Detection of Credit Card Fraudulent Transactions using Boosting Algorithms

[1] Prathima Gamini, [2] Sai Tejasri Yerramsetti, [3] Gayathri Devi Darapu, [4]Vamsi Kaladhar Pentakoti ,[5] Vegesna Prudhvi Raju
[1]Assistant Professor, [2]Student, [3]Student, [4]Student, [5]Student
[1]Electronics and Communication,
[1]Sagi Ramakrishnam Raju Engineering College, Bhimavaram, India

***Abstract*: In today's socio-economic scenario, people rely heavily on credit cards. Moreover, credit cards are a requisite financial tool that enables its holders to make assets. It is true that credit cards, as a new method of payment, have become socially amenable to the masses. But nowadays, improvement in technology directs to growth in illegal activities. During credit card transactions many fraudsters can breach security and make fraudulent transactions to withdraw or transfer funds from one's account or e-wallets. In this paper, three Boosting algorithms such as CatBoost, XGBoost and Stochastic gradient boosting algorithms are applied for the identification of frauds achieved using credit cards. Boosting assists to obtain an accurate result. For CatBoost algorithm, the evaluation of metric parameters namely precision, recall and confusion matrix is the best when compared to XGBoost and SGB boosting algorithms for the classification of fraudulent or non-fraudulent transactions.**

***Index Terms*: Machine learning; Credit card fraudulent transactions; Boosting algorithms.**

I. INTRODUCTION

Frauds are deliberate act to obtain unauthorized benefits which are being accelerated every day. In recent days, online and offline credit card transactions have been the key method of payment. That's why these frauds have been expedited.

Credit card fraud is an act of criminal deception. In this paper, the common methods of fraud and the fraud detection methods are discussed. This paper also shows how machine learning algorithms assist to find fraud detection. From an ethical frame of reference, it can assert that banks and credit card agencies should try to detect all these frauds.

The credit card data is prone to attack by the cyberpunk. Illegal activists will try to make every fraudulent transaction permissible which is very difficult to detect. In essence, for these illegal activists the effective choices are to seek potential documentation of fraud from the available data using mathematical algorithms. There are various algorithms for credit card fraud detection such as machine learning, genetic algorithms, artificial and neural networks, etc. During this fraud detection, so many problems are to be faced. The process for a successful or failed transaction takes very little time that is in seconds the transaction will be done. Therefore, to find a fraud transaction the process must be very fast and accurate to find the detection.

But there is another problem, with too many transactions happening simultaneously, it is difficult to identify the fraudulent transaction. Hereby, analysation of every transaction. Thus, an efficient fraud detection system works on the principle of learning user-specific card usage behaviour.

Machine Learning is a subclass of AI where the machine is trained to find out from its past experience. The past experience is developed through the information collected. Then it combines with algorithms like Naive Bayes, Support Vector Machine (SVM) to deliver the ultimate results. Machine learning is the study of computer algorithms which upgrades consistently through experiences. Machine learning algorithms are used in many applications such as email filtering and computer vision, where it is unachievable to evolve typical algorithms to perform the needed task. Machine learning is also known as predictive analytic. Teaching the machines involves a systemic process where every stage builds a better version of the machine. For simplification purpose, the process of teaching machines can be classified into three groups:

