Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest

Article *in* INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING · April 2019

DOI: 10.26438/ijcse/v7i4.10601064

2 authors, including:

Sameena Naaz
Jamia Hamdard University
63 PUBLICATIONS **194** CITATIONS

SEE PROFILE

IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest

Hyder John[1], Sameena Naaz[2*]

[1,2] Dept. of Computer Science and Engineering, School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi, India

[*]*Corresponding Author: snaaz@jamiahamdard.ac.in, Tel.: +91-9899368267*

DOI: https://doi.org/10.26438/ijcse/v7i4.10601064 | Available online at: www.ijcseonline.org

Accepted: 18/Apr/2019, Published: 30/Apr/2019

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, Isolation Forest, Local Outlier Factor, Credit card, Dataset.

I. INTRODUCTION

Credit cards are used for shopping in everyday life. This shopping can be both offline and online for purchasing goods or services. It provides cashless shopping for online and offline shopping with the benefit of buy now and pays later. With this predominant usage of credit cards, credit card fraud is also growing (proportionally). Credit Card Fraud is one of the biggest threats to business establishments today. Credit card fraud begins either with the theft of the physical card or with the important data associated with the account, such as card account number or other information that necessarily be available to a merchant during a permissible transaction. The fraudsters employ a large number of techniques to commit frauds. The losses which occur due to these frauds do not only affect financial companies but affects the customers a lot tot. As per the United States Federal Trade Commission report till the mid-2000s the theft rate of identity was holding stable, but it got increased by 21% during 2008. According to the Nilson Report [1], worldwide losses from card fraud rose to US$21 billion in 2015, up from about US$8 billion in 2010. By 2020, that number is expected to reach US$31 billion. In order to reduce the losses which occur due to these credit card frauds, we need to build efficient fraud detecting systems.

In this paper analysis of the dataset is performed which is taken from Kaggle [2]. The dataset contains Credit card transactions which are made by customers during September 2013 in Europe. By monitoring the behaviour of the transactions Credit card transactions are characterized into two categories fraudulent and non-fraudulent. Anomalies are created based upon these two classes and use machine learning algorithms to detect the fraudulent transactions. Then by using Local Outlier Factor and Isolation Forest the behaviour of these anomalies can be analysed and compare their final results to check which algorithm is best.

Rest of the paper is organized as follows: Discussion of existing literature is done in section 2. Section 3 gives the details of the methodologies used in this work, the experimental setup and the results are discussed in section 4. Finally section 5 gives the conclusions that can be drawn from this work.

II. RELATED WORK

As the usage of credit cards is 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 data mining, genetic programming etc. which are used for detecting fraud transactions. This paper [3] uses genetic algorithms which consist of techniques for detecting optimal solution for the problem and implicitly generating result of fraudulent transactions. This paper has mainly focussed 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 short span of time and minimizing the number of false alert.

As data science is emerging as means of identifying fraudulent behaviour, present-day methods depend on applying data mining techniques to the skewed datasets which contain confidential variables. In paper [4] authors determined optimal algorithm for analysis as well as best performing combination of factors to detect credit card fraud. It has inspected various classification models which are trained on a public dataset to analyse interrelationship of certain factors with fraudulence. This paper has proposed better metrics for finding out false negative rate and measured the effectiveness of random sampling to diminish imbalance of the dataset. This paper has also determined best algorithms to utilize with high class imbalances and it was found that Support Vector Machine algorithm had highest performance rate for detecting credit card fraud under realistic conditions because this algorithm analyses the purchase time in order to detect weather a credit card transaction is fraudulent or not more accurately .

One of the statistical tools for scientists and engineers to solve various types of problems is Hidden Markov Model. Paper [5] states 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 analysed and any deviation from regular pattern is considered as fraudulent transaction. It makes detection handling very easy and tries to eliminate the complexity.
This paper has also described how they can 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] the model to solve the credit card fraud detection for both online and offline transactions by Local Outlier Factor using MATLAB and used purchasing amount as the examination of frauds is proposed. They have performed analysis on two datasets and accuracy for dataset 1 is 60-69%, for dataset 2 it is 96% with variation in neighbours.

