Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Credit Card Fraud Detection Using Machine Learning

Ruttala Sailusha
Department Of Information Technology
Velagapudi Ramakrishna **Siddhartha** Engineering College
Vijayawada, India
rsailusha99@gmail.com

V. Gnaneswar
Department Of Information Technology
Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada, India
v.gnaneswar123@gmail.com

R. Ramesh
Department Of Information Technology
Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada, India
rameshraparla59@gmail.com

G. Ramakoteswara Rao
Department Of Information Technology
Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada, India
grkraoganga@gmail.com

***Abstract*—Credit card fraud detection is presently the most frequently occurring problem in the present world. This is due to the rise in both online transactions and e-commerce platforms. Credit card fraud generally happens when the card was stolen for any of the unauthorized purposes or even when the fraudster uses the credit card information for his use. In the present world, we are facing a lot of credit card problems. To detect the fraudulent activities the credit card fraud detection system was introduced. This project aims to focus mainly on machine learning algorithms. The algorithms used are random forest algorithm and the Adaboost algorithm. The results of the two algorithms are based on accuracy, precision, recall, and F1-score. The ROC curve is plotted based on the confusion matrix. The Random Forest and the Adaboost algorithms are compared and the algorithm that has the greatest accuracy, precision, recall, and F1-score is considered as the best algorithm that is used to detect the fraud.**

Keywords—credit card fraud, fraudulent activities, Random Forest, Adaboost, ROC curve

I. INTRODUCTION

Credit card fraud is a growing concern in the present world with the growing fraud in the government offices, corporate industries, finance industries, and many other organizations. In the present world, the high dependency on the internet is the reason for an increased rate of credit card fraud transactions but the fraud has increased not only online but also offline transactions. Though the data mining techniques[6] are used the result is not much accurate to detect these credit card frauds. The only way to minimize these losses is the detection of the fraud using efficient algorithms which is a promising way to reduce the credit card frauds. As the use of the internet is increasing[Figure.1], a credit card is issued by the finance company. Having a credit card means that we can borrow the funds. The funds can be used for any of the purposes. When coming to the issuance of the card, the condition involved is that the cardholder will pay back the original amount they borrowed along with the additional charges they agreed to pay.

A credit card is said to be a fraud when some other person uses your credit card instead of you without your authorization. Fraudsters steal the credit card PIN or the account details to perform any of the unauthorized transactions without stealing the original physical card. Using the credit card fraud detection we could find out whether the new transactions are fraud one or a genuine one.

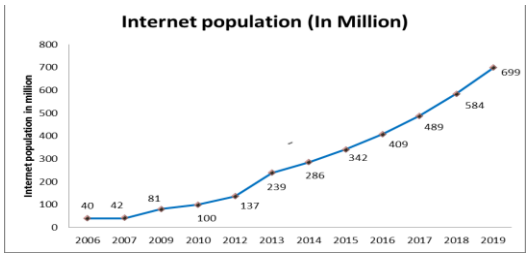


Figure1.Growth of Internet users[2]

The fraud that is committed may involve the card such as a credit card or debit card. In this, the card itself acts as a fraudulent source in the transaction. The purpose of committing the crime may be to obtain the goods without paying money or to obtain the unauthorized fund. Credit cards are a nice target for fraud. The reason is that in a very short time a lot of money can be earned without taking many risks and even the crime will take many weeks to be detected.

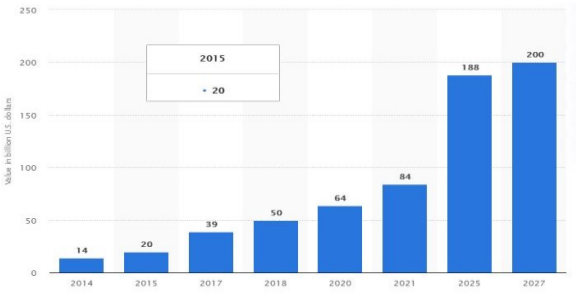


Figure.2 Growth of E-Commerce sites[9]

As the use of the internet nowadays [Figure.2] is very much increasing there may be many chances for the fraudsters to commit the credit card frauds. The main fraud cases that are ongoing in the present world are in those of the e-commerce sites. In the present generation, people are showing much interest in getting things online rather than going and purchasing them, and due to this, the growth of the e-commerce sites is increasing and thereby there is a huge chance of credit card fraud. So to avoid such credit card frauds, we need to find out the best algorithm that reduces credit card frauds.

