International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V69I8P101

Volume 69 Issue 8, 1-3, August 2021

Original Article

Credit Card Fraud Detection using Unsupervised Machine Learning Algorithms

Hariteja Bodepudi

Irving, USA.

Received Date: 01 July 2021
Revised Date: 02 August 2021
Accepted Date: 13 August 2021

Abstract - *In the modern era, the usage of the internet has increased a lot these days and becomes an essential part of the life. As the e-commerce has increased, the buying and selling the products over the internet becomes more easy and flexible. The usage of online shopping, online bill payment has increased a lot these days with the introduction of the modern technology like online banking, credit card payments.*

Due to the increase of the online payment and online shopping, the risk of credit card usage also increased as the credit card was used in many places as it becomes hard for the bank to distinguish the real transactions of the consumer versus the fraud transactions. Also, the credit card fraud transaction can also happen if the customer accidentally loses the credit card. So, it becomes complex for the banks to stop the fraud transactions at that point.

This paper talks about how to detect the fraud transactions and block the payments before processing by using the Machine Learning on a real-time basis. This paper clearly explains about how the anomalies can be detected in an unsupervised approach and to achieve the higher accuracy. This Anomaly detection process used in this paper can be applied in a wide range of applications like fraud detections in banking, underbilling/overbilling for customers in telecommunications, security monitoring, network traffic, health care, and a wide range of manufacturing industries.

Keywords *- Anomaly, Machine Learning, Supervised Learning, Unsupervised Learning.*

I. INTRODUCTION

Ecommerce is playing a crucial role in the day-to-day activities of humans. The Usage of the internet and IOT devices has increased a lot these days, and that results in massive volumes of data . [1] As the data increases due to the collection of data from different forms like structured and semi-structured data, which is commonly referred to as Big Data . [2]This large volume of data becomes more essential for the organizations for the data analysis, especially for the banks to track the transactions on a real-time basis. Security of credit cards has become a big threat in the modern world due to the wide range of online usage and wide growing data transactions to be analyzed to detect fraud. It becomes easy for the users to do shopping, paying bills online in the busy life rather than the physical visits.

As the usage of credit cards across the globe and online has increased and data have grown drastically so, it becomes a challenge for the bank to manually detect the fraudulent transactions and block the usage. This paper clearly explains the usage of Artificial Intelligence Machine Learning Unsupervised Algorithms to detect the Anomalies.

II. ANOMALY DETECTION

Anomaly detection, which is also referred to as Outlier Detection, helps us to identify the events, data points that are far different from the other normal events. These Anomalies are the data points that deviate from the normal data points behavior. [3]

Anomalies can be detected in many ways. In this paper, I'm focusing on detecting the Anomalies, i.e., Fraudulent Transactions of the Bank, using the Unsupervised Machine Learning Algorithms.

III. MACHINE LEARNING

It belongs to the branch of Artificial Intelligence . The concept of Machine Learning algorithms is to learn from the data, identify the behaviors and predict the future with the minimal intervention of the mankind. [4]

There are two popular methods for the Machine Learning widely used across the globe

They are :

- Supervised Learning Algorithms
- Unsupervised Learning Algorithms

A. Supervised Learning Algorithms

Supervised Learning Algorithms uses the approach of the Learning the data patterns using the labeled data, i.e., using the output. These algorithms are trained with the training data set, which has both inputs and outputs. The algorithm is used to learn how the input features are related to the target variable, i.e., output, in this approach . As the output is fed into this algorithm, it will adjust the weights until the model has fitted perfectly . Once the model is trained using the train data set and that model is used to predict the dataset, which is unseen before .

B. Unsupervised Learning Algorithms

Unsupervised Algorithms are the algorithms that are used for unlabeled data set, i.e., which doesn't have the output variable . These algorithms are used to find the unknown patterns in the data and are used to analyze and segment them into clusters based on the behaviors. Similar behavioral patterns of the data are clustered into similar groups . These algorithms are widely useful for the unlabeled dataset . [5]

IV. IMPLEMENTATION

In this paper, I'm using the Kaggle credit card data set to find the fraudulent transactions using the unsupervised algorithms . [6] So, the credit card data set is not trained with the output variable, i.e., the target variable . It is directly trained on the actual dataset without any labels . After training on the dataset and the next step is to predict on the same dataset to find the fraudulent transactions in the whole dataset .