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 dataset to evaluate model efficiency. Then they have analysed 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 voting method achieves good accuracy rates in detecting fraud cases in credit cards by comparing the values generated by Matthews Correlation Coefficient (MCC) metrics which is adopted as performance measure for these algorithms. The best MCC score is 0.823, achieved using majority voting. A perfect MCC score of 1 has been achieved 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 voting method has yielded the best MCC score of 0.942 for 30% after evaluation.

In past few years it has become very difficult for banks to detect credit card frauds. For detecting frauds in credit card system machine learning plays an important role. Banks are using various methodologies of machine learning for predicting these frauds. Banks have collected past data of transactions and used new features for enhancing predictive power of algorithms. Sampling approach on dataset, 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 performance of Decision Tree, Random Forest and Logistic Regression using R language on the dataset which is obtained from Kaggle for detecting frauds in credit card system. The data set contains a total of 2,84,808 credit card transactions of a European bank data set. Fraud transactions are considered as "positive class" and genuine ones as "negative class". This dataset is highly imbalanced, having about 0.172% of fraud transactions and the rest are genuine transactions. To balance the dataset they performed oversampling on the dataset, which resulted in 60% of fraud transactions and 40% as genuine ones. For different variables the performance of the techniques used is based on sensitivity, specificity, accuracy and error rate. The accuracy results shown for Logistic Regression, Decision Tree and Random Forest classifier are 90.0, 94.3, 95.5 respectively, and these comparative results show that the Random Forest has higher performance rate as compared to Logistic Regression and Decision Tree.

For data mining association rules are considered to be best studied models. This article [9] proposes the use of association rules on credit card dataset which is obtained from some important companies in Chile, in order to extract knowledge so that normal behaviour patterns may be obtained in unlawful transactions from credit card transaction database in order to detect and prevent fraud. This model helps to make the results more intuitive by optimizing the execution time, reducing the use of excessive generation of rules and overcomes the difficulties of minimum support and confidence.

Paper[10] focuses on real-time fraud detection and presents a new and innovative approach in understanding spending patterns to decipher potential fraud cases. It makes use of self-organization map to decipher, filter and analyse customer behaviour for detection of fraud.

Fraudulent e-card transactions are among foremost prevailing and disturbing activities occurring in commercial business. To detect suspicious and non-suspicious transactions paper[11] discusses the supervised based mostly classification. When pre-processing the dataset using normalization and Principal element Analysis, all the classifiers achieved over 95.0% accuracy compared to results reached before pre-processing the dataset.

III. METHODOLOGY

We analyse the dataset and classify the transactions as fraud or legit. In this paper we used two different algorithms for our proposed model on the Kaggle dataset for detecting frauds in credit card system using python. Which are briefly explained below and compared their performance. Comparison are made for these algorithms to determine which algorithms gives better results and can be adapted by credit card merchants for identifying frauds.

A. Local Outlier Factor

In 2000 M. Breunig, Hans-peter Kriegel, Raymond T. Ng and Jörg Sander introduced the Local Outlier Factor (LOF) algorithm to find the anomalous data points by measuring the local deviation of a given data point with respect to its neighbours.

Outliers based on the local density are detected using this algorithm [12]. Locality is given by nearest neighbours and density is calculated by their distance. By comparing the local density of an object to the local densities of its neighbours, one can identify regions of similar density, and points that have a substantially lower density than their neighbours. The data point is considered as an outlier if it has very small density as compared to its neighbours.

Outlying patterns may be divided into two types: global and local outliers. The object which is significantly having a large distance with respect to its k-th neighbour is called Global outlier while as object whereas a local outlier has a distance to its k-th neighbour that is large relatively to the average distance of its neighbours to their own k-th nearest neighbours.

