Credit Card Fraud Detection using Machine Learning Algorithms

Mudita Sharma
HMR Institute of Technology & Management
New Delhi, India
mudita10251@gmail.com

Harshita Sharma
HMR Institute of Technology & Management
New Delhi, India
harshi.28sharma@gmail.com

Pawan Bhutani
Assistant Professor
HMR Institute of Technology & Management
New Delhi, India
pvbhutani@gmail.com

Ira Sharma
HMR Institute of Technology & Management
New Delhi, India
irasharma408@gmail.com

ABSTRACT:

Fraud: fraud is any activity that causes financial loss to any other person. With the advancement of technology in the modern era, fraud related to the credit card is continuously increasing. The credit card frauds are costing a huge amount of dollars from the consumer and customers as well. Now a day's scammers are continuously finding new ways to do the scams. Also, to minimize the credit card fraud losses, different types of systems have been developed which detect the frauds. We have to make a highly sensitive and powerful fraud detection system up to such an extent that it detects the fraud even before it takes place and in a sequential and exact manner.

In this paper, we have mentioned three important algorithms named Isolation Forest, Support Vector Machine and Local Outlier Factor for credit card fraud detection. Also, we have discussed methodologies used in all three algorithms as mentioned above. We have also discussed the applications of the fraud detection system and how it will be implemented in the future.

Keywords: Fraud, Local Outlier factor, Support Vector Machine, Fraud detection.

1. INTRODUCTION

Credit card fraud: When someone uses one's credit card or credit card information for making transactions without his permission or which he doesn't authorize. The Fraudster (one who does the fraud) can steal one's credit card number, PIN number, etc. Now, credit card can

be defined as a card which helps a customer by giving a pre-set credit limit which he can use for making the payment. There are many benefits of having a credit card. Some are listed:

- Chance to gain credit points.
- Earn rewards such as cashback and miles point.
- Protection against **credit card fraud.**
- Free information about credit score.
- No foreign transaction fees.
- Increased purchasing power.
- Not linked to checking and savings account.

Because of advancements in electronic technology, the use of credit card transactions has been increased. There are basically two types of transaction:

(i) **Virtual Transaction:** In virtual transaction only some important information is necessary to make the payment.

(ii) **Physical Transaction**: In the physical transaction, the cardholder physically presents the card for making payment and the fraud is discovered after a few days the transaction has been made.

Various Credit Card Frauds are:

Application frauds: By entering the details or information of the user such as password, pin number, etc. scammers gain control over the application system of the user.

Electronic credit card Imprints: The scammers can steal the information that is present on the magnetic stripe of the card. The information present on the magnetic stripe of the card is confidential, so one should handle this cautiously.

CARD NOT PRESENT (CNP): In some cases, the card can also be used without card number, this can be possible when the scammer has the information about the expiry date and account number.

Card ID theft: In this type of fraud the scammer gets all the details about the credit card and uses that information to open another new (fake) account.

Credit card fraud occurs when one person uses credit card information of the other person. We have to minimize the presently available limit on the credit card as a solution, the present problem statement of fraud-related credit card transactions. The best way to reduce the frequency of frauds is to analyze the present or normal purchase data of cardholders. Every person follows a particular pattern in the transactions and when there is a deviation in that pattern it indicates fraud. **There are three basic types of algorithms i.e. (Isolation Forest, Local Outlier Factor and Support Vector Machine) which are important ways to solve the frauds.** There are various advantages that came into light after practicing these algorithms in the detection of frauds caused by credit card transactions. These are, after using these algorithms; first, there is a drastic change that can be seen very easily in the frauds. The second advantage is, based on the previous pattern followed by the user, fraud can be easily detected. Next is that we can add more updates in our system, i.e. the fraud detection system will block the user after certain attempts which are formerly set by the system organization.

2. DATASET

We have taken the dataset from Kaggle. It contains information of transactions made by European Cardholders in September 2013. The total number of records is 2, 84,807 out of which there are 492 fraud transactions. There are total of 31 features out of which we know only three features i.e. Time, Amount and Class. Rest twenty-eight features are not disclosed due to the confidentiality issue and are named as V1, V2…V28. Time represents the seconds elapsed between the first transaction and the consequent transaction. Amount refers to the amount that has been transacted. Class is the final prediction. Class 0 represents the Normal transaction and Class 1 represents Fraud Transactions.