1. Data as input
2. Abstracting the data
3. Generalization

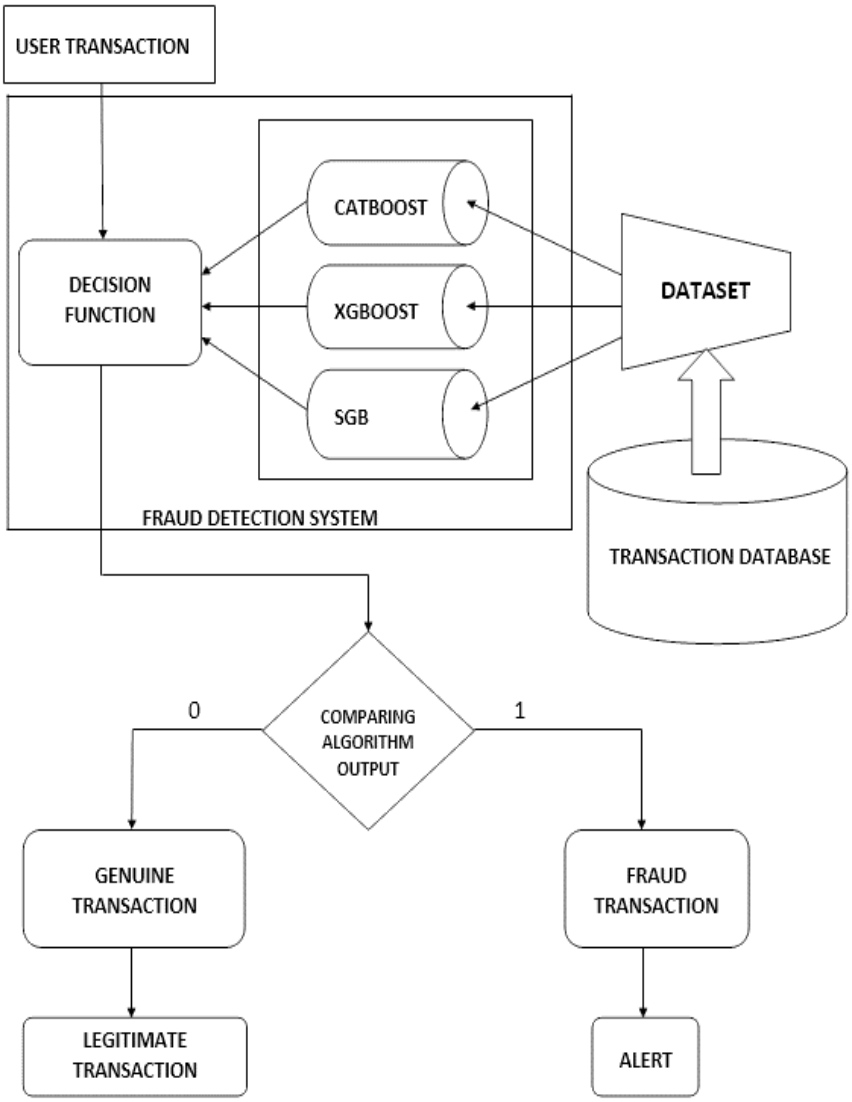

Fig.1.1 Fraud detection system using boosters

The applications of machine learning algorithms are not restricted to these sectors only. Even many researchers are still working on many areas in which machine learning algorithms are applicable and will give better results.

In this paper, machine learning algorithms are applied for detecting frauds during credit card transactions. The next section deals with the related work.

Most Common credit card frauds:

There are five salient types of credit card fraud, according to the Australian Payments Network, they are:

- Card-Not-Present (CNP) fraud
- Counterfeit and skimming fraud
- Lost and stolen card fraud
- Card-never arrived-fraud
- False application fraud

The dataset used comes under counterfeit fraud. Counterfeit fraud takes place when a criminal skim or copies the data on the magnetic stripe of a legalized credit card and places the copied data on a fake credit card that carries all the details. By using these fake cards, they can buy things or can withdraw money at banks etc.

II. RELATED WORK

In this paper, some determined algorithms based on artificial intelligence and neural networks are suggested and implemented to forecast the credit card fraud detection. In neural networks there are two methods are supervised and unsupervised learning but, in this paper, we use supervised learning only. The allocation of datasets used in detection is imbalanced. So, to overcome this we use some data mining techniques

to balance the data. Several advanced techniques based on Sequence Alignment, Machine learning, Artificial Intelligence, Genetic Programming and Data mining etc. has been developed and is being developed to detect fraudulent transactions. In this paper, three algorithms namely XGBoost, CatBoost, and SGB are used to detect fraudulent detections. These algorithms are implemented using a training dataset of approximately two lakh credit card transactions.