B. Isolation Forest

Isolation forest is a tree-base model that is developed to detect outliers [13]. This algorithm is based upon the fact that anomalies are the data points which are few and different. These properties result in susceptible mechanism to anomalies which is known as Isolation. This method is basically different from all other existing methods and is highly useful. To detect the anomalies rather than the basic distance and density measures it introduces the use of isolation as an efficient and more effectively. This algorithm has small memory requirement and low linear time complexity. It builds a good performing model with a small number of trees using small sub-samples of fixed size, regardless of the size of a data set.

C. Dataset

In this paper we have analysed the dataset which is taken from Kaggle[2]. The dataset is in CSV format (creditcard.csv), it contains Credit card transactions which are made by customers during September 2013 in Europe containing 284,807 transactions. By monitoring the behaviour of the transactions Credit card transactions are characterized into two categories fraudulent and non-fraudulent. Original features and more background information are not provided in the dataset because of confidentiality issues. Only numerical input variables are provided which are the results of Principal Component Analysis (PCA) Transformation. Features V1, V2 ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

D. Tools

The list of tools used to explore credit card fraud detection analysis is as:

This proposed model is implemented in Python. Numpy and Pandas are used for simpler tasks such as data storage and transformation. For data analysis and visualization Matplotlib is used. Seaborn is used for statistical data visualization and for algorithms we used Skitlearn.

IV. 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:

- Precision
- Recall
- F1-score
- Support

These all standard of correctness are depend upon the Actual and Predict class, so we draw a 2×2 confusion matrix to know more about them

Table 1: Confusion Matrix

		Predicted Class	
		Negative	Positive
Actual Class	Negative	True Negative (TN)	False positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

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.

$$Precision = \frac{TP}{TP \times FP}$$

Recall: it is ratio of correctly predicted positive observations to the all observations in actual class YES.

$$Recall = \frac{TP}{TP + FP}$$

F1 Score: It is the weighted average of Precision and Recall. Therefore this score takes both false negatives and false positives into account.

$$F1\ Score = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

Support: it is number of occurrences of each class in correct target values.

Isolation Forest: 71
0.99750711000316

Table 2: Values calculated by Isolation forest

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	28432
1	0.28	0.29	0.28	49

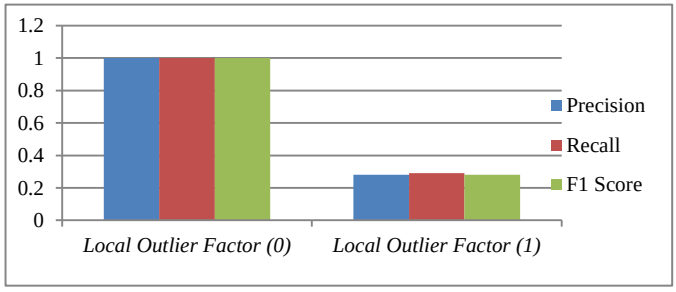

Fig1: variation in precision, recall and f1-score

Local Outlier Factor: 97
0.9965942207085425

Table 3: Values calculated by local outlier factor

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	28432
1	0.02	0.02	0.02	49

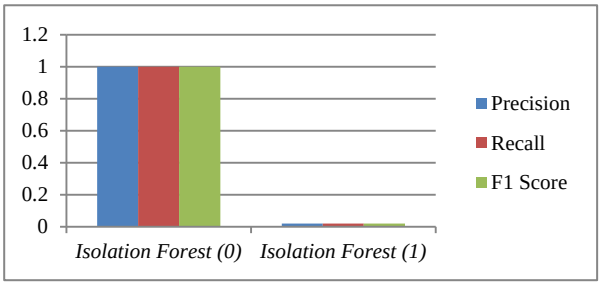

Fig2: variation in precision, recall and f1-score

B. Experimental Results

By comparing the results of Local Outlier Factor and Isolation Forest algorithms, from the above table it is clear

that the Local Outlier Factor is best for detecting the frauds in credit card transactions with the accuracy of 97%.

