Isolation Forest and Local Outlier Factor for Credit Card Fraud Detection System

V. Vijayakumar, Nallam Sri Divya, P. Sarojini, K. Sonika

Abstract: ***Fraud identification is a crucial issue facing large economic institutions, which has caused due to the rise in credit card payments. This paper brings a new approach for the predictive identification of credit card payment frauds focused on Isolation Forest and Local Outlier Factor. The suggested solution comprises of the corresponding phases: pre-processing of data-sets, training and sorting, convergence of decisions and analysis of tests. In this article, the behavior characteristics of correct and incorrect transactions are to be taught by two kinds of algorithms local outlier factor and isolation forest. To date, several researchers identified different approaches for identifying and growing such frauds. In this paper we suggest analysis of Isolation Forest and Local Outlier Factor algorithms using python and their comprehensive experimental results. Upon evaluating the dataset, we received Isolation Forest with high accuracy compared to Local Outlier Factor Algorithm***

Keywords: ***anomaly detection, isolation, local outlier, fraudulent, credit card.***

I. INTRODUCTION

Credit cards have been used in people's everyday lives to go shopping. For buying products and services this purchasing can be offline as well as internet. It offers online and offline electronic payment shopping with the option of ordering now and paying later. Credit-card fraud is also growing (substantially) with this prevalent use of credit cards Credit Card Theft is one of contemporary biggest risks to corporate institutions [1]. Credit card fraud occurs either with actual card stealing or with sensitive account-related details, such as payment card number or any other information that is automatically accessible to a dealer in the process of a legal purchase. The frauds use a whole range of methods to conduct fraud. The damages that arise as a consequence of such frauds impact not only financial institutions but also the consumers. According to the U.S. Federal Trade Commission survey, the identity fraud rate stayed steady until the mid-2000s, but throughout 2008 it rose by 21 points.such frauds impact not only financial institutions but also the consumers [1]. As per the Nilson Report [1], card fraud losses globally rose to US 21 billion dollars in 2015, up from around US dollars. This article analyzes the dataset which is taken from Kaggle [2]. The dataset includes Credit Card purchases made by consumers in Europe during September 2013. Credit card purchases are defined by tracking the conduct of purchases into two classifications: fraudulent and non-fraudulent. Depending on these two groups correlations are generated and machine learning algorithms are used to identify suspicious transactions. Instead, the action of such anomalies can be evaluated using Isolation Forest and Local Outlier Factor and their final results can be contrasted to verify which algorithm is better.

The key problems involved in the identification of credit card fraud are: Immense data is collected on a regular basis and the model construct must be quick sufficiently respond to the scandal in time. Imbalanced data, i.e. most purchases are not fraudulent, which renders it extremely challenging to identify fraudulent ones. Data transparency is important because the data is still confidential. Another big problem could be mislabeled records, because not every suspicious activity is detected and recorded. The fraudsters used advanced tactics against the system [3].

To handle these challenges, we go for the following: The model used would be easy and accurate sufficiently identify the phenomenon and recognize it as a suspicious activity as easily as possible. Imbalance can be done by the correct application of certain techniques. The dimensionality of the data should be minimized to preserve the user's privacy. It is important to take a more legitimate source which will cross-check the results, at least for model training. We will keep the model easy and interpretable, and we will get a fresh model up and running to implement as the fraudster adapts to it with only a few changes.

We used the isolation forest and local outlier factor. Isolation Forest algorithm is a supervised method for the classification. It is used for issues of both regression and classification kinds. Local Outlier factor is an algorithm used to find anomalies. The two main categories of outliers are outlier regional, and outlier local. Diverse systems produce a huge volume of data continuously. Outlier detection is a data analysis strategy whose activity is altered from usual activity or planned behavior. This paper reflects on the static and streaming data identification techniques. The work often focuses on different identification methods at local and global stage. The remainder of the paper is structured as follows: Section 2 addresses current literature. Section 3 includes descriptions of the methodologies used in this study, the experimental design and the findings are described in section 4.

Revised Manuscript Received on March 17, 2020.

*Correspondence Author

Dr. V. Vijayakumar*, Computer Science and Engineering Department, Sri Manakula Vinayagar Engineering College, Puducherry, India, Email: vijayakumarv@smvec.ac.in

Nallam Sri Divya, Computer Science and Engineering Department, Sri Manakula Vinayagar Engineering College, Puducherry, India, Email: nallam.nr@gmail.com

The conclusions that can be taken from this research are finally presented in section 5.