A. XGBOOST ALGORITHM:

It stands for extreme gradient boosting. It is designed for speed and performance and is an implementation of gradient boosted machines. It is a software library that has access to various interfaces such as C++, Python interface, R interface, Command-Line interface, Java, etc. It supports gradient boosting, stochastic gradient boosting, regularized gradient boosting. This algorithm implements features like sparse aware, block structure, continued training. The XGBoost library enables a system to be used in a wide range of environments such as cache optimization, out-of-core computing, distributed computing, parallelization. XGBoost implements the gradient boosting decision tree algorithm or multiple additive regression trees or stochastic gradient boosting. Gradient boosting is an approach, where new models are added to the existing model that can predict the errors of the prior model. Models are added sequentially until no further improvements can be made, so that it gives the improvised prediction. Here the Gradient descent algorithm is used to minimize the errors. Since it is used for binary classification, the model will be XGB Classifier.

B. CATBOOST ALGORITHM:

CatBoost refers to category and boosting. CatBoost is used in many frameworks such as TensorFlow, core ML and used for business problems. It produces output without extensive data training, provides out-of-the-box support for more descriptive data formats, easy to use, and has high performance. It lowers overfitting and reduces the need for extensive tuning of hyper-parameters, doesn't require conversion of the dataset and can be used for both classification and regression. For classification, "CatBoost Classifier" is used. The main difference between this and other boosting algorithms is that, it implements symmetric trees. CatBoost trains log (number of data points) models to calculate the residuals for each data point. It considers only past data points to calculate, so that it requires less prediction time. But if it has many numerical features then it takes so much time to train.

C.STOCHASTIC GRADIENT BOOSTING (SGB) ALGORITHM:

A stochastic gradient boosting is the modification of gradient boosting by injecting randomness. Injection of randomness will improve the performance. At each iteration, a subsample of the training data from the full training dataset is drawn at random without replacement. Instead of full sample, this randomly selected subsample is used to fit the base learner. The subsamples allow one to define an out-of-bag error of the prediction performance improvement by evaluating predictions on those observations which is not employed in the building of the upcoming base learner. Out-of-bag estimates and helps to avoid the necessity for an independent validation dataset, but often underestimate actual performance improvement, and therefore, the optimal number of iterations.

III. PARAMETERS:

Confusion Matrix: It defines the performance of a classification model on a set of test data. Confusion matrix defines true positives, true negatives, false positives, false negatives.

Precision: It is also called positive predictive value. It is a fraction of true positives and all the positives. It gives the measure of relevant data points.

$$precision = \frac{TRUE\ POSITIVES}{TRUE\ POSITIVES+FALSE\ POSITIVES}$$

Recall: It is also called sensitivity or true positive rate. It is a measure of correctly identifying the true positives.

$$recall = \frac{TRUE\ POSITIVES}{TRUE\ POSITIVES+FALSE\ NEGATIVES}.$$

These parameters are used for comparison of performance. The prediction accuracy as well as the recall among the three techniques is very high and almost equal to each other but CATBOOST algorithm has the highest prediction accuracy among the three.

IV. RESULT AND ANALYSIS

Machine Learning (ML) algorithms are implemented on the data set of 2,48,007 credit cards. These machine learning algorithms are applied using the python programming language and platform used is Anaconda IDE. The results of three machine learning algorithms are as follows:

A. XGBOOST ALGORITHM:

Figure 4.1 shows the confusion matrix of the prediction results obtained by applying XGBoost machine learning algorithm. From the figure 4.1 it is observed that the XGBoost algorithm correctly predicts zeroes in the final output 78 times and 1 time it predicts the zero incorrectly. XGBoost algorithm correctly predict one's 50 times and 11 times it predicts one's incorrectly.

