International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.8, Issue.6, June 2020 E-ISSN: 2347-2693

Comparative Analysis of Machine Learning Algorithms for Credit Card Fraud Detection

Jerin Ignatious[1*], Yogita Kulkarni[2], Shruti Bari[3], Deepali Naglot[4]

[1,2,3,4]Dept. of Computer Science & Engineering, Jawaharlal Nehru Engineering College, Aurangabad, India

**Corresponding Author: jerinignatious1998@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v8i6.69 | Available online at: www.ijcseonline.org

Received: 05/June/2020, Accepted: 20/June/2020, Published: 30/June/2020

Abstract- In this age of growing digitization, many of our day to day activities are being transformed with the help of the internet and everything is now available online. With the advent of the internet, nowadays online transactions have become an important and necessary part of our lives. As the number of transactions are increasing, the number of fraudulent transactions are also increasing rapidly. To reduce fraudulent transactions, machine learning algorithms like Local Outlier Factor and Isolation Forest are discussed in this paper. An online dataset is used to implement and test these algorithms. Finally with comparative analysis we tried to conclude which algorithm works better.

Keywords- Credit Card, Fraud, Machine Learning, Isolation Forest, Local Outlier Factor

I. INTRODUCTION

With the emergence of online transactions people started using credit cards on a frequent basis which has led to high amounts of frauds involving credit cards. This has in turn led people to come up with solutions to prevent these kinds of frauds and machine learning is one such solution among them[1].

Detecting a fraud involves studying and analysing the user's behaviour pattern in order to identify the patterns that the user may not tend to use. In order to differentiate a fraud transaction from a legitimate transaction, we need to understand various technologies, algorithms and types involved in identifying credit card frauds[4].

Algorithms can differentiate between transactions which are fraudulent or not. To find fraud transactions they need a dataset and the knowledge of fraudulent transactions. They analyze the dataset and learn to classify the transaction as fraudulent or non fraudulent[7].

In this paper we discuss the various algorithms which we have used to help us differentiate between fraudulent and legitimate transactions. We employ unsupervised machine learning algorithms like Local Outlier Factor and Isolation Forest to find some visible patterns in the fraudulent and non fraudulent transactions

The paper is organised as follows, Section I contains the introduction of the importance to prevent fraudulent transactions, Section II contains the proposed system architecture which we have employed, Section III contains methodology used in the proposed system, Section IV explains the results and discussion of our system and Section V concludes research work with future scope.

II. RELATED WORK

Laxmi S. V. S. S. and Selvani Deepthi Kavila used machine learning algorithms like Random Forest, Logistic Regression and decision tree to detect frauds and presented a comparison between them. As the dataset was highly imbalanced, they used oversampling to balance the dataset and thereafter applied the above mentioned algorithms and implemented them in R language. By comparing the results produced by all the three algorithms they concluded that Random forest classifier had the highest accuracy of 95.5%. Sensitivity, specificity, accuracy and error rate were used to evaluate the performance of the proposed system[7].

Vaishnavi Nath Dornadula and Geetha S clustered the cardholders into different groups based on their transaction amounts and then used a sliding window strategy to aggregate the transactions made by cardholders from different groups so that behaviour patterns of each group can be extracted respectively. Later they used different classifiers on each group separately They concluded that Logistic Regression, Decision Tree and Random Forest were the better algorithms compared to the others[1].

Suresh K. Shirgave, Cheta J. Awati, Rashmi More and Sonam S. Patil reviewed various machine learning algorithms and examined them based on the metrics such as precision, accuracy and specificity. After the examination they selected the supervised machine learning algorithm, Random Forest to to classify the transaction as fraudulent or authorized. They trained the classifier using feedback and delayed supervised samples which later aggregated each probability to detect frauds[5].

Heta Naik and Prashasti Kanikar tested many machine learning algorithms such as Naive Bayes, Logistic

Regression, J48, K- nearest Neighbour, Random Tree, Outlier and AdaBoost to detect fraudulent transactions. They evaluated the above mentioned algorithms using the metrics such accuracy and time taken. They concluded that logistic regression and adaboost had the highest accuracy and the execution time taken by adaboost was very less. Hence AdaBoost was the better algorithm than the others which were tested[4].