To predict the anomalies in the dataset, I'm used 3 unsupervised algorithms to test the accuracy and see the best performing algorithm . The reason for using the unsupervised approach in this paper is mostly in the real world; there won't be any labeled data available, and also for the fraud detection, the unsupervised approach is the best approach .

The algorithms I'm using in this paper are:

- Isolation Forest
- Local Outlier Factor
- One Class SVM

A. Isolation Forest

Isolation Forest belongs to the unsupervised Algorithm . Isolation Forest algorithm works in the same methodology as the Random Forest and is constructed on the concept of decision trees. The main idea and methodology of the isolation forest algorithm are to detect the anomalies, i.e., fraud transactions, instead of using the normal characteristics of data points. It randomly selects the features and then randomly splits the feature between min and max values. [7]

The important parameter in the Isolation Forest Algorithm is the contamination parameter which we can add the outlier value like 0.1 if it is 10%, and if we are not sure about the value can default it to auto for the algorithm to take care of the outliers while fitting .

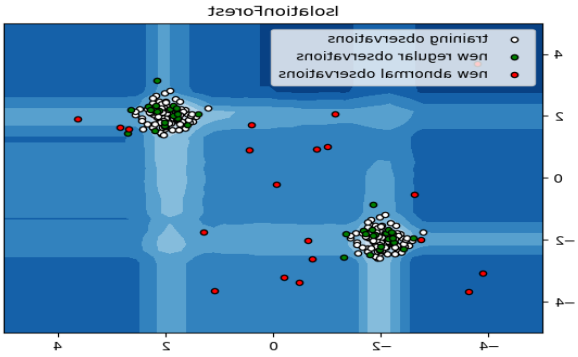

Fig. 1 Isolation forest [8]

From the figure1, it is clear how the outliers are separated from the normal data points in the isolation factor . Red data points that are far from the other data points are considered as the outliers

B. Local Outlier Factor

Local Outlier Factor (LOF) is an unsupervised algorithm used in the identification of the anomalies or fraud transactions in the data . The methodology of the Local outlier Factor Algorithm measures the density score of each sample and weighs the scores. It compares the scores with the other data points, and the lowest score data points are the anomalies . [9]

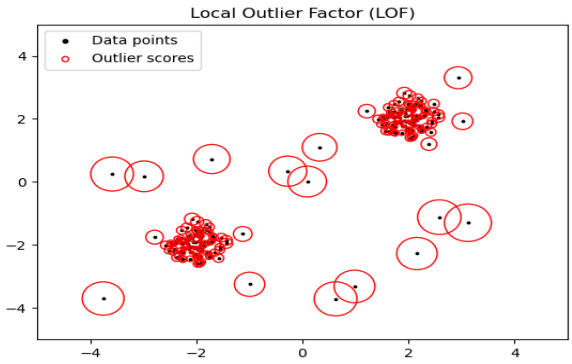

Fig. 2 Local Outlier Factor (LOF) [10]

From the figure2 we can see the data points and the outlier scores, and the lower density from the neighbor points are considered as outliers.

C. One Class SVM

One Class Support vector machine is the unsupervised algorithm used to detect the outliers in the given dataset . The idea of the One class SVM is to classify the whole data into binary classes, i.e., inliers and outliers . It takes the density of the higher class and the extremity points, i.e., lower density is considered as outliers . [11]

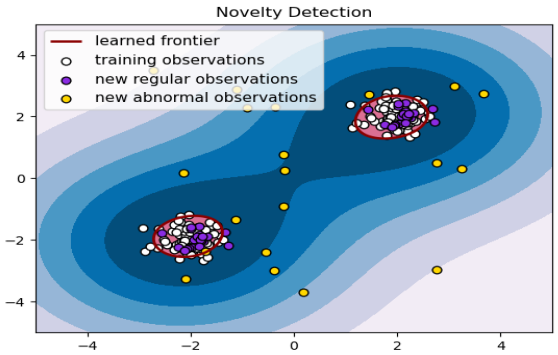

Fig. 3 One-Class support vector machine [12]

It is clear from the figure 3 how the One class SVM separates the outliers . The yellow-colored data points in the figure are considered as anomalies .

I trained and predicted the Kaggle credit card data using the above 3 unsupervised algorithms to identify the anomalies, i.e., to separate the fraud transactions from the normal transactions.

