost and majority voting methods to detect credit card fraud. These models are applied on a publically available credit card data set to evaluate model efficiency. Then they have analyzed real world dataset from a financial institution. To assess the robustness of algorithms further they have added noise in data samples and finally proposed that majority of voting method achieves good accuracy rates in detecting fraud cases in credit cards by comparing the values which is generated by Matthews Correlation Coefficient (MCC) metrics and used as performance measure for these algorithms. 0.823 is The best MCC score , achieved using majority voting. A perfect MCC score of 1 has been pull off using Ada Boost and majority voting methods for real credit card data set which is taken from a financial institution. After adding the noise from 10% to 30% in the majority of voting method has yielded the best MCC score of 0.942 for 30% after evaluation. In past few years, it has become very difficult task for banks to detect credit card frauds. For frauds in the credit card system. Machine learning plays a very important role. Banks are using different methodologies of machine learning for predicting such frauds. Bank have been collecting past transactions data and used new features for enhancing predictive power of algorithms. Sampling approach on data-set, selection of variables and detection techniques which are used highly affect the performance of fraud detection in credit card transactions. Paper [8] has examined the performances of Random Forest, Decision Tree and Logistic Regression using R language on the data-set which is obtained from Kaggle for detecting frauds in the credit card system. The data-set contains a total of 2,84,808 credit card transactions of a European banks data set. Fraud transactions are considered to be ―positive class‖ and genuine ones as ―negative class‖. This data-set is highly imbalanced, having about 0.172% of fraud transactions and the rest are genuine transactions. To balance these data, they performed oversampling on the data-set, which resulted in 60% of fraud transactions and 40% as genuine ones. For distinct variables the efficiency of the techniques used is based on accuracy, specificity, sensitivity and error rate. The accuracy results shown in the division of Logistic Regression, Decision Tree and Random Forest are 90.0, 94.3, 95.5 respectively, and these comparative results show that Random Forest has a higher level of performance compared to Logistic Regression and Decision Tree. For data mining association rules are considered to be the best studied model. This article [9] proposes the use of association rules on credit card data-set which is obtained from some important companies in Chile, to extract information in order to detect common misconduct transactions from a credit card database in order to detect and prevent fraud. This model helps to make the results more accurate by increasing the execution time, reducing the use of excessive rules and overcoming the difficulties of low

support and confidence. Paper[10] focuses on the real time frauds detection and presents new and innovative approaches in understanding spending patterns to decipher potential fraud cases. It uses self-organizations map to decipher, filter and analyse customer behavior for detection of fraud. Fraudulent e-card transaction are among foremost prevailing and disturbing activities occurring in commercial business. Finding suspicious and suspicious transaction paper [11] deals with surveillance based on separation in particular. When pre-analyzing the database using standard and Principal element Analysis analysis, all classifiers achieved more than 95.0% accuracy compared to the results achieved prior to pre-data analysis

Problem Statement
Credit card fraud detection becomes difficult to detect because credit card information sets are extremely skewed and imbalance. Because of Skewed and scattered data point it is extremely difficult to detect fraud transaction. In this model we will try to figure out which transaction is fraud and which is not.

Methodology
Data collection is checked and payments are marked as a criminal or legal. In this paper we have used two ways to detect fraud for the proposed model on the Kaggle data set for detecting frauds in credit card system using python .Because it is not possible to get a real-time database, we tend to use credit card database csv from Kaggle i.e. includes 2,83,807 entries. Variety parameters used in the database time, category, value, location & etc. The same type of 30 parameters is used. V1 -V28 are the fields of a Decreasing the size of the PCA to protect identity and sensitive user options. there performance were compared. These is compared to determining which algorithms are offered better results .Therefore , a test is performed to determine which algorithms offered accurate results and can be used to detect credit fraud.

A **Local Outlier Factor (LOF)**
Local Outlier Factor (LOF):Anomaly score of every sample is named the local Outlier factor. It calculates the local deviation of the density of a given sample in relation to its neighbors. It's known as local because the anomaly score depend on the object isolation, the thing is with relation to the encircling neighborhood.