II. RELATED WORK

When the utilization of credit cards for both online and offline shopping grows exponentially growing, the frauds connected with it are also rising. Each day a huge amount of people talk about their bank fraud purchases, there are many new methods used to identify fraud purchases, such genetic engineering, data mining etc. This paper [4] utilizes genetic algorithms comprising of techniques for predicting the best solution to the problem and extracting tacit findings from fraudulent transactions. This work concentrated primarily on identifying illegal transactions and creating a test generation methodology.Genetic models are well designed for fraud detection. This approach proves effective in identifying very short time fraudulent activity and reducing the amount of false alarms.

When data analysis evolves as a way of detecting deceptive activity, existing approaches remain based on the application of data processing strategies to distorted databases comprising sensitive variables. In paper [5] authors defined the ideal computational method as well as the best-performing combination of variables to identify credit card fraud. It has examined specific classification models based on a general dataset to examine the interconnections with fraud of some variables. This paper suggested improved measures for detecting false negative levels and assessed the efficiency of randomized sampling to decrease data set variance. The article also defined the best algorithms to use for large-class imbalances, and it was noticed that the Support Vector Machine has the best success rating for detecting financial fraud in practical circumstances as this algorithm analyzes the payment period in order to identify the environment suspicious or not more effective a credit card transaction.

Secret Markov Model is one of the mathematical methods for engineers and scientists to overcome the specific sorts of problems. Paper [6] notes that bank card scams can be identified during purchases using the Hidden Markov Model**.** deviate further apart. This model aims to obtain broad fraud activity coverage at a relatively small false alert rate and manage huge amounts of purchases, thereby offering a simpler and more efficient means of identifying credit card frauds and delivering clearer and quicker outcomes with less time. Using this model, transaction behavior for consumers is evaluated and any divergence from standard pattern is called fraud. The paper further explained how to determine not whether the incoming transaction is illegitimate and mentioned that certain additional protection features such as MAC address identification and mailing address authentication are offered for improved security and stronger fraud detection.

In Paper [7] the concept for solving counterfeiting detection by Local Outlier Factor both for offline and online purchases utilizing MATLAB and the payment number used as the fraud test is suggested. They conducted analyzes on two samples, and data set 1 precision is 60-69 percent, data set 2 is 96 percent with community variance. Paper[8] used default models like NB, SVM, and DL models as well as advanced machine learning models such as Ada Boost, and ranked voting approaches to identify credit card fraud. In order to further test the scalability of the algorithms, they introduced noise in data tests and finally suggested that the proportional voting approach achieves high accuracy levels in the identification of cases of fraud in credit cards by analyzing the values produced by the parameters of Matthews Correlation Coefficient (MCC) embraced as a quality measure for such algorithms. The highest ranking of MCC is 0.823, achieved by supermajority support. Use Ada Boost and plurality voting strategies with actual credit card data collection obtained from a bank's, a total MCC score of 1 has been achieved. Upon incorporating the disturbance in the plurality voting system from 10 percent to 30 percent, it produced the highest MCC result of 0.942 upon assessment for 30 percent.

It has been really challenging for banks to track payment card scams over the last few years. Machine learning plays an significant part in identifying fraud of credit-card scheme. Banks use different machine learning techniques to forecast such frauds. Banks also gathered past purchase details and used modern technologies to improve algorithm explanatory power.Dataset sampling strategy, collection of variables and identification methods that are used significantly influence the efficiency of fraud detection during credit card purchases. Paper [9] analyzed the output of Random Forest and Logistic Regression using R language on the Kaggle dataset for predicting financial fraud.

The data collection contains a minimum of 2,84,808 payment card purchases with a range of data from a European Financial institution. Scam transactions are deemed to be optimistic, and legitimate transactions to be bad. This dataset is somewhat imbalanced, with around 0.172 percent of payments containing theft and the remainder being legitimate transactions.

They conducted over-sampling on the dataset to align the data collection, resulting in 60 percent as scam transactions and 40 percent as legitimate transactions. The efficiency of the methods employed is dependent on flexibility, precision, consistency and error rate for various variables. The consistency figures reported for the grouping of Decision Tree, Logistic Regression and Random Forest are 90.0, 94.3, 95.5 respectively and these comparison findings indicate that the Random Forest has a better success as relative to the Logistic Regression and Decision Tree. Laws related to the relationship of data mining are deemed better learned models.