II. RELATED WORK

New methods for credit card fraud detection with a lot of research methods and several fraud detection techniques with a special interest in the neural networks, data mining, and distributed data mining. Many other techniques are used to detect such credit card fraud. When done the literature survey on various methods of credit card fraud detection, we can conclude that to detect credit card fraud there are many other approaches in Machine Learning itself.

The research on credit card fraud detection uses both Machine Learning[1][2] and Deep Learning algorithms[7]. In this section, we enhance the work done in two different points:(i) the methods that are readily available for fraud detection, and (ii) The techniques that are available to handle the imbalanced data. To handle the imbalanced data A[11] some of the techniques are available. They are (a) classification methods (b) sampling methods (c) resembling techniques. Here are some of the Machine Learning algorithms that are used for credit fraud detection are support vector machine(SVM), decision trees, logistic regression, gradient boosting, K-nearest neighbor, etc;

In 2019, Yashvi Jain, NamrataTiwari, Shripriya Dubey, Sarika jain have researched various techniques[10] for credit cards fraud detection such as support vector machines(SVM), artificial neural networks(ANN), Bayesian Networks, Hidden Markov Model, K-Nearest Neighbours (KNN) Fuzzy Logic system and Decision Trees. In their paper, they have observed that the algorithms k-nearest neighbor, decision trees, and the SVM give a medium level accuracy. The Fuzzy Logic and Logistic Regression give the lowest accuracy among all the other algorithms. Neural Networks, naive bayes, fuzzy systems, and KNN offer a high detention rate. The LogisticRegression, SVM, decision trees offer a high detection rate at the medium level. There are two algorithms namely ANN and the Naïve Bayesian Networks which perform better at all parameters. These are very much expensive to train. There is a major drawback in all the algorithms. The drawback is that these algorithms don't give the same result in all types of environments. They give better results with one type of datasets and poor results with another type of dataset. Algorithms like KNN and SVM give excellent results with small datasets and algorithms like logistic regression and fuzzy logic systems give good accuracy with raw and un-sampled data.

In 2019, Heta Naik, Prashasti Kanikar, has done their research on various algorithms [4] like Naïve Bayes, Logistic Regression, J48, and Adaboost. Naïve Bayes on among the classification algorithm. This algorithm depends upon Bayes theorem. Bayes's theorem finds the probability of an event that is occurring is given. The Logistic regression algorithm is similar to the linear regression algorithm. The linear regression is used for the prediction or forecasting the values. The logistic regression is mostly used for the classification task. The J48 algorithm is used to generate a decision tree and is used for the classification problem. The J48 is the extension of the ID3 (Iterative Dichotomieser). J48 is one of the most widely used and extensively analyzed areas in Machine Learning. This algorithm mainly works on constant and categorical variables. Adaboost is one of the most widely used machine learning algorithms and is mainly developed for binary classification. The algorithm is mainly used to boost the performance of the decision tree. This is also mainly used for the classification of the regression. The Adaboost algorithm is fraud cases to classify the transactions which are fraud and non-fraud. From their work they have concluded that the highest accuracy is obtained for both the Adaboost and Logistic Regression. As they have the same accuracy the time factor is considered to choose the best algorithm. By considering the time factor they concluded that the Adaboost algorithm works well to detect credit card fraud.

In 2019 Sahayasakila V, D.Kavya Monisha, Aishwarya, Sikhakolli Venkatavisalakshiswshai Yasaswi have explained the Twain important algorithmic techniques [8] which are the Whale Optimization Techniques (WOA) and SMOTE (Synthetic Minority Oversampling Techniques). They mainly aimed to improve the convergence speed and to solve the data imbalance problem. The class imbalance problem is overcome using the SMOTE technique and the WOA technique. The SMOTE technique discriminates all the transactions which are synthesized are again re-sampled to check the data accuracy and are optimized using the WOA technique. The algorithm also improves the convergence speed, reliability, and efficiency of the system.