B **Data set**
In this paper we have analyzed the dataset which is taken from Kaggle. The data set is in CSV format it contains 284,807 Credit card transactions By monitoring the behavior of the transactions are characterized into two categories fraudulent and non fraudulent. Original feature and more background information are not provided in the data set because of confidentiality issues. Features like V1, V2V3 ... V27 are the principal components obtained with PCA, the features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Class' is the response variable and it takes value 1 in case of fraud and otherwise 0.

We will import our the dataset from a csv file as Pandas Data Frame. And then, we will begin to explore the data-set to gain understanding of the type, quantity, and distribution of data in our data-set. For this purpose, we will be using Pandas' built-in describe feature.

3. MODELING & ANALYSIS

3.1 Requirements

Application Used : Python
Operating System : Windows 10
Tools-
List of tools used to assess credit card fraud detection analysis is as follows: This proposed model is used in Python, Numpy and Pandas are used for simple tasks like data storage and to perform operation. Matplotlibrary is used for visualizing data and to plot graph. Other tools used a sci-kit.3.2.

Phases of project
We are completing this fraudulent activity visual activity in the next three phases,
1) Data Checking Steps:
a) Upload the database
b) Database processing
c) Do a graph d) Show the database

2) Data Processing Steps:
a) Upload the database
b) Remove Null values
c) Divide the database
d) Go to the training phase

3) Data Classification:
a) Train the database
b) Improve classification
c) Isolation Forest
d) Perform Classification

4. EXPERIMENTAL SETUP AND RESULTS

A. Evaluation Metrics :- Many classification tasks use simple evaluation metrics such as Accuracy to compare performance between models, because accuracy is simple measure to implement and generalizes to more than just binary labels. But there is one major drawback of accuracy that it is assumed that there is an equal representation of examples from each class, and for skewed datasets like in our case accuracy is a misleading factor. It does not provide accurate results. So accuracy is not a correct measure of efficiency in our case. To classify the transactions as fraud or non-fraud we need some other standards of correctness which are as :

1 Precision
2 Recall
3 F1-score
4 Support

- These all standard of correctness are depend upon the Actual and Predict class,
- True Positive (TP): These values are correctively predicted positive that means value of both actual class and predicted class are YES.
- True Negative (TN): These values are correctively predicted Negative that means value of both actual class and predicted class are NO.
- False Positive (FP): when value of actual class is NO and value f predicted class is YES.
- False Negative (FN): when value of actual class is YES and value f predicted class is NO.
- False Positive and False Negative classes occur when actual class contradicts with predicted class.

Precision: It is Ratio of correctly predicted Positive observations to the Predicted positive observations

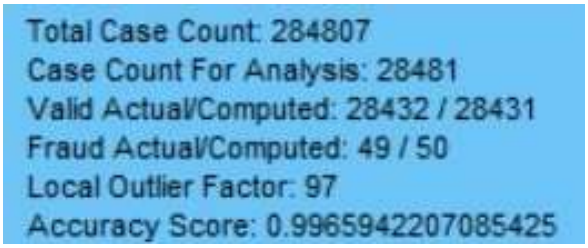
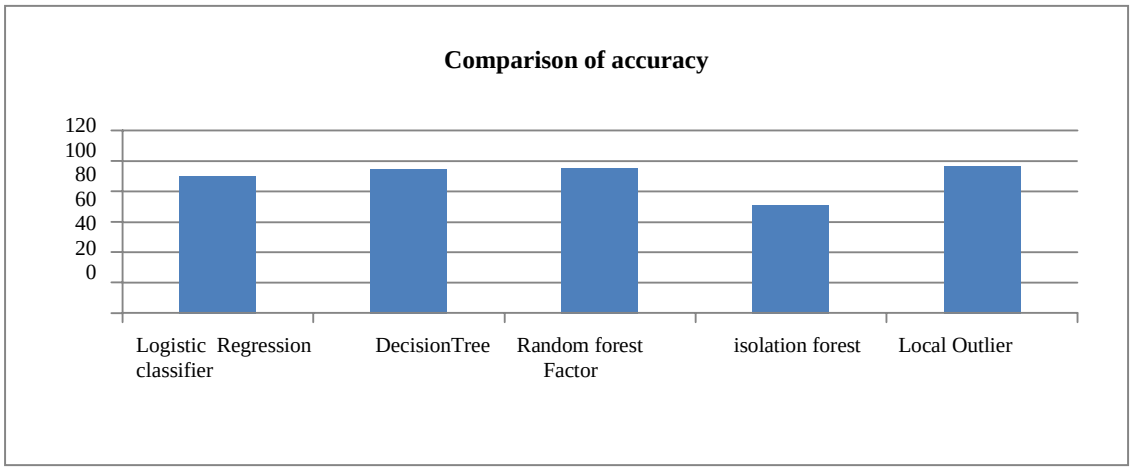