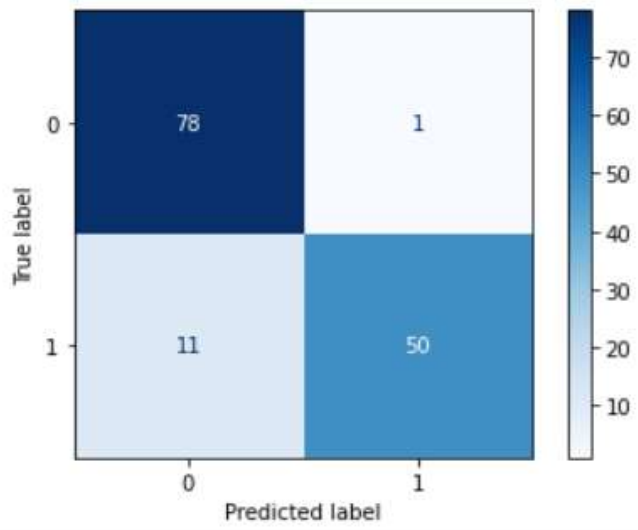

Fig.4.1 Confusion matrix for XGBoost algorithm

B. CATBOOST ALGORITHM:

Figure 4.2 shows the confusion matrix of the prediction results obtained by applying the CatBoost machine learning algorithm. From figure 4.2 it is observed that the CatBoost algorithm correctly predicts zeroes in the final output 79 times and 0 times it predicts the zero incorrectly. CatBoost algorithm correctly predict one's 50 times and 11 times it predicts one's incorrectly.

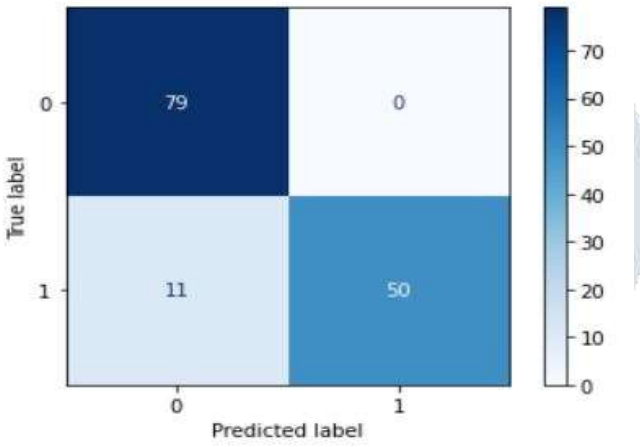

Fig.4.2 Confusion matrix for CatBoost algorithm

C. SGB ALGORITHM:

Figure 4.3 shows the confusion matrix of the prediction results obtained by applying the SGB machine learning algorithm. From figure 4.3 it is observed that the SGB algorithm correctly predicts zeroes in the final output 77 times and 2 times it predicts the zero incorrectly. SGB Algorithm correctly predict one's 50 times and 11 times it predicts one's incorrectly.

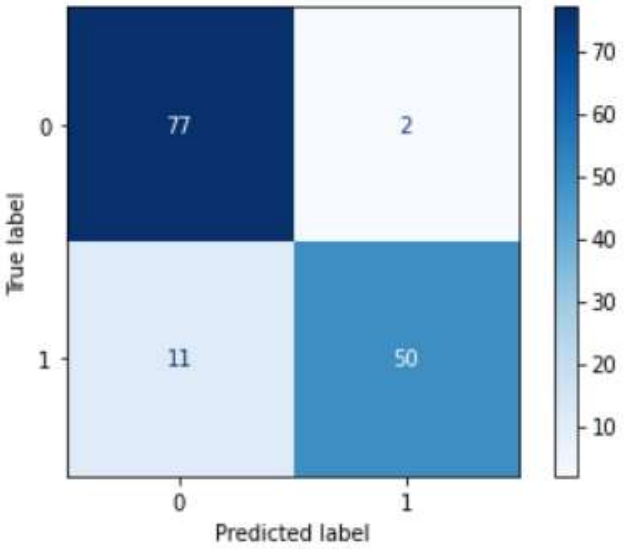

Fig.4.3 Confusion matrix for SGB algorithm

TABLE I. Accuracy of Machine Learning Algorithms for Credit Card Fraud Detection

S. No	Boosting Algorithm	Prediction Accuracy of Machine learning Algorithm	Recall
1	XGBOOST	0.9142	0.99
2	CATBOOST	0.9214	1.00
3	SGB	0.9071	0.97

The three algorithms are tested on the basis of their prediction accuracy, recall and confusion matrix.