In 2018 Navanushu Khare and Saad Yunus Sait have explained their work [5] on decision trees, random forest, SVM, and logistic regression. They have taken the highly skewed dataset and worked on such type of dataset. The performance evaluation is based on accuracy, sensitivity, specificity, and precision. The results indicate that the accuracy for the Logistic Regression is 97.7%, for Decision Trees is 95.5%, for Random Forest is 98.6%, for SVM classifier is 97.5%. They have concluded that the Random Forest algorithm has the highest accuracy among the other algorithms and is considered as the best algorithm to detect the fraud. They also concluded that the SVM algorithm has a data imbalance problem and does not give better results to detect credit card fraud.

III. PROPOSED WORK

The main aim of this paper is to classify the transactions that have both the fraud and non-fraud transactions in the dataset using algorithms like that the Random Forest and the Adaboost algorithms. Then these two algorithms are compared to choose the algorithm that best detects the credit card fraud transactions. The process flow for the credit fraud detection problem [Figure.3.]includes the splitting of the data, model training, model deployment, and the evaluation criteria.

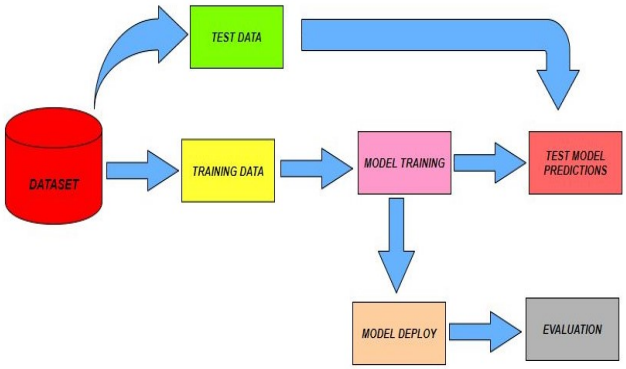


Figure.3 Process Flow

The detailed architecture diagram for the credit card fraud detection system [Figure. 4.] includes many steps from gathering dataset to deploying model and performing analysis based on results. In this model we take the Kaggle credit card fraud dataset and pre-processing is to be done for the dataset. Now to prepare the model we have to split the data into the training data and the testing data. We use the training data to prepare the Random Forest and the Adaboost models. Then we develop both the models. Finally, the accuracy, precision, recall, and F1-score is calculated for bot the models. Finally the comparison of the credit card fraud transactions more accurately.

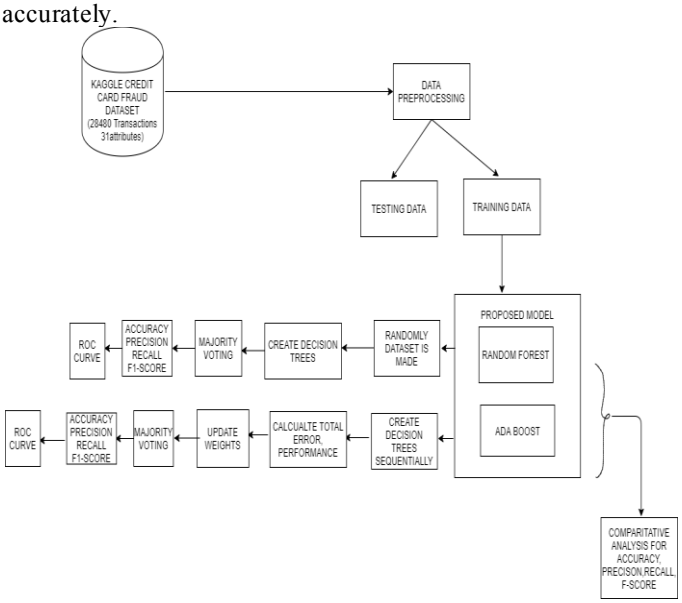


Figure.4 Architecture Diagram

A. Random Forest Algorithm

The Random Forest algorithm [Figure. 5]is one of the widely used supervised learning algorithms. This can be used for both regression and classification purposes. But, this algorithm is mainly used for classification problems. Generally, a forest is made up of trees and similarly, the Random Forest algorithm creates the decision trees on the sample data and gets the prediction from each of the sample data. Then Random Forest algorithm is an ensemble method. This algorithm is better than the single decision trees because it reduces the over-fitting by averaging the result.