3. RELATED WORK

Electronic commerce technology has reached new heights in the past few years and so has the use of credit cards, net banking, debit cards and other mobile payment platform has increased at a fast pace. Credit Card being the most versatile mode of payment for both online and regular purchases is somewhat insecure too if we do not handle it with care. Hence, the cases of credit card frauds are also surging. To resolve this issue we have various techniques; Isolation Forest, Support Vector Machine, Random Forest, Logistic Regression, etc. The paper by Massimiliano Zanin and his colleagues [1], has concluded in the result that the score obtained by a standard ANN classification model can be improved by the features that are extracted from network based representation of data and plays a vital role in rectifying the score. To improve data mining models, this paper tested a hypothesis that complex networks can be used for the same. These data mining models are specifically used for detecting fraud

transactions in credit card transactions. This hypothesis provides a new approach to the combined usage of complex networks and data mining tools for detecting fraud cases.

In the paper [2], the Authors- Yashvi Jain, Namrata Tiwari, Shripriya Dubey, and Sarika Jain had done a comparative analysis on various fraud detection techniques such as Support Vector Machine, Artificial Neural Network, Bayesian network, K-nearest neighbor; Fuzzy Logic based system, Decision tree and logistic regression. They concluded that we have major gaps in the techniques and methods of fraud detection. Artificial Neural Network (ANN) and Naive Bayesian Network have high accuracy and better detection rates, but they are costly to train because they need GPUs for training purpose, otherwise they'll work very slowly and can take a lot of time to train. Some algorithms like KNN and SVM give GIVES tremendous results with small data sets but large datasets cannot be trained properly using them. Decision Tree and SVM being a supervised machine learning algorithms works well on sampled and pre-processed data. While raw data can be used efficiently with Logistic Regression and Fuzzy Systems. Authors have suggested using the hybrid of various techniques. According to them, to develop a good hybrid model there is a need to pair an expensive model with an optimization technique that will take care of the cost of the model. The application and environment of the fraud detection system play an important role in choosing the right algorithm. Some Examples of hybrid techniques that have been mentioned in the paper are: a hybrid of Decision Tree and Neural Network; Combination of Neural Network and Genetic Algorithm; hybrid of Fuzzy Clustering and Neural Network; hybrid Bayesian Network and Artificial Neural Network and the combination of Support Vector Machines and Decision Trees.

The Authors- Ishu Trivedi, Monika, MrigyaMridushi, of the paper [3], found that to lower the number of false alerts and finding out fraudulent transactions, the method of Genetic Algorithm gives more accurate results. This algorithm requires a very short span of time after the transactions have been made, to detect whether the transaction is fraud or not, which will ultimately prevent the banks and customers from huge losses and also will reduce risks.

The paper [4] by Linda Delamaire (UK), Hussein Abdou (UK), and John Pointon (UK) gives information about various types of frauds such as theft fraud, bankruptcy fraud, application fraud, counterfeit fraud, theft fraud and behavioral fraud. In today's world, it is important to know about various frauds in order to be careful while making any transaction. The methods that they have used for detecting fraud transactions are Genetic Algorithms, Clustering Techniques, Pair-Wise Matching, Decision Trees, and Neural Networks.

The Authors- SamanehSorournejad, Zahra Zojaji, Reza EbrahimiAtani, and Amir Hassan Monadjemihave stated some difficulties of Credit Card Detection Techniques in the paper [5]. These are Imbalanced data, Different misclassification importance, Overlapping data, Lack of adaptability, Fraud detection cost, Lack of standard metrics. They have also given the complete classification of Credit Card Fraud Detection Techniques. Their paper has shown that Fraud Detection Techniques are of two types: 1. Supervised/user's behavior/ misuses, 2. Unsupervised/Transaction Analysis/ Anomaly Detection. In which Supervised Techniques are Inductive logic programming, Artificial Immune System, Expert System, Artificial Neural Network (BP), and the Case Based Reasoning, Support Vector Machines, Bayesian Network, Rule Base, and Fuzzy System. And Unsupervised Techniques are Hidden Markov Model, Support Vector Machines, Artificial Immune System, Artificial Neural Network (Self-Organizing Map), the Fuzzy System, and Bayesian Network. They have also described the complete classification of the dataset's attribute. Credit card transaction datasets are usually divided into two types i.e. Numerical and Categorical Attributes. The numerical attributes may consider features such as time, amount, etc. while the Categorical attributes can be class which is the final output. They identified Decision Tree and Rule Induction- algorithms as Categorical Based