III. PROPOSED SYSTEM ARCHITECTURE

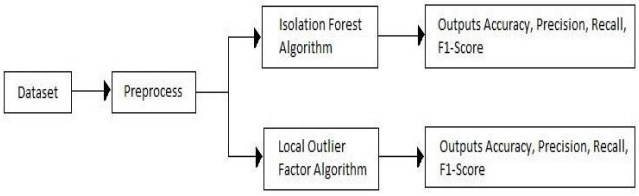

Figure 1: Proposed System Architecture

Dataset:
The dataset that we used to train our models is of the shape (284807, 31) i.e it has 284807 training examples. Each training example consists of 31 features namely Time, Amount, Class and 28 other columns named V1 to V28. The columns named V1 to V28 have been transformed using PCA transformation for security purposes of the user and to keep the identity and personal information of the user secure. The feature named Class denotes whether the training example is legitimate transaction or a fraud transaction. Fraud transactions are marked as 1 and non fraudulent are marked as 0. The dataset has 492 fraudulent transactions and 284315 non fraudulent transactions. In this paper we have analysed the dataset which is taken from Kaggle. 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.

Preprocessing the Data:
The dataset used to train the models mentioned in this paper is already preprocessed to some extent using PCA transformation. The only task left behind is to check for correlation between the features and to remove the Class feature and store it in a separate NumPy array.

We needed to check for correlation between the features to make sure that no two features are directly dependent on each other. This kind of a dependency causes abnormalities in the working of the learning model and to prevent these abnormalities we need to check for correlation between the features of the dataset and remove the features that are correlated.

As we implemented Unsupervised Learning Algorithms, we needed to remove the feature Class from the dataset which contains the label for all the training examples and store it in a NumPy array to use it to check the accuracy of the values predicted by our algorithms.
Implementing the Algorithms:

To detect potential fraud cases we decided to use unsupervised machine learning algorithms, namely Isolation Forest and Local Outlier Factor. After training the dataset on these two algorithms we compare the results to determine which of the two produces the best result. The details of both the algorithms and the dataset is discussed in the next section.

IV. 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 dataset for detecting frauds in credit card systems using python which are briefly explained below and compared their performance. Comparisons are made for these algorithms to determine which algorithms give better results and can be adapted by credit card merchants for identifying frauds [2].

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. 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.

Isolation Forest:
Isolation forest is a tree-base model that is developed to detect outliers. This algorithm is based upon the fact that anomalies are the data points which are few and different. These properties result in a 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 effective. This algorithm has small memory requirements and low linear time complexity.

V. RESULTS AND DISCUSSION

Evaluation Metrics Many classification tasks use simple evaluation metrics such as Accuracy to compare performance between models. 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

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

Precision = TP/ (TP * FP) (1)

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

Recall = TP/ (TP + FP) (2)

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 * (Recall * Precision) / (Recall + Precision) (3)

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

Isolation Forest: 645 (Wrong predictions)

Accuracy: 0.997735308472053

Table 1: Values calculated by Isolation forest

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	284315
1	0.34	0.35	0.35	492

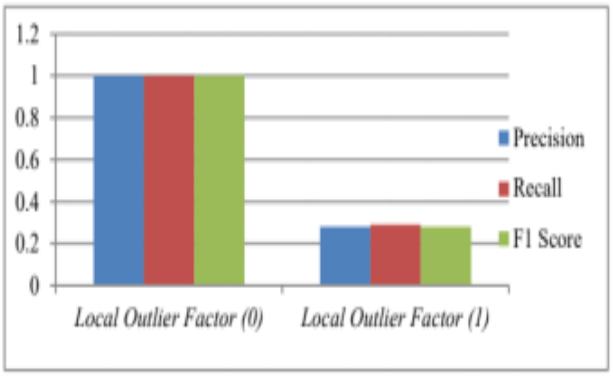

Figure 2: Local Outlier Factor Results

Local Outlier Factor: 935 (Wrong predictions)

Accuracy: 0.9967170750718908

Table 2: Values calculated by local outlier factor

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	284315
1	0.05	0.05	0.05	492

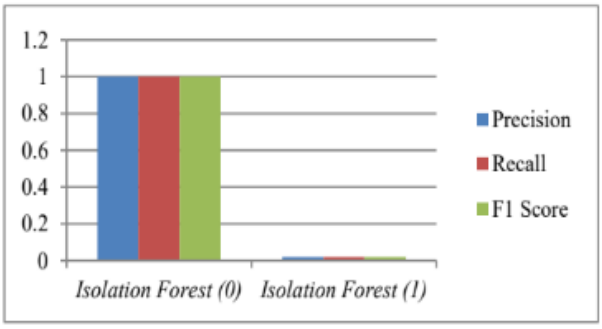

Figure 3: Isolation Forest Results

Experimental Results

By comparing the results of Local Outlier Factor and Isolation Forest algorithm,it is clear that the Isolation Forest is best for detecting the frauds in credit cards.