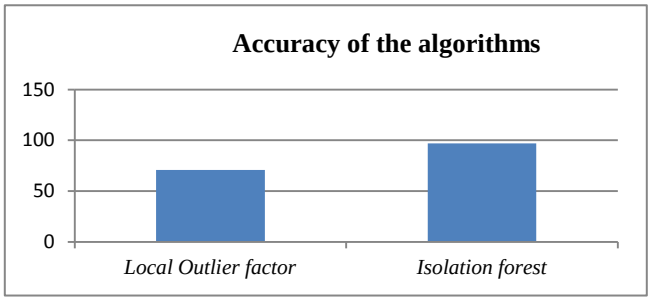

Fig3: Accuracy values of LOF and Isolation Forest

We have also made the comparison of thee algorithms with some other algorithms and the values are given in below table:

Table 4: Comparison 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%

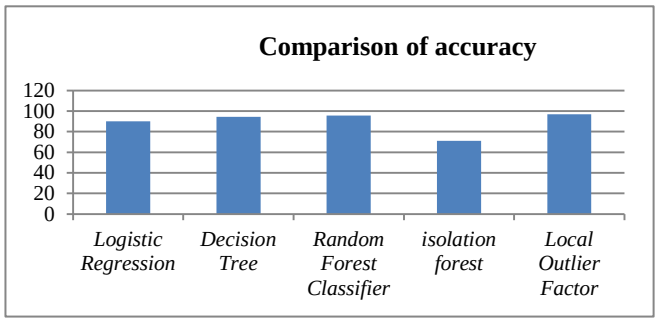

Fig 4: graph depicting accuracy of various fraud detecting algorithms.

V. CONCLUSION AND FUTURE SCOPE

Chances of credit card frauds are increasing massively with the increase in usage of credit cards for transactions. A study of credit card fraud detection on a publically available dataset using Machine Learning algorithms such as Local outlier factor and Isolation Forest has been presented in this paper. The proposed system is implemented in PYTHON. On analysing the dataset Local outlier factor gave highest accuracy rate of 97% followed by the Isolation forest 76%.

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.

[12] "Local outlier factor", *En.wikipedia.org*, 2019. [Online]. Available:https://en.wikipedia.org/wiki/Local_outlier_factor.[Accessed: 06- May- 2019].

[13]. "Isolation forests for anomaly detection improve fraud detection.", *Blog Total Fraud Protection*, 2019. [Online]. Available: https://blog.easysol.net/using-isolation-forests-anamoly-detection/. [Accessed: 12- Apr- 2019].

Authors Profile

Hyder John has completed B.TECH computer sciences and engineering from Islamic University of Science and Technology Awantipora, J&K in 2017. He is currently pursuing MTECH computer sciences and engineering from Jamia Hamdard New Delhi. His research interests are Machine learning, Data mining and Artificial Intelligence.

Dr. Sameena Naaz is working as an Assistant Professor (Senior Grade) in the Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India. She has a total experience of more than 18 years with one-year overseas experience. She received her Bachelor of Science (Computer Engg.) from Aligarh Muslim University, in 1998 and the M.Tech. degree from Aligarh Muslim University, in 2000. She completed her Ph. D from Jamia Hamdard in the field of distributed systems in year 2014. Sameena Naaz has published several research articles in reputed International Journals and Proceedings of reputed International conferences published by IEEE and Springer. Her research interests include Distributed Systems, Big Data, Cloud Computing, Data Mining and Image Processing. She is life member of Indian Society for Technical Education (ISTE) and a member of International Association of Computer Science and Information Technology (IACSIT). She serves as reviewer of various Journals of International repute. She is also member of program committee of various reputed International conferences. She is in editorial Board of some reputed Intentional Journals in Computer Sciences.