whereas SVM, KNN, AIS (NSA), GA, ANN- algorithms are Numerical Based. Naive Bayes, CBR, AIS (DCA), Fuzzy System- are some algorithms that come under both types. The paper has also stated about the open issues like Nonexistence of extensive credit card benchmark dataset, Nonexistence of standard algorithm, Nonexistence of suitable metrics, and the deficiency of adaptive credit card fraud detection systems.

The Authors- Navanshu Khare and Saad Yunus Sait have shown in their research paper [6] that Random Forest will give the highest accuracy whereas Logistic Regression, SVM, Decision Tree gives less accuracy than Random Forest. Though Random Forest gives the highest accuracy, but it doesn't hold well in case of speed during testing. In the case of SVM, Datasets requires pre-processing in order to get better results.

In the Research Paper [7], Authors- Masoumeh Zareapoor, Pourya Shamsolmoali compared various standard classifiers with Bagging Ensemble Classifier, which is an interestingly new technique, based on Decision Tree. They also identified that it works very well in the detection of fraudulent transactions. Bagging Ensemble has the highest fraud catching rate and a relatively low false alarm rate. They concluded that the Bagging Ensemble Classifier has performed stably during training, testing, and evaluation. Also, it is independent of the rate of fraud. It also has the capability to handle imbalanced datasets.

The Authors of the Research paper [8] has identified that the Isolation Technique gets high AUC score when compared to K- Means, OCSVM, and LOF. It is more accurate, performs much better than others in detecting fraud transactions, and detects few errors than the other three methods. Isolation Forest is proficient in detecting anomalies of credit card fraud transactions.

In the Research Paper [9], the Authors clearly stated that the Local Outlier Factor has the highest accuracy than Isolation Forest. LOF performed much better and give the best results when compared with Isolation Forest.

4. ALGORITHMS USED

Credit Card Fraud is a matter of great concern. Everyone wants a secure transaction and hence people take care of various things to keep their cards safe but still sometimes can get trapped in a situation that can lead to a great loss. Hence, the total numbers of fraud transactions are very less as compared to genuine transactions. So, we can say that in any dataset, the fraud transactions act like an outlier.

"*Observation which deviates so much from other observations as to arouse suspicion it was generated by a different mechanism" — Hawkins (1980).*

An object which deviates remarkably from the rest of the objects can be defined as an **Outlier**.

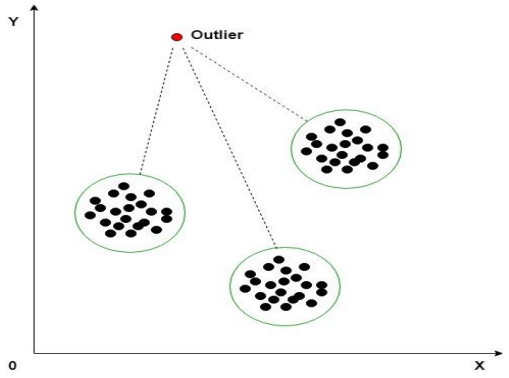

Figure-1: Outliers

These points represented within a circle form a Cluster.