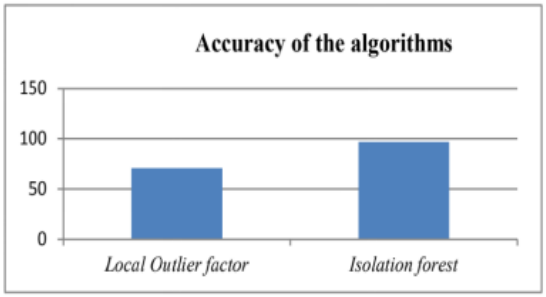

Figure 4: Accuracy Comparison of Algorithms

VI. 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 Isolation Forest gave the highest accuracy rate of 97% followed by the Local Outlier Factor 76%.Accuracy is calculated by F1-score of false positives. In this paper we discussed two unsupervised machine learning algorithms to detect credit card frauds, but for future scope we recommend using more unsupervised algorithms and even suggest the use of neural networks and deep learning to predict the fraudulent transaction accurately without raising any false alarms. The use of deep learning hasn't yet been discussed or sought after when we discuss credit card frauds, so deep learning may be the next big thing after machine learning that can help us understand the patterns of a user in a better way and help us detect frauds with greater efficiency. Hence in the future scope we recommend the use of Deep Learning.

REFERENCES

[1] Vaishnavi Nath Dornadula, Geetha S, "Credit Card Fraud Detection Using Machine Learning Algorithms", International Conference on Recent Trends in Advanced Computing 2019

[2] Samaneh Sorournejad, Zahra Zojaji, Reza Ebrahimi Atani, Amir Hassan Monadjemi, "A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective"
[3] Ramyashree K, Janaki K, Keerthana S, B. V. Harshita, Harshita Y. V, "A Hybrid for Credit Card Fraud Detection Using Machine Learning Algorithm", International Journal of Recent Technology And Engineering, Volume-7, Issue-6S4, April 2019
[4] Heta Naik, Prashasti Kanikar, "Credit Card Fraud Detection Based on MAchine Learning Algorithms", International Journal of Computer Applications, Volume 182 - No. 44, March 2019
[5] Suresh K. Shirgave, Chetan J. Awati, Rashmi More, Sonam S. Patil, "A Review on Credit Card Fraud Detection Using Machine Learning", International Journal of Scientific and Technology Research, Volume 8, Issue 10, October 2019
[6] S. P. Maniraj, Aditya Saini, Swarna Deep Sarkar, Shadab Ahmed, "Credit Card Fraud Detection Using Machine Learning and Data Science", International Journal of Engineering Research and Technology, Volume 8, Issue 9, September 2019
[7] Laxmi S. V. S.S, Selvani Deepthi Kavila, "Machine Learning for Credit Card Fraud Detection System", International Journal of Applied Engineering Research, Volume 13, Number 24(2018)

Author's Profile

Ms.Deepali Naglot is an Assistant Professor at MGM's Jawaharlal Nehru Engineering College, Aurangabad, India. She pursed Bachelor of Engineering in Computer Engineering from Savitribai Phule Pune University, Pune in year 2014 and Master of Technology in Computer Science and Engineering from Savitribai Phule Pune University, Pune in year 2016. Her Area of Interest is Data Science & Machine Learning.

Mr.Jerin Ignatious is currently pursuing his bachelor in Computer Science & Engineering from MGM's Jawaharlal Nehru Engineering College, Aurangabad, India. His research area of interest includes Data Science and competitive Programming

Ms. Shruti Bari is currently pursuing her bachelor in Computer Science & Engineering from MGM's Jawaharlal Nehru Engineering College, Aurangabad, India. Her research area of interest includes Machine learning and Cloud Computing

Ms.Yogita Kulkarni is currently pursuing her bachelor in Computer Science & Engineering from MGM's Jawaharlal Nehru Engineering College, Aurangabad, India. Her research area of interest includes Machine Learning.