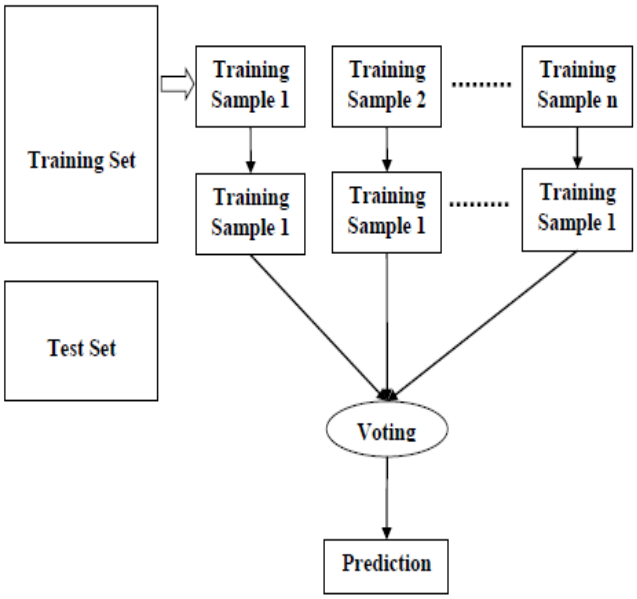


Figure.5 Random Forest Algorithm

Steps for Random Forest Algorithm

1. Take the Kaggle credit card fraud dataset that is trained and randomly select some of the sample data.
2. Using the randomly created sample data now creates the Decision Trees that are used to classify the cases into the fraud and non-fraud cases.
3. The Decision Trees are formed by splitting the nodes, the nodes which have the highest Information gain make it as the root node and classify the fraud and non-fraud cases.
4. Now the majority vote is performed and the decision Trees may result in 0 as output which includes that these are the non-fraud cases.
5. Finally, we find the accuracy, precision, recall, and F1 -score for both the fraud and non-fraud cases.

Random Forest algorithm

Algorithm Random Forest :
To generate c classifiers:
For i=1 to c do
Randomly select the training data D with replacement to produce Di

Create a root node N containing Di and cell
Build Tree(N)
End for
Majority Vote

Build Tree(N)
Randomly select x% of all the possible splitting features in N
Select the features F that has the highest Information A gain for further splitting
Gain (T,X)=Entropy (T)-Entropy(T,X)
Now to calculate the entropy we use,
$E(S) = \sum_{i=1}^{c}(-Pi \log Pi)$
Create f child nodes
For i=1 to f do
Set contents f N to Di
Call Build Tree(Ni)
End for

End

B. Adaboost Algorithm

Boosting is one of the ensemble techniques. This algorithm is used to build strong classifiers from weaker classifiers. This can be done by building a strong model by using a weak model in the series. Initially, a model is built from the training data. Then the second model is built from the first model by correcting the errors that represent in the model that is created before. This is a repetitive process and is continued until either the maximum number of models is added or the complete training dataset is predicted correctly. Adboost was one of the most successful boosting algorithms that were developed for the binary classification.