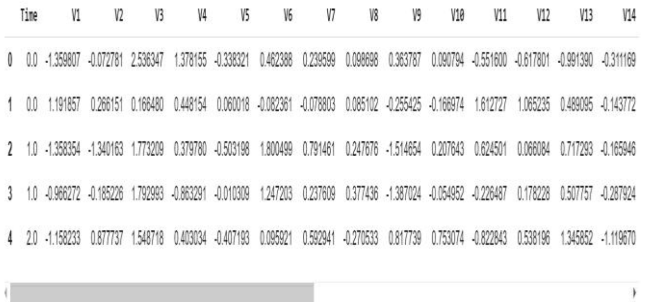
This article [10] suggests the usage of credit card registry association guidelines acquired from certain big Chilean firms to collect information such that regular activity trends can be retrieved from the bank transaction database in illegal transactions to track and deter fraud. This model aims to render the outcomes more understandable by maximizing the implementation period, the usage of needless rule creation and addressing the constraints of limited support and trust amounts of labeled data already on more complex datasets. Their results suggest semi-supervised methods [11].

III. METHODOLOGY

The data collection is evaluated and the payments are marked as scam or legal. In this paper we used two separate methodologies to detect frauds in credit card framework using python for our new methodology on the Kaggle dataset. These are discussed briefly below and their efficiency contrasted. These algorithms are compared to decide which algorithms offer better results and can be adapted as shown below. With such algorithms, a test is made to decide which algorithms offer stronger results and can be modified to detect fraud by credit card dealers.

A. Dataset analyzing and preprocessing

We evaluated the sample taken from Kaggle in this paper [2]. The report is in CSV type (creditcard.csv), it includes credit card purchases comprising 284,807 payments made by consumers in Europe throughout Sept 2013. Credit card purchases are defined by tracking the purchase activity into two types: fraud and non-fraudulent purchases. According to security concerns original functionality and further context details are not included in the training data. The findings of the Principal Component Analysis (PCA) Conversion are given with only quantitative input variables. Traits V1, V2... V28 (Fig 1) are the main components obtained with PCA, only' Time' and' Number' are the properties not converted with PCA.



	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	0.090794	-0.551600	-0.617801	-0.991390	-0.311169
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	-0.166974	1.612727	1.065235	0.489095	-0.143772
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	0.207643	0.624501	0.066084	0.717293	-0.165946
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	-0.054952	-0.226487	0.178228	0.507757	-0.287924
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	0.753074	-0.822843	0.538196	1.345852	-1.119670

Fig 1: Analyzing the dataset

The' Time' attribute includes the seconds from each transaction to the first transaction in the dataset. The' Amount ' task is the Amount fee, this functionality may be used for value-sensitive learning, for example. App' Rank' is the answer vector and in the case of theft it takes value 1 and 0 otherwise.

Just 0.17 per cent of the sales are illegal. The evidence were extremely skewed. First let's add our models without optimizing them and if we don't get a strong consistency then we will find a way to fit this dataset. Because we can easily see from this, for the dishonest ones the typical money expenditure is higher. This renders this topic critical to solving.

Graphically, the matrix of correlation as shown in Fig 2 provides one an understanding of how characteristics interact with each other, which may help one determine which features are more important to the forecast.

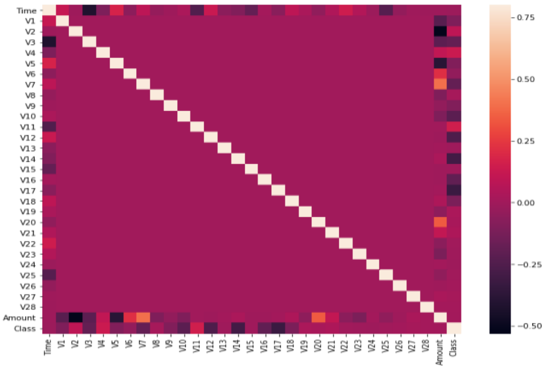


Fig 2: Correlation matrix for features co-relation

B. Training the dataset

The dataset obtained from Kaggle is trained using two algorithms one is Isolation Forest and one more is Local Outlier Factor Algorithm. The results are compared between these two algorithms.

