Credit Card Fraud Detection System Using Density Based and Classification Based Algorithms

Sanjana Jagdish[1], Mayank Singh[2], Vikash Yadav[3]

Department of Computer Science and Engineering, ABES Engineering College, Ghaziabad, India

sanjana.16bcs1014@abes.ac.in[1], mayank.16bcs1102@abes.ac.in[2], vikash.yadav@abes.ac.in[3]

Abstract- In today's world, billions of transactions are done via credit card. From last decade, there has been an ease in online payment, and due to its promptness, it has opened a new way for different opportunities. These opportunities are not only in the form of businesses, lowering distance differences and for fast transaction, but also it led to a new increasing problem called fraudulent transaction.

While these factors increased efficiency and led the e-commerce to its new height per year, but also it opened a gateway for fraudsters to highly misuse the transparency of online transactions as well as credit card transactions. In this project, we aim to improve the method, and the state of transactions. Other than scalability and efficiency, fraud detection system also has other technical errors like, training data, collection of data; both of them led to great causes of error in the system.

Via this paper, we aim to improve the method, and the state of transactions. Other than scalability and efficiency, fraud detection system also have other technical errors like, training data, collection of data; both of them led to great causes of error in the system. In this paper, we aim to discuss the algorithms which are more efficient in the fraud detection system. The proposed model works upon different algorithms and regression techniques; those are Local Outlier Factor, Isolation Forest Algorithm, Naïve Bayes Classifier, and Logistic Regression. This system has helped lot of banks and financial firms to pre-empt the fraud transactions, and predict the fraud that may occur, via analysis.

Keywords- Isolation Forest, Local Outlier Factor, Naïve Bayes, Logistic Regression

I. INTRODUCTION

In this century, credit cards are used to perform almost every transaction that we process, and these transactions may be of lesser or huge amounts. And due to its vast use, credit card fraud is becoming more uncontrolled. According to some banks in 2019, the total fraud level increased from $130 million USD to $843 million USD in single area of payment from previous years. Moreover, payments have change to new technology such as wire transfer, credit card,

gifts and reload cards, UPI transactions etc. which therefore lead in increased level of fraud. Due to which we have to introduce new detection system which is better than previous one.

Addition to this, the fraudsters keep changing their techniques to get through the firewall, thus the previous fraud detection techniques go waste. Also, if old models don't change or don't adapt new methodologies, the machine learning algorithms fail to work. So due to this we have to introduce new detection system which is better than the previous one [1].

During the development of Credit Card Fraud Detection System, there are lots of factors that highly affect the precision of system such as skewness of the data, redundancy, and short-time replication of the system cost-sensitivity of the system and how to progress the feature.

For detecting fraud transaction using credit cards we design a system. The system we design provides most of the needed features which required finding transaction are valid one or fraud one. As time changes, technology also changes and make it difficult to find patterns and behavior of fraud transaction. For averting these unknown frauds is not an easy task.

Making a system that is more secure, and to avoid fraud transaction we can use certificates for both side of merchants and customers which enhances the system and updating and maintaining it as per the new technological developments.

II. RELATED WORK

Our aim is to identify the frauds in the credit card transaction, to achieve that; we need to classify the transactions, which increase the efficiency. Thus, we review the various credit card fraud detection publications.

Bolton and Hand (2001) divide credit card frauds as: Behavioural frauds and Application frauds. Sometime unauthorized person gets new card obtain by falsification called application fraud. Cards having lost or get mail theft these are part of behavioural fraud. So, increment of time gets more problems due to advancement of technology. Both have tried to challenge this issue by developing system using Supervised Machine Learning Algorithm [2].

Sahin and Duman (2011) study using algorithms such as (LR) logistic regression and (ANN) Artificial Neural Network to scale transactions as per their flagship in legal or fraud transactions. They found that ANN is better than LR based on results. However, as the data get larger the performance gets decreased. In the studies of related, the false negative cost (such as legal transaction as fraud) and false positive cost (such as fraudulent transaction as legal) are taken equally. However, in the domain costs of false negative is greater than costs of false positive and in every transaction it changes. To maintain that researcher, use adjustment cost matrix in training [4].