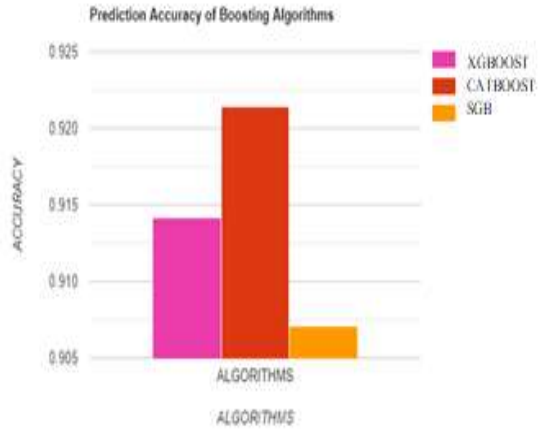

Fig.4.4. Performance comparison of boosting algorithms.

Figure 4.4 shows the representation of the results using a bar chart graph. From the graph the performance is compared easily and CATBOOST algorithm gives the highest accuracy.

V. CONCLUSION AND FUTURE SCOPE:

Boosting algorithms are sensitive to outliers and are difficult to scale up. More room for improvement can be found by improving the performance of scaling and increasing the size of the dataset. The precision of the algorithm increases as the size of the dataset increases. In this paper, boosting algorithms namely XGBoost algorithm, CatBoost algorithm and SGB algorithm are analysed by using metric parameters namely precision, confusion matrix and recall. Finally, it turns out that the performance of CatBoost algorithm is the best when compared to SGB algorithm and XGBoost algorithm.

In future, different models of supervised and unsupervised learning can be examined while handling highly imbalanced credit card fraud data in order to acquire better precision and accuracy.

VI. REFERENCES

[1]. Mohammed, Emad, and Behrouz Far. "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study." IEEE Annals of the History of Computing, IEEE, 1 July 2018, doi.ieeecomputersociety.org/10.1109/IRI.2018.00025.

[2]. Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. Fraud detection system: A survey. Journal of Network and Computer Applications, 2016

[3]. Gustavo Batista, Andre Carlos Carvalho, and Maria Carolina Monard. Applying one-sided selection to unbalanced datasets. MICAI 2000: Advances in Artificial Intelligence, 2000.

[4]. Nitesh V. Chawla. Data mining for imbalanced datasets: an overview. Data mining and knowledge discovery handbook, 2005.

[5]. A. Shen, R. Tong, Y. Deng, "Application of classification models on credit card fraud detection", Service Systems and Service Management 2007 International Conference, pp. 1-4, 2007.

[6]. Raj S.B.E., Portia A.A., Analysis on credit card fraud detection methods, Computer, Communication and Electrical Technology International Conference on (ICCCET) (2011), 152-156.

[7]. Jain R., Gour B., Dubey S., A hybrid approach for credit card fraud detection using rough set and decision tree technique, International Journal of Computer Applications 139(10) (2016).

[8]. K. Chaudhary, J. Yadav, and B. Mallick, "A review of Fraud Detection Techniques: Credit Card," Int. J. Comput. Appl., vol. 45, no. 1, pp. 975-8887, 2012.

[9]. O. S. Yee, S. Sagadevan, N. Hashimah, and A. Hassain, "Credit Card Fraud Detection Using Machine Learning as Data Mining Technique," vol. 10, no. 1, pp. 23-27.

[10]. J. Awoyemi, A. Adetunmbi and S. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis", 2017 International Conference on Computing Networking and Informatics (ICCNI), 2017.

[11]. S. Mittal and S. Tyagi, "Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection", 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019.

[12]. "Survey Paper on Credit Card Fraud Detection by Suman", Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014

[13]. Kou, Y., Lu, C-T., Sinvongwattana, S. and Huang, Y-P., (2004). Survey of Fraud Detection Techniques, In Proceedings of the 2004 IEEE International Conference on Networking, Sensing & Control, Taipei, Taiwan, March 21-23.

[14]. Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2013). Cost sensitive credit card fraud detection using Bayes minimum risk. In Machine Learning and Applications (ICMLA), 2013 12th International Conference on (Vol. 1, pp. 333-338). IEEE.