First, we have to calculate the mean for each cluster and then, initialize a threshold value. Hence, an outlier can be identified in the following way:

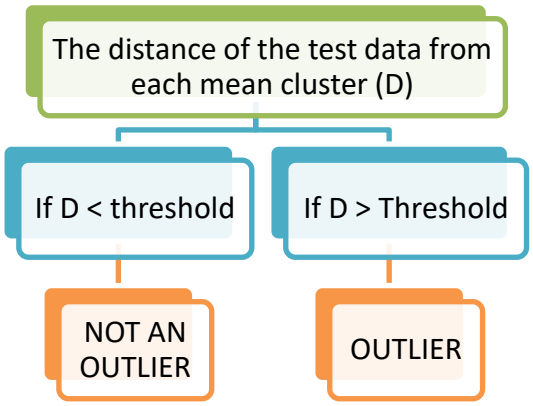

Figure-2: Detection of Outlier

In our project, we have used three methods to find these outliers or frauds.

(I) Local Outlier factor

It is an unsupervised outlier detection technique which works on the principle of density measurement and comparison. It measures the density of different data-points, or their clusters and compares them with one another. The anomaly or the outlier will have less density in comparison with the other data points, or clusters. So, by this technique, we can differentiate the outliers with normal data-points, and hence detect them. There are certain terms associated with LOF:

(a) k

"k" is defined as the number of neighbors. It can be considered as a threshold value for calculations of the local outlier factor. Its value should not be either too small or too large. If k is very small, then it would be able to consider very few points. If k is very large, then it can miss some of the outliers.

(b) k-distance

It is the distance which is based on the value of k. A point's distance to its k^{th} neighbor is called k-distance.

(c) Reachability Distance

If '**a**' and '**b**', are two points, then Reachability distance can be described using the formula:

reach-dist (a,b) = max{k-distance(b), dist(a,b)}

(d) Local Reachability Distance (LRD)

It can be found out using the reachability distance by using the formula:

LRD (a) = 1/(sum(reach-dist(a,n))/k)

(e) Local Outlier Factor (LOF)

The ratio of the LRD of the neighbors of point the **'a'** to the LRD of point '**a**' is known as Local Outlier Factor.

LOF ≈1 ⇒ no outlier

LOF ≫1 ⇒ outlier

(II) Isolation Forest Algorithm

Isolation Forest Algorithm is also a type of anomaly detection algorithm. It is very much similar to Decision trees. The basic difference between the two algorithms is, Decision trees start with a parent node and then keep on partitioning on the basis of information gain, whereas in Isolation forest algorithm, the partitioning occurs randomly on the basis of the value of the selected feature. The selected feature will have a maximum and minimum value which perhaps gives the range of the split value. We continue splitting until we isolate the outlier from other data points.

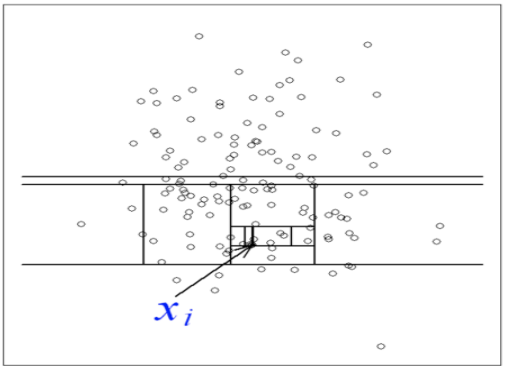

Figure-3: Isolating a Normal Point

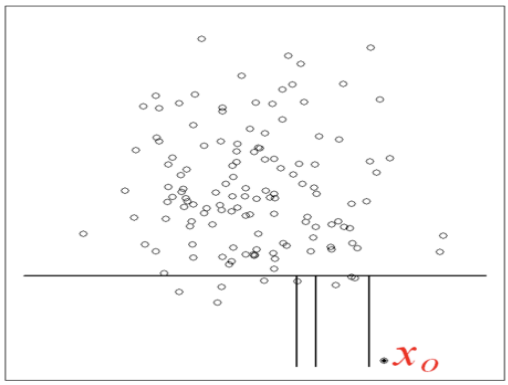

Figure-4: Isolating an Outlier

(a) Isolating a Normal Point

A normal point will be much clustered i.e. it would be surrounded by many data points. Many data points will be there in close proximity to a normal point. So, more splits would be required to isolate a normal point from other data points.