So, there are still need some places where it lacks due to time efficiency, large data processing, etc. for working with rapidly developing technology and new mode of transaction it needed to

reorganised and reform the system [3]. For this, we introduced this study in which classification is used for better development of system which guarantees to detect fraudulent and valid transactions done in a given set of data, with highest accuracy and also is able to identify the group of fraudsters who do it, by studying their characteristics, according to their gender, and age group.

A vast survey has been done in this area [8]. In which author's have discussed Fisher Discriminant Analysis and Modified Fisher Discriminant Analysis. In which it was depicted that Normal Fisher Discriminant Analysis (FDA) shows results generated with little less sensitivity input values were not that accurate. But, Modified version of it resolved that issue, and thus yielded the most accurate value amongst all the results generated by all the algorithms used till now.

III. METHODOLOGY

Machine learning algorithms help us in manipulating the date on-premise. And gives us the results with the highest accuracy. Various iterations of these algorithms help us in getting more refined accuracy of the desired. The algorithms used in this test case are:

* Isolation Forest Algorithm
* Local Outlier
* Naïve Bayes Classifier
* Logistic Regression

3.1 Isolation Forest Algorithm

Isolation forest is classified as unsupervised learning algorithm which is used for anomaly detection.

Isolation forest algorithm isolates the transaction having the chances of detecting anomaly in them. These transactions which is isolated having checked with different parameters for fraud or valid transaction [7].

3.2 Local Outlier Factor

Local Outlier Factor is proposed by Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng and Jörg Sander in 2000. To find anomaly points, it is used to measure the local deviation of a data point with their neighbors.

The local outlier factor is based on local density, where k-NN (k nearest neighbors) gives locality, whose distance used for calculating density. Comparison of local density between an object and its neighbors, we can identify the region of similar density, and points having lower density then their neighbors. These are called as outliers [6].

3.3 Naïve Bayes Classifier

Naïve Bayes classifier is the part of classification algorithms, which is based on Bayes Theorem. It is the collection of algorithms not a single one where everyone shares similar principle, i.e. every single one have their own features [5].

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

Where A and B are events

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

3.4 Logistic Regression

Logistic regression is a type of classification algorithm, where certain type of class must exist such as win/lose, pass/fail, alive/dead or healthy/sick. These models can be further classified or extended to different classes of event such as to find out whether an image contain dog, cat, elephant, etc. And every object having point between 0-1 and sum is one.

IV. EXPERIMENTAL RESULTS

The sample of dataset is taken from daily life transactions where there are both type of transactions occurs such as: Valid transaction and Fraud transaction. Precisely, dataset contain 284807 transactions with different attributes, where there are 284392 are valid transaction and 415 are fraud transaction.

We have tons of data; thus, computation will be slow and hence we take 10% of total data for ease of computation. For determining the cases of fraud transaction, we use classification and made 'class = 1' for 'fraud' and 'class = 0' for 'Valid' transactions.

When we start to first use machine learning algorithms there are two types of algorithms, we use one is Isolation forest and Local Outlier which is density-based algorithms which also used for anomaly detection.

Then firstly we use Isolation Forest and Local Outlier Factor for data processing and run the classification matrices for it to get precision scores.

ISOLATION FOREST				
Column1	precision score	recall score	f1-score	support
class 0	1.00	1.00	1.00	28432
class 1	0.28	0.29	0.28	49
avg/total	1.00	1.00	1.00	28481

In Isolation Forest algorithm it isolated transaction which have high chances of anomaly detection and scores them. This step iterated for every transaction to find out transactions is valid or fraud and get scores in matrix or graphical form.

