International Journal of Scientific Research and Engineering Development-– Volume 4 Issue 1, Jan-Feb 2021

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Detecting Credit Card Frauds Using KPCA with Classification Methods

Sumira Jabin*, Hamid Ghous**

*(Institute of Southern Punjab, Multan, Pakistan

Email: sumirajabin786@gmail.com)

** (Institute of Southern Punjab, Multan, Pakistan

Email: hamidghous@gmail.com)

----------**********----------

Abstract:

Regardless of various techniques adopted for credit card fraud detection. Credit card frauds are increasing day by day. Credit card fraud detection is a serious issue and incurs loss to financial institutions and customers. As credit card has become a popular method of payment. There are various ways to steal money like skimming, stealing card and making duplicate card. This study has tried to find out and make a review that how these frauds could be detected using different techniques. Different methods have used till now to tackle this issue. Some of these are classification techniques like Artificial Neural Network, Biological Neural Network and Outlier Detection. But instead of all this it has observed that the credit card frauds are increasing, that has caused a great loss to banks and customers. Due to imbalanced and high dimensional data sets fraud detection in the field of credit card became difficult and ambiguous. Our main objective is to find a way to detect credit card frauds using feature selection and classification methods. In this study, we are using Kernel Principal Component Analysis (KPCA) as a feature selection method with decision tree and boost as classification methods. We have found better and improved results than the previous studies.

Keywords — **Credit Card, Kernel Principal Component Analysis (KPCA), Receiver Operating Characteristic (ROC), Decision Tree, Boost.**

----------**********----------

I. INTRODUCTION

The formation of credit card goes back to 1946 when New York National bank issued charge card to its loan customers. In 1950 Diners club card were issued when Frank Mc Namara forgot his purse in a dinner. The bill of this kind of cards had paid at the

end of the month. By the end of 1951, the users of dinner card were about 42,000. In 1958 American Express Card was launched. In 1966 California bank issued maser card. In 1980 magnetic strip that was on the back of card was issued. Now days credit cards issued by the banks are used for purchasing all over the world. The era of credit card has facilitated to banks and to the customers. The banks charges against the purchase that the card holder made, that benefits the banks financially. On the other side the consumers will be able to make purchases, pay bills and can also buy tickets even when they have no money. The regular card use will also be able to get extra rewards. Transfer of balance from one credit card to other is also possible [18].

Due to the increasing rate of frauds in credit card industry the fraud detection measures are introduced by different organizations. In this study we are using data mining methods to detect credit card frauds. Data mining is also recognized as knowledge discovery (KDD) in data. Data mining field uses datasets and features selection as a tool. In the field of data mining different machine learning methods are used as feature selection. With feature selection methods large data is splinted into small data and important features are detected. People use credit cards for shopping and other payments. But alongside other fields credit card is not secure. Some of the types of credit card frauds are listed below. These types help to understand credit card frauds.

- Without the permission of authorized persons the use of account information.
- The use of a person's account is a criminal act by using a person's information without his permission.
- Skimming the card is also an act of credit card fraud.
- Stealing the card and making a duplicate copy.

In another research the author studied that the growth rate of the use of plastic money has increased more than the economic growth. This paper studied about the innovations timeline to find out overview of credit card frauds. For this purpose of fraud detection, the secondary data is taken from newspapers and journals. Methods to detect frauds adopted by the author are Address Verification Service (AVS). This method tries to match customers ZIP code and address while any transaction is made [6]. A study Proved that there are different types of credit card frauds. How we will be able to detect these frauds is the key point of this study. The methods used by the author are ANN, Decision Tree, Fuzzy Logic, SVM, Bayesian Network, K - NN, Hidden Markov Model and Logistic Regression. The results showed that the neural network and the bayesian network gives the highest level of accuracy that is 99.71% and 97.52 % respectively. K-Nearest Neighbour

(97.15%), SVM (94.65%) and decision tree (97.93%). Fuzzy logic system and logistic regression gives low level of accuracy that is 95.2% and 94.7% respectively as compared to others [20]. A smilar research studied the performance comparison of previously used machine learning techniques. The author has listed the results of all machine learning techniques and datasets. This study identified that NB, KNN and KNN-DRF