Local Outlier Factor

Hans-peter Kriegel,M. Breunig, Raymond T. Ng and Jörg Sander implemented the Local Outlier Factor (LOF) algorithm for finding abnormal data points by evaluating the local variability of a given data point in comparison to its neighbors. Local-density outliers are observed with this algorithm [13]. Locality is defined by nearest neighbors and distance is used to measure density. When matching an object's local density with its neighbors ' local densities, one can distinguish areas with similar consistency, and points that have a substantially lower density than their neighbors. The data point is called an outlier because opposed to its surroundings it has very low scale.

External trends can be classified into 2 sorts: global outliers and local outliers. The entity which has a considerable distance from its k-th neighbor is called Global outlier while as an entity whereas a local outlier has a distance from its k-th neighbor which is large compared to its neighbors ' average distance from its own k-th closest neighbors.

Isolation Forest

Isolation Forest is a tree-based model capable of detecting outliers [14]. This approach is compounded by the fact that the data points are the anomalies that were few and many. Such properties originate in system that is vulnerable to phenomena known as Isolation. This approach is significantly different from all other approaches already in use and is extremely useful. It promotes the use of insulation as an inexpensive and more reliable to locate the irregularities instead of the usual distance and density controls. This algorithm has a small memory demand and a low complexity in linear time. It constructs a good reliable model, utilizing small sub-samples of fixed size with a specific number of trees, irrespective of a data set.

Tools

The set of methods used to evaluate the study of credit card fraud identification is as follows: This suggested approach is built in Python.

Numpy and Pandas are used for simplified tasks like the storing and manipulation of data. Matplotlib is used for the interpretation and visualisation of results. Seaborn is used for the analysis of statistical information and we used Sckitlearn for algorithms.

IV. EXPERIMENTATION AND RESULTS

The comparison results and performance metrics for different algorithms i.e., Local Outlier Factor and Isolation Forest are shown below

A. Evaluation Metrics

Most labeling activities utilize basic measurement measures such as accuracy to evaluate results between templates, since precision is a simple measure to apply and generalizes to more than binary labels. But one major downside to consistency is that it is presumed that there is an equal representation of instances from each class, and a limiting consideration for distorted data points like in our case. It fails to deliver accurate data. So in our situation, precision is not an appropriate measure of efficiency. We need some other norms of correctness to categorize the payments as fraud or non-fraud that are as follows:

Precision: Percentage of accurately expected Positive findings to the Positive Findings foreseen.

Recall: It is the percentage of accurately expected affirmative findings to all actual class Valid observations.

F1 Score: Accuracy and Recall is measured average. The ranking also takes into account all false negatives and false positives.

Support: The number of instances in the relevant target values for each class is.

The isolation forest showed the total number of errors as 71 and the accuracy was 99.72 percent while Local Outlier Factor showed the total number of errors as 107 and Accuracy as 99.62 percent. Accuracy is not a good metric for anomaly detection. It is important to look at precision, recall and f1-score. The precision=0.02, recall=0.02 and f1-score=0.02 are very low for Local outlier Factor as shown in Table 2 and fig 4. This suggests there is 2% chance of actually predicting a fraudulent transaction and there is 2% chance for a predicted fraudulent transaction to be actually true.

Table 1: 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

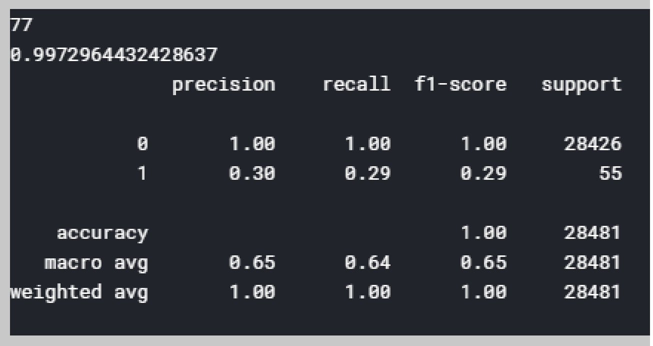
The precision = 0.3, recall = 0.29 and f1-score = 0.29 for Isolation Forest as shown in table 1 and fig 3 is better compared to Local Outlier Factor.

Table 2: 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

B. Experimental Results

When looking at the results of Local Outlier Factor and Isolation Forest algorithms, it is obvious from the above table that the Isolation Factor is better observed with an accuracy of 97 percent in online transactions.