(b) Isolating an Outlier

While, if we want to isolate an outlier or an anomalous point which would not be in the vicinity of other normal data points, less number of splits would be required. An Outlier is a point which lies away from other normal points. There will be no clusters as well as the density will be very less around an outlier. So, it would be easier to find the anomalous point or the outlier.

(III) Support Vector Machine

Support Vector Machine is a supervised learning classifier. It is used for dividing the dataset into two or more classes, but it depends on a number of factors such as Kernel, Hyperplane, Support vectors, etc.

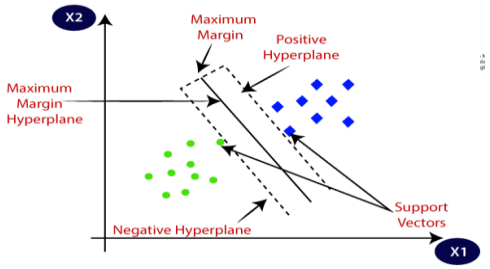

Figure-5 SVM

(a) Hyperplane

Support Vector Machine or SVM can be thought of as a classifying machine. Suppose, we are working in2 dimensional space, having two features X1 and X2, then these two features can be separated using a plane known as Hyperplane which is also known as Maximum Margin Hyperplane. This plane separates the two features linearly. The main aim is to choose a hyperplane with maximum margins. The plane which lies on the right of hyperplane is known as Positive Hyperplane, whereas the plane which lies on the left is known as Negative Hyperplane.

(b) Margins

The distance between the two lines on either side of the features from the hyperplane (center line) is known as Margins. And the combination of the margins of both side of the hyperplane is known as Maximum Margin.

(c) Support Vectors

These are the points which help in making the hyperplane. It supports the positive and negative hyperplanes on either side of the features.

(d) Kernel

Kernels are the tools for analyzing the different types of patterns in SVMs and hence partition the data. They are useful in separating different features from one another.

In our project, we have specifically used **One Class SVM,** which is an unsupervised machine learning algorithm. It trains on normal data and its boundary and hence identifies all the points that lie outside this boundary. It is suitable for Outlier detection.

In the table-1 given below, we have compared these three algorithms on the basis of their advantages and disadvantages. The table shows a theoretical comparison between these algorithms

i.e. Local Outlier Factor, Isolation Forest and Support Vector Machine. The Support Vector Machine algorithm works well with small dataset but here, in Credit Card Fraud Detection, the dataset is humungous, it contains around two lakhs eighty-four thousand records. So, the SVM method took more than 14 hours for its execution and its result was also not appreciable which we will see in the result section.

This comparison table will help the users to unleash the basic concepts of these algorithms as well as will enhance the understanding of readers.

CLASSIFIER	ADVANTAGES	DISADVANTAGES
Local Outlier factor	• LOF identifies outliers in a data set locally. • It works well for unlabelled data.	• It works better for larger datasets rather than smaller ones.
Isolation Forest	• Values need not be scaled in feature space. • Value distributions need not be assumed. • It is robust and easy to optimize because of fewer parameters.	• The result visualization is complicated. • Training time can be very long and computationally expensive if it's not correctly optimized.
Support Vector Machine	• Accuracy is good. • Works well on smaller and cleaner datasets. • It can use a subset of data points, hence making it efficient.	• It is less efficient for noisier datasets with overlapping classes. • Training time is large for larger data sets.

Table-1: Comparison of algorithm

5. METHODOLOGY

We have used three algorithms to find the fraud transactions in our dataset which contains around 284315, transactions. We proceeded according to the given flow chart. For data visualization, we have used 4 different plots. The first plot gives us the total number of fraud as well as normal transactions. The second plot is a histogram between the Amount and Number of Transactions for both Fraud and Normal transactions. The third plot is a scatter plot between the Time of the transaction and the Amount for both Fraud and Normal transactions. Also, we have a heatmap, which clearly explains the total transactions. The fraud transactions are classified as "1" while normal transactions are classified as "0". We have used three classifiers i.e. Local Outlier Factor, Isolation Forest and Support Vector Machine. We fit the model with the dataset and find the result. Isolation Forest has the highest accuracy of 99.74% whereas SVM has the lowest accuracy. Also, SVM takes a large amount of time as compared to the other two methods.