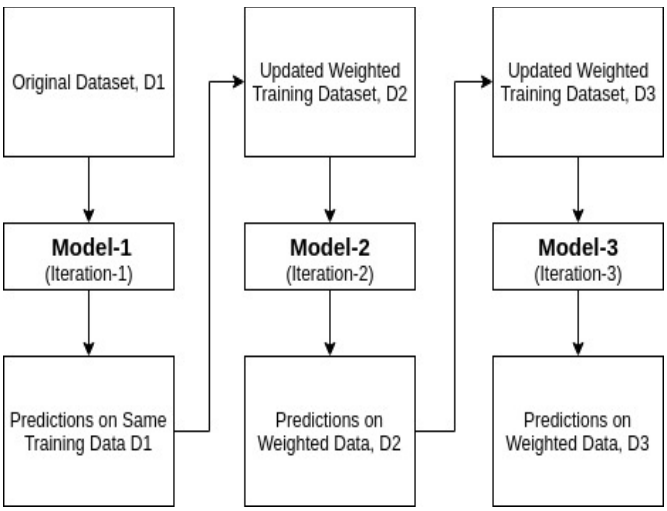


Figure.6 Adaboost Algorithm

The short name for Adaboost is adaptive boosting. It is best used with weak learners. This Adaboost boosting technique [Figure. 6]combines the multiple weak classifiers into a strong classifier. Adaboost algorithm can be used with short decision trees. The way the Adaboost is created is such that initially at first the nodes are created and the tree is made, then the performance of the tree on each of the instances is checked. Also, a weight is assigned. The training data that is hard to predict is the one that gives more weight. The Adaboost algorithm is a powerful classifier that works well on both the basic and complex problems. The disadvantage of this algorithm is that this algorithm is mostly sensitive to noisy data. This algorithm is also sensitive to outliers.

Steps for Adaboost Algorithm

1. The Kaggle credit card fraud dataset is taken and is trained. Randomly select some of the sample data.
2. Using the randomly created sample data now creates the decision trees sequentially for classifying the fraud and non-fraud cases.
3. The decision trees are formed initially. This can be done by splitting the node based on which has the highest information gain, make it as the root node, and classify the fraud and non-fraud cases.
4. Now calculate the error rate, performance, and update the weights of the fraud and non-fraud transactions that are incorrectly classified.
5. Now majority vote is performed and the decision trees may result as output which indicates the non-fraud cases.
6. The decision trees may output 1 which indicates that it is a fraud case.
7. Finally, we find the accuracy, precision, recall, and F1-score for both the fraud and non-fraud cases.

Adaboost Algorithm

Algorithm Adaboost :
IINPUTdataset
Initialize weights, w1(n)=1/n
Create a decision tree
Select the one that has the lowest Entropy
If Incorrectly classified
Calculate Total Error (TE)= sum of up incorrectly Classified sample weights
Calculate Performance, $P = \log\frac{1-TE}{TE}$
For each
Incorrectly classified, increase weights:
Weights incorrect =old weight $*e^{p}$
Correctly classified, decrease the weights:
Weight correct =old weight $*e^{-p}$
Normalized weight of each sample:

$$\text{Normalized weight} = \frac{updated\ weight}{sum\ of\ updated\ wight}$$

End for
End if

IV. EVALUATION AND RESULT ANALYSIS

A. Dataset

The dataset, credit card fraud data is taken from the European credit card company. The dataset is obtained from the Kaggle. The dataset holds the transactions that are done by the credit cardholders in the year 2013 September. The dataset contains the transactions that are done in two days. The data set contains 284,807 transactions in which 492 transactions are a fraud. These fraud transactions account for only 0.172%of all the transactions. The dataset having the input variable are converted into the numerical values by the PCA transformation. This is done due to confidentiality reasons. The features 'Time' and 'Amount 'can't be PCA transformed. The class 'Time 'represents the difference in the seconds elapsed between the particular transaction and the first transaction. The class 'Amount ' represents the money transaction that had occurred. Another important feature 'Class' shows whether the transaction is fraudulent or not. The number indication 1 shows that it is a fraud transaction and 0 indicates the non-fraud transactions.

B. Evaluation Criteria

To compare various algorithms, we need to evaluate metrics like accuracy, precision, recall, and F1-score. The confusion matrix is also plotted. The confusion matrix is a 2*2 matrix. The matrix contains four outputs which are TPR, TNR, FPR, FNR. Measures such as sensitivity, specificity, accuracy, and error-rate can be derived from the confusion matrix. Then we that best suit to detect the credit card fraud.