V. RESULTS

After predicting with unsupervised algorithms Isolation Forest, Local Outlier Factor, and One class SVM. The Isolation Forest outperforms than other 2 algorithms

A. Accuracy of Model

Algorithm	Accuracy
Isolation Forest	99.74%
Local Outlier Factor	99.65%
One-Class SVM	70.09%

Fig. 4 Results Table from the Algorithms

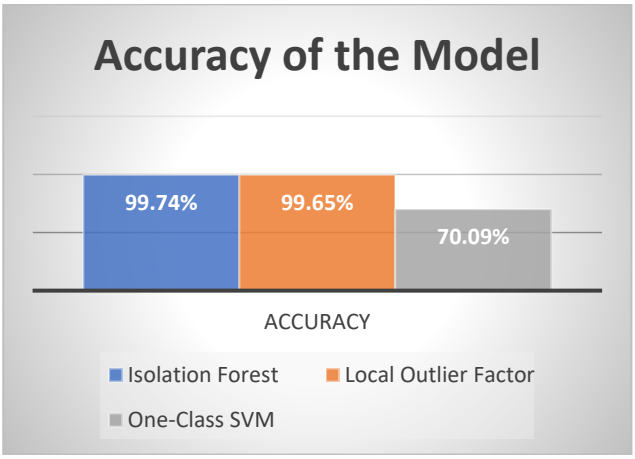

Fig. 5 Results plot from the algorithms

From the above figure4 and figure5 Results table and plot , we can see the accuracy of Isolation Forest is 99.74%, Local Outlier Factor is 99.65%, and One-Class SVM is 70.09%.

So, when compared with all the 3 models, results of Isolation Forest and Local Outlier Factor Algorithms are almost performing similarly and better than the One class SVM, which is poorly performing .

VI. CONCLUSION

This paper talks about how the credit card fraudulent transactions can be detected with the unsupervised algorithms. In the real world, we won't be having the outputs for the inputs, so this paper concentrates on unsupervised learning so that way these algorithms can detect the anomalies without even observing the output variables .

Unsupervised algorithms are very useful and good detectors of the anomalies . The algorithms used in this paper can be applied to any field for anomaly detection, i.e., to detect the outliers or fraudulent transactions .

REFERENCES

[1] H. Bodepudi., Faster The Slow Running RDBMS Queries With Spark Framework, (2020). [Online]. Available: https://www.researchgate.net/publication/347390599_Faster_The_Slow_Running_RDBMS_Queries_With_Spark_Framework. [Accessed 07 2021].

[2] H. Bodepudi., Data Transfer Between RDBMS and HDFS By Using The Spark Framework In Sqoop For Better Performance, Internation Journal of Computer Trends and Technology, 69(3) (2021) 1.

[3] Anaomaly-Detection., [Online]. Available: https://avinetworks.com/glossary/anomaly-detection/. [Accessed July 2021]., (2021).

[4] What is Machine Learning., [Online]. Available: https://www.ibm.com/cloud/learn/machine-learning. [Accessed July 2021].

[5] Popular Machine Learning Methods, [Online]. Available: https://www.sas.com/enus/insights/analytics/machine-learning.html. [Accessed July 2021].

[6] CreditCard Data, [Online].Available:https://www.kaggle.com/mlg-ulb/creditcardfraud?select=creditcard.csv. [Accessed July 2021].

[7] H2o.ai Data Science, ,[Online]. Available: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/if.html.

[8] Isolation Forest Example, [Online]. Available: https://scikit-learn.org/stable/auto_examples/ensemble/plot_isolation_forest.html#sphx-glr-auto-examples-ensemble-plot-isolation-forest-py.

[9] Anaomaly Detection With Local Outlier Factor, [Online]. Available: https://www.datatechnotes.com/2020/04/anomaly-detection-with-local-outlier-factor-in-python.html.

[10] Outlier detection with Local Outlier Factor (LOF), [Online]. Available:https://scikitlearn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html. [Accessed August 2021].

[11] One-Class Support Vector Machine., [Online]. Available: https://www.xlstat.com/en/solutions/features/1-class-support-vector-machine.

[12] One-Class SVM.M [Online]. Available: https://scikit-learn.org/stable/auto_examples/svm/plot_oneclass.html#sphx-glr-auto-examples-svm-plot-oneclass-py.