The flowchart given below explains the process flow very efficiently and in a simple and readable format.

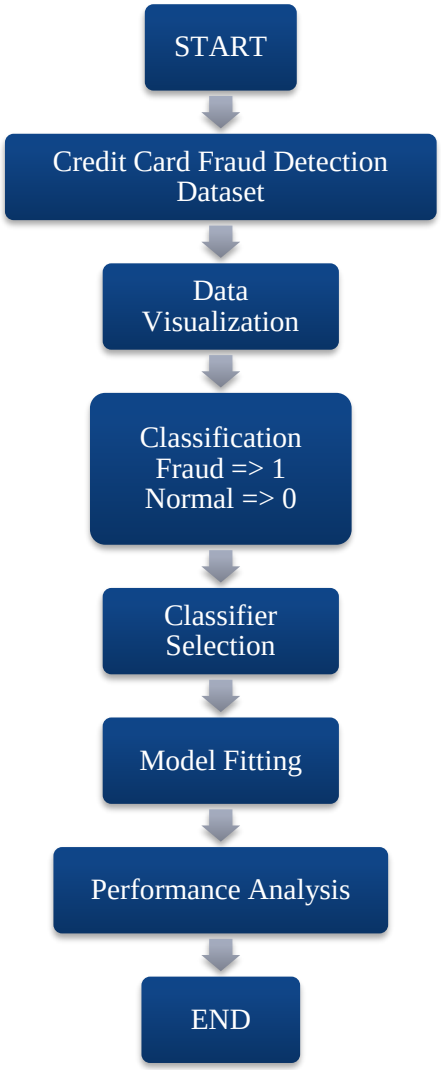

Figure-6 Process Flow

6. RESULTS

The quality of predictions from a classification algorithm can be measured using a **Classification Report**.

The classification report uses four different metrics, namely, **Precision, Recall, F1-score and Support**. These metrics are calculated by using some parameters called true positive, false positive, true negative and false negative. The values of these four parameters can be found out on the basis of True class and Predicted class using the confusion matrix shown in table 2.

		Predicted Class	
		Negative	**Positive**
Actual Class	**Negative**	True Negative	False Positive
	Positive	False Positive	True Positive

Table-2: Confusion Matrix

1. **True Positive (TP):** These are the values which have been predicted correctly. It includes only the positive values which belong to the class "True or Yes".
2. **True Negative (TN):** These are the values which have also been predicted correctly.It includes only the negative values which belong to the class "False or No".
3. **False Positive (FP):** It signifies the wrong prediction. The value of the actual class is NO and the value of the predicted class is YES.
4. **False Negative (FN):** It also signifies the wrong prediction. Here, the value of the actual class is YES and the value of the predicted class is NO.

False Positive and False Negative classes occur in the case of class mismatch between the actual and the predicted class.

Classification Report Metrics are as follows:

1. Precision
2. Recall
3. F1-score
4. Support

Precision: Precision in simple words can be defined as the accuracy of positive predictions. It

is the number of the correct predictions made by the classifier.

Precision = TP / (TP + FP)

Recall: It is the fraction of positive predictions made by the classifier that are correctly identified. It can be used to find the true positive rate.

Recall = TP/(TP+FN)

F1-Score: It is the mean of Precision and Recall.

F1-Score = 2*(Recall * Precision) / (Recall + Precision)

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

The table given below shows the accuracy as well as the number of wrong predictions made by that particular algorithm.

ALGORITHM	NO. OF WRONG PREDICTIONS	ACCURACY
Isolation Forest	683	99.76%
Local Outlier Factor	935	99.67%
Support Vector Machine	71,078	75.52%

Table-3: Accuracy & wrong Predictions

As described above, we have used the method of the classification report which gives us the result on the basis of four parameters i.e. Precision, Recall, F1 score and Support. **Here, Class 1 denotes fraudulent transactions while Class 0 represents the Normal transactions.** We have shown the classification report in table-4.