Table 1: Comparisn of various algorithms

Algorithm	Accuracy
Logistic Regression	90.0%
Decision Tree	94.3%
Random Forest Classifier	95.5%
Isolation Forest	71%
Local Outlier Factor	97%



Graph Depicting accuracy of various detection algorithms

5. CONCLUSION

Finding fraudulent credit card transactions is really important, especially in today's society. There are lots of methods to capture these instances, and it's really cool to see how companies deal with this on a day-to-day basis. The chances of credit card frauds increases massively with the increase in usage of credit cards for transactions. A study of credit card fraud detection on a public available data-set using Machine Learning algorithms such as Isolation Forest and Local outlier factor has been presented in this paper. The proposed system is implemented using PYTHON. On analysis the dataset Local outlier factor gave highest accuracy rate of 97% followed by the Isolation forest 76%

6. REFERENCES

[1] Nilsonreport.com. (2019). [online] Available at: https://nilsonreport.com/upload/content_promo/The_Nilson_Report_10-17-2016.pdf [Accessed 6 May 2019].

[2] Machine Learning Group, ―Credit Card Fraud Detection,‖ Kaggle,23-Mar-2018. [Online].Available:https://www.kaggle.com/mlgulb/creditcardfraud. [Accessed: 06-May-2019].

[3] I. Trivedi, M. M, and M. Mridushi, ―Credit Card Fraud Detection,‖ Ijarcce, vol. 5, no. 1, pp. 39–42, 2016.

[4] R. Banerjee, G. Bourla, S. Chen, S. Purohit, and J. Battipaglia, ―Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection,‖ pp. 1–10, 2018.

[5] T. Patel and M. O. Kale, ―A Secured Approach to Credit Card Fraud Detection Using Hidden Markov Model,‖ vol. 3, no. 5, pp. 1576–1583, 2014.

[6] D. Tripathi, T. Lone, Y. Sharma, and S. Dwivedi, ―Credit Card Fraud Detection using Local Outlier Factor,‖Int. J. Pure Appl. Math., vol. 118, no. 7, pp. 229–234, 2018.

[7] C. P. Lim, M. Seera, A. K. Nandi, K. Randhawa, and C. K. Loo, ―Credit Card Fraud Detection Using AdaBoost and Majority Voting,‖ IEEE Access, vol. 6, no. 11, pp. 14277–14284, 2018.

[8] I. Sohony, R. Pratap, and U. Nambiar, ―Ensemble learning for credit card fraud detection,‖ vol. 13, no. 24, pp. 289–294, 2018.

[9] D. Sánchez, M. A. Vila, L. Cerda, and J. M. Serrano, ―Association rules applied to credit card fraud detection,‖ Expert Syst. Appl., vol. 36, no. 2 PART 2, pp. 3630–3640, 2009.

[10] J. T. S. Quah and M. Sriganesh, ―Real time credit card fraud detection using computational intelligence,‖ IEEE Int. Conf. Neural Networks -Conf. Proc., vol. 35, pp. 863–868, 2007.

[11] H. A. Shukur, ―Credit Card Fraud Detection Using Machine Learning methodologies,‖ vol. 8, no. 3, pp. 257–260, 2019

AUTHORS PROFILE

Snehal Shinde
JSPM Narhe Technical Campus, Pune, Maharashtra, India

Aishwarya Shirole
JSPM Narhe Technical Campus, Pune, Maharashtra, India

Riya Thakur
JSPM Narhe Technical Campus, Pune, Maharashtra, India

Yash Singh
JSPM Narhe Technical Campus, Pune, Maharashtra, India

Harsh Singh
JSPM Narhe Technical Campus, Pune, Maharashtra, India