The output of the confusion matrix is

1. True Positive Rate, which can be defined as the number of fraudulent transactions that are even classified by the system as fraudulent.
2. True Negative Rate, which can be defined as the number of legitimate transactions that are even classified as legitimate by the system.
3. False Positive Rate, which can be defined as a number of the legal transactions which are wrongly classified as fraud.
4. False Negative Rate is defined as the transactions that are fraud but are wrongly classified as legal.

The Receiver Operating Characteristics curve is created by plotting the TPR against the FPR. This can be done at various thresholds. ROC curve is a graph in which the FPR is the horizontal axis and the TPR is the vertical axis. The graph under the ROC curve is the AUC.

C. Results Analysis

The confusion matrix and the ROC curve is plotted for both the algorithms. The dataset, when applied for different algorithms, gives different outputs. Firstly we apply the dataset for the random forest model and the results are as below:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	93825
1	0.95	0.77	0.85	162
accuracy			1.00	93987
macro avg	0.97	0.89	0.93	93987
weighted avg	1.00	1.00	1.00	93987

Figure.7 Output for Random Forest

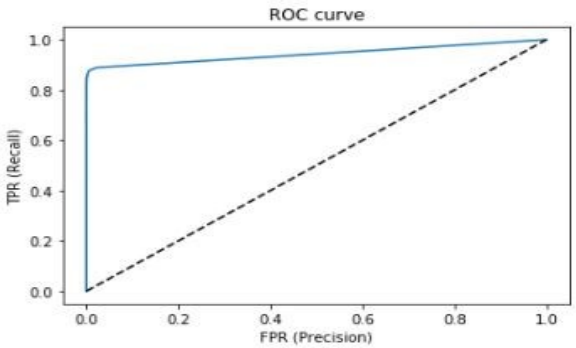
The evaluation criteria are explained[Figure.7] and the precision, recall, F1-score are the same for that of the non-fraud cases and differ for that of the fraud cases.

```
Confusion Matrix on train data
[[190490      0]
 [     0    330]]

Confusion Matrix on test data
[[93818    37]
 [    7   125]]
```

Figure.8 Confusion Matrix for Random Forest

The confusion matrix[Figure.8] shows us that for the train data the true positives are 190490 and false positives are 0, the true negatives are 0 and the false negatives are 330. For the test data, the true positives are 93818 and false positives are 37, the true negatives are 7 and the false negatives are 125.




```
Area under curve (AUC):  0.9429434888303349
```

Figure.9 ROC curve for Random Forest

Now the dataset is applied for the Adaboost algorithm. The results are obtained similar to that of the Random Forest Algorithm.

```
Accuracy = 0.9990743400683073
              precision    recall  f1-score   support

           0  0.99938202 0.99969091 0.99953644     93825
           1  0.78195489 0.64197531 0.70508475       162
```

Figure.10 Output for Adaboost

The evaluation criteria[Figure.11] shows us that the evaluation criteria like the precision, recall, and F1-score differ less in the case of the non-fraud cases and differ greatly in those of the fraud cases.

```
Confusion Matrix on train data
[[190464    120]
 [    26    210]]
Confusion Matrix on test data
[[93811    65]
 [   14    97]]
```

Figure.11 Confusion Matrix for Adaboost

The confusion matrix [Figure.11] shows us that for the train data the true positives are 190464 and false positives are 120, the true negatives are 26 and false negatives are 201. For the test data, the true positives are 93811 and false positives are 65, the true negatives are 14 and false negatives are 97.

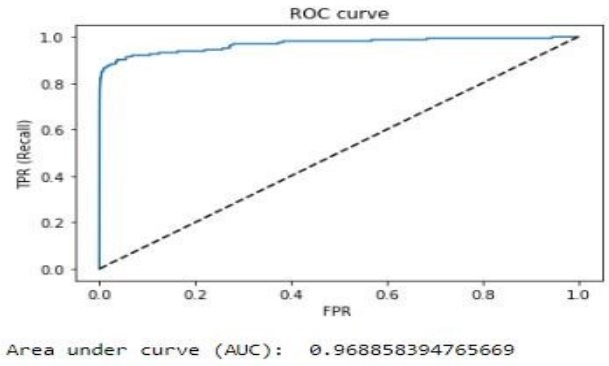


Figure.12 ROC curve for Adaboost

Now the comparison of the random forest and the Adaboost algorithms is shown [Figure.12]. The two algorithms have the same accuracy but the precision, recall, and the F1-score of the two algorithms differ. The random forest algorithms have the highest precision, recall, and F1-score.