The Isolation Forest method performs the best out of all the algorithms with an accuracy of 99.76 %. Out of 284315 data entities, While the Support Vector Machine shows the poorest result. It has just the accuracy of 75.52 % and it predicts 71078 records wrongly. Local Outlier Factor also shows the good result, its result is close to that of Isolation forest. It has an accuracy of 99.67 % and misclassifies only 935 values. Both the methods, Isolation Forest and Local Outlier factor also shows a good result and is close to that of the Isolation Forest whereas the Support Vector Machine is not that accurate. Also, SVM is a very slow algorithm. It took more than 14 hours to run which is not at all beneficial.

7. CONCLUSION & FUTURE SCOPE

Fraud Transaction cases can be eliminated to a greater extent with the help of machine learning algorithms. In this paper, we have compared three fraud detection techniques, i.e. Support Vector Machine, Local Outlier Factor and Isolation Forest. In all these fraud detection techniques, the Isolation forest is proficient.

Isolation Forest gives 99.7% accuracy in results when applied on the complete dataset whereas LOF gives 99.6% accuracy and SVM being the slowest one gives 99.7% accuracy. During the execution of code, SVM takes the longest time to give results as compared to other techniques. Its accuracy is also very low as compared to the other two techniques.

We can conclude here that to get the highest accuracy and fast results, use the Isolation Forest-fraud detection technique. It will give you a precise and efficient result.

Despite these outstanding techniques, there will always be a percentage of fraud prevailing, until we integrate good security software/hardware with present software/hardware available in the market. There is always a possibility of improvement. So, for the future scope, we can add a fingerprint sensor in ATM machines to avoid fraudulent transactions. Nowadays, we all use mobile phones or laptops for making transactions. The inbuilt camera in these devices can act as the most useful aid for security. We can also add Iris Detection System in the bank's software or mobile application. Whether the user uses a mobile phone or laptop for the transactions, it will be a safe transaction if the bank's software is integrated with Iris Detection

System. We can also add a Face detection System with the bank's software.

ALGORITHM	CLASS	PRECISION	RECALL	F1 SCORE	SUPPORT
Isolation Forest	0	1.00	1.00	1.00	284315
	1	0.31	0.31	0.31	492
Local Outlier Factor	0	1.00	1.00	1.00	284315
	1	0.05	0.05	0.05	492
Support Vector Machine	0	1.00	0.75	0.85	284315
	1	0.00	0.32	0.00	492

Table-4 Classification Report

8. REFERENCES

[1] Massimiliano Zanin, Miguel Romance, Santiago Moral and ReginoCriado "Credit Card Fraud Detection through Parenclitic Network Analysis", Volume 2018, Article ID 5764370.

[2] Yashvi Jain, Namrata Tiwari, Shripriya Dubey, Sarika Jain, "A Comparative Analysis of Various Credit Card Fraud Detection Techniques",International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-7 Issue-5S2, January 2019.

[3] Ishu Trivedi, Monika, Mrigya Mridushi, "Credit Card Fraud Detection", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016.

[4] Linda Delamaire (UK), Hussein Abdou (UK), John Pointon (UK), "Credit card fraud and detection techniques: a review", Banks and Bank Systems, Volume 4, Issue 2, 2009.

[5] Samaneh Sorournejad, Zahra Zojaji, Reza Ebrahimi Atani, Amir Hassan Monadjemi, "A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective."

[6] Navanshu Khare and Saad Yunus Sait, "Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models", International Journal of Pure and Applied Mathematics Volume 118 No. 20 2018, 825-838.

[7] Masoumeh Zareapoor, Pourya Shamsolmoali, "Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier", ICCC-2015 Procedia Computer Science 48 (2015) 679 – 685.

[8] Soumaya Ounacer, Hicham Ait El Bour, Younes Oubrahim, Mohamed Yassine Ghoumari1 and Mohamed Azzouazi, "Using Isolation Forest in anomaly detection: the case of credit card transactions" Periodicals of Engineering and Natural Sciences ISSN 2303-4521 Vol.6, No.2, December 2018, pp.394-400.

[9] Hyder John, Sameena Naaz, "Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest", Journal of Computer Sciences and Engineering Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693.