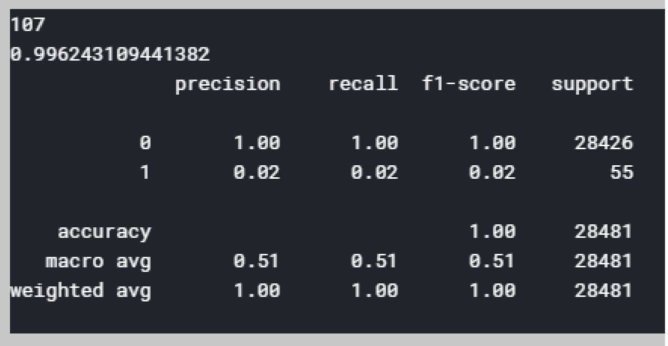


```
77
0.9972964432428637
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28426
           1       0.30      0.29      0.29        55

    accuracy                           1.00     28481
   macro avg       0.65      0.64      0.65     28481
weighted avg       1.00      1.00      1.00     28481
```

Fig 3: Results obtained with Isolation Forest



```
107
0.996243109441382
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28426
           1       0.02      0.02      0.02        55

    accuracy                           1.00     28481
   macro avg       0.51      0.51      0.51     28481
weighted avg       1.00      1.00      1.00     28481
```

Fig 4: Results obtained with Local Outlier Factor

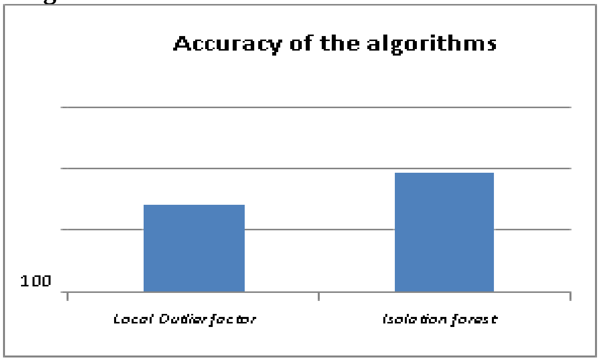


Fig 5: Accuracy values of used algorithms

V. CONCLUSION

It is essential for credit card businesses to be able to recognize fraudulent credit card transactions so that consumers are not paid for things they have not purchased. With the growing use of credit cards for purchases, the risks of credit card frauds grow rising significantly. In this paper an analysis of credit card fraud identification was described on a publicly available dataset utilizing Machine Learning techniques such as Local outlier factor and Isolation Forest. In PYTHON the framework introduced is enforced. When analyzing the data set Isolation Forest provided the highest precision rate than Local Outlier Factor algorithm.

Our future working will be with Neural Networks for efficient finding of fraud when deployed in the any financial institution server.

REFERENCE

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.
3. https://www.geeksforgeeks.org/ml-credit-card-fraud-detection/
4. I.Trivedi and M.Mridushi, ―Credit Card Fraud Detection, Ijarcce, vol. 5, no. 1, pp. 39–42, 2016.
5. R. Banerjee, G. Bourla, S. Chen, S. Purohit, and J. Battipagli ―Comparative Analysis of Machine Learning Algorithm through Credit Card Fraud Detection,‖ pp. 1–10, 2018.
6. 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.
7. 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.
8. 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.
9. I. Sohony, R. Pratap, and U. Nambiar, ―Ensemble learning for credit card fraud detection,‖ vol. 13, no. 24, pp. 289–294, 2018.
10. 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.
11. 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.
12. H. A. Shukur, ―Credit Card Fraud Detection Using Machine Learning methodologies,‖ vol. 8, no. 3, pp. 257–260, 2019.
13. "Local outlier factor", En.wikipedia.org, 2019. [Online].Available:https://en.wikipedia.org/wiki/Local_outlier_factor.[Accessed: 06- May- 2019].
14. "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

Dr. V. Vijayakumar graduated from Pondicherry University with a Bachelor's and his Master's degree. He did his research work at Pondicherry University in the area of Vehicle Ad-hoc Network on 2019. He currently works in the Department of Computer Science Engineering as an Associate Professor, and has 7 years of experience in teaching.

Nallam Sri Divya is currently pursuing her Under Graduate Degree in Sri Manakula Vinayagar Engineering College in the field of Computer Science and Engineering. She is interested in research works on Machine Learning and ERP.

P. Sarojini is currently pursuing her Under Graduate Degree in Sri Manakula Vinayagar Engineering College in the field of Computer Science and Engineering. She is active in research works on Block-chain and ML.

K. Sonika is currently pursuing her Bachelor of Technology in Sri Manakula Vinayagar Engineering College in the field of Computer Science and Engineering. She is interested in learning AR and interested to do research in Computer Vision.