LOCAL OUTLIER FACTOR				
Column1	precision score	recall score	f1-score	support
class 0	1.00	1.00	1.00	28432
class 1	0.02	0.02	0.02	49
avg/total	1.00	1.00	1.00	28481

The Local Outlier Factor defines different parameters to identify criteria for fraud transactions. This process gets iterated on every transaction, which gives each transaction their own scores between 0 and 1.

After getting these scores we compute the results of both the algorithms:

ALGORITHM	precision score	recall score
Isolation Forest (	1.00	1.00
Isolation Forest (	0.28	0.29
Local Outlier (0)	1.00	1.00
Local Outlier (1)	0.02	0.02

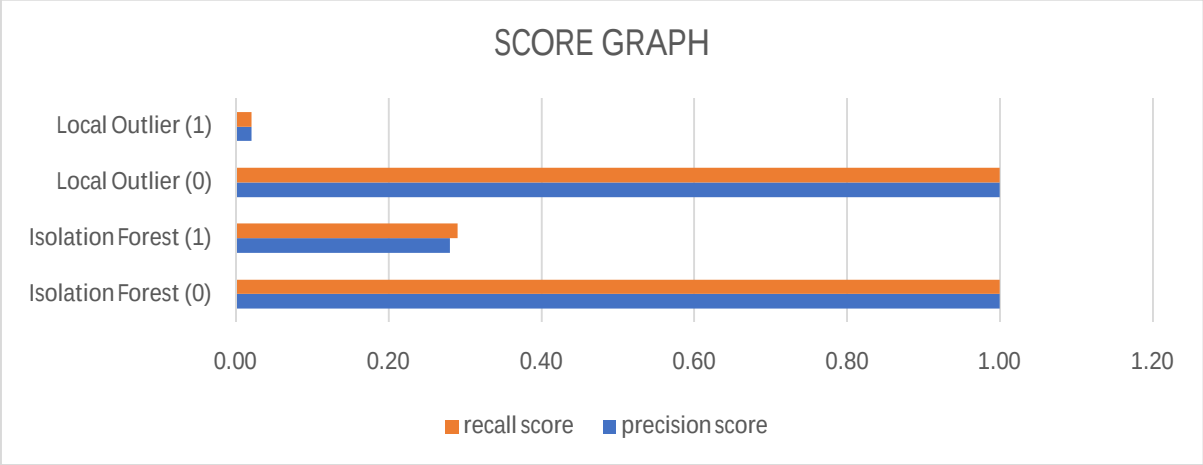

Fig.1: Score Graph for comparing Isolation forest and Local Outlier Factor

Naïve Bayes and Logistic Regression is a part of classification algorithm, so we use Naïve Bayes Classifier and Logistic Regression to get the precision scores from given dataset to print the classifier's scores.

NAÏVE BAYES CLASSIFIER				
Column1	NB 1	NB 2	NB 3	NB 4
precision score	0.05878187	0.084729064	0.084812623	0.086345382
recall score	0.846938776	0.87755102	0.87755102	0.87755102
f1-score	0.109933775	0.154537287	0.154676259	0.157221207
accuracy score	0.976405323	0.983480215	0.98349777	0.983813771
ROC AUC	0.963247972	0.96220341	0.961361264	0.961155618

LOGISTIC REGRESSION				
Column1	LR 1	LR 2	LR 3	LR 4
precision score	0.808823529	0.930693069	0.062526584	0.06137931
recall score	0.56122449	0.959183673	0.896341463	0.908163265
f1-score	0.662650602	0.944723618		
accuracy score	0.999016888	0.944162437		
ROC AUC	0.974786243	0.978045764		

By computing the classifier's scores, we get:

ALGORITHM	precision score	recall score
Naïve Bayes (NB4	0.086345382	0.87755102
Logistic Regressic	0.06137931	0.908163265