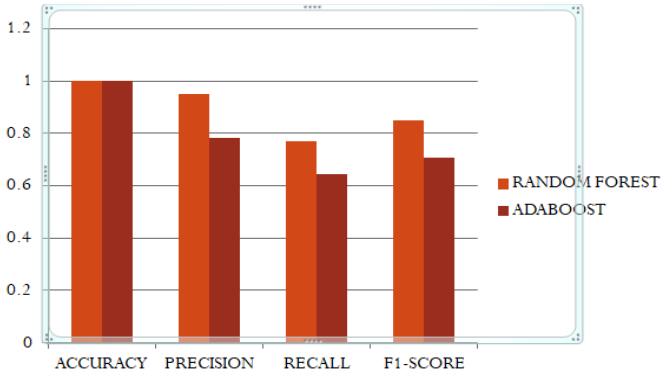


Figure.13 Comparision of Algorithms

V. CONCLUSION

Even though there are many fraud detection techniques we can't say that this particular algorithm detects the fraud completely. From our analysis, we can conclude that the accuracy is the same for both the Random Forest and the Adaboost algorithms. When we consider the precision, recall, and the F1-score the Random Forest algorithm has the highest value than the Adaboost algorithm. Hence we conclude that the Random Forest Algorithm works best than the Adaboost algorithm to detect credit card fraud.

VI. FUTURE SCOPE

From the above analysis, it is clear that many machine learning algorithms are used to detect the fraud but we can observe that the results are not satisfactory. So, we would like to implement deep learning algorithms to detect credit card fraud accurately.

REFERENCES

1. Adi Saputra1, Suharjito2L: Fraud Detection using Machine Learning in e-Commerce, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 9, 2019.
2. Dart Consulting,Growth Of Internet Users In India And Impact On Country's Economy: https://www.dartconsulting.co.in/market-news/growth-of-internet-users-in-india-and-impact-on-countrys-economy/
3. Ganga Rama Koteswara Rao and R.Satya Prasad, "- Shielding The Networks Depending On Linux Servers Against Arp Spoofing, International Journal of Engineering and Technology(UAE),Vol. 7, PP.75-79, May 2018, ISSN No: 2227-524X, DOI - 10.14419/ijet.v7i2.32.13531.
4. Heta Naik , Prashasti Kanikar: Credit card Fraud Detection based on Machine Learning Algorithms,International Journal of Computer Applications (0975 – 8887) Volume 182 – No. 44, March 2019.
5. Navanshu Khare ,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 ISSN: 1314-3395.
6. Randula Koralage, , Faculty of Information Technology, University of Moratuwa,Data Mining Techniques for Credit Card Fraud Detection.
7. Roy, Abhimanyu, et al:Deep learning detecting fraud in credit card transactions, 2018 Systems and Information Engineering Design Symposium (SIEDS), IEEE, 2018.
8. Sahayasakila.V, D. Kavya Monisha, Aishwarya, Sikhakolli VenkatavisalakshiseshsaiYasaswi: Credit Card Fraud Detection System using Smote Technique and Whale Optimization Algorithm,International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-5, June 2019.
9. Statista.com. retail e-commerce revenue forecast from 2017 to 2023 (in billion U.S. dollars). Retrieved April 2020, from India : https://www.statista.com/statistics/280925/e-commerce-revenueforecast-in-india/
10. Yashvi Jain, NamrataTiwari, ShripriyaDubey,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.
11. Yong Fang1, Yunyun Zhang2 and Cheng Huang1, Credit Card Fraud Detection Based on Machine Learning, Computers,

Materials & Continua CMC, vol.61, no.1, pp.185-195, 2019.

12. Kaithekuzhical Leena Kurien, Dr. Ajeet Chikkamannur: Detection And Prediction Of Credit Card Fraud Transactions Using Machine Learning , International Journal Of Engineering Sciences & Research Technolog.