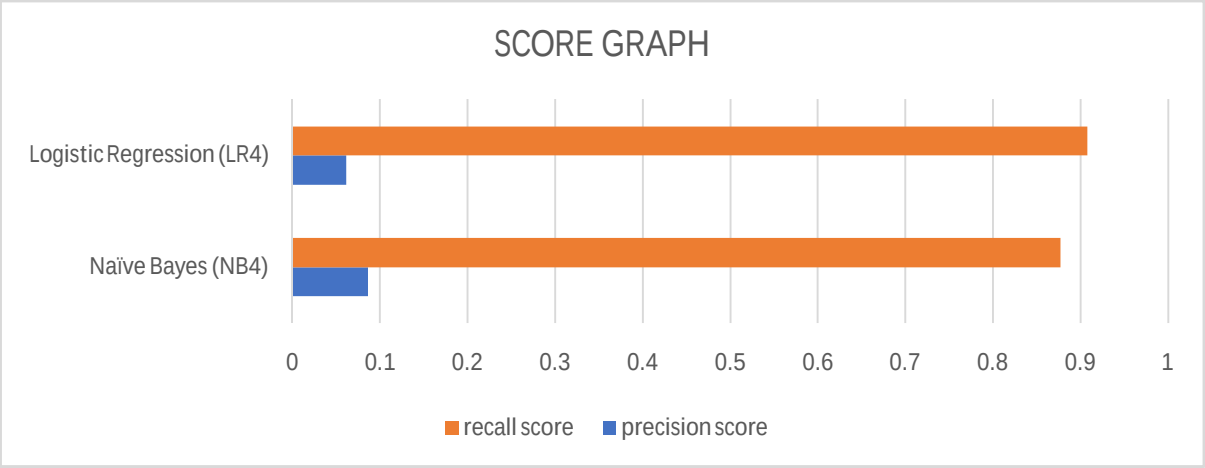

Fig.2: Score Graph for comparison of LR and Naïve Bayes Classification

The output is the occurrences of frauds that happened during the record of our data set. This system is valuable to bank systems, or financial firms to study the abnormalities in the transactions, and improve their system. Consultancy firms and individuals can also use this system to design more profitable system for business.

V. CONCLUSION

As rapidly developing technology, the mode of online payment and credit card uses increase exponentially. Now a day's credit card is normally used in any transaction either online or regular purchase increased. As a result, cards issued by banks or any type of online transaction where credit card are used needed efficient fraud detection system.

In the case of finding the fraudulent and valid transactions with the help of machine learning algorithms, we hence conclude that, In case on density based algorithms, Isolation Forest is better than Local Outlier Factor as its precision score is much better than the later in precision and in case of classifier algorithms, Naïve Bayes detected more 'True Positive' and less 'False Negative'. So, Naïve Bayes Classifier gives the best results for both cases, False Negatives and True Positives.

References

[1] Bahnsen, Alejandro Correa, Djamila Aouada, Aleksandar Stojanovic, and Björn Ottersten. "Feature engineering strategies for credit card fraud detection." *Expert Systems with Applications* 51 (2016): 134-142. [Accessed 11 Oct. 2019]

[2]https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjqx5PL3NDpAhUN6XMBHeVSCkQQFjABegQIAxAB&url=https%3A%2F%2Fprojecteuclid.org%2Fdownload%2Fpdf_1%2Feuclid.ss%2F1042727940&usg=AOvVaw31VJee_vjRdqiM8ccIW34F.

[3] https://www.aarp.org/money/scams-fraud/info-2020/ftc-fraud-complaints-rise.html.

[4] https://en.wikipedia.org/wiki/Logistic_regression

[5] https://www.geeksforgeeks.org/naive-bayes-classifiers/

[6] https://en.wikipedia.org/wiki/Local_outlier_factor

[7] https://en.wikipedia.org/wiki/Isolation_forest.

[8] Sanjana Jagdish, Vikash Yadav et al, "Credit Card Fraud Detection: A Survey", Journal of Xidian University, Science Press Publication, ISSN: 1001-2400, Vol. 14, No. 5, pp. 5528-5534, May 2020, DOI:10.37896/jxu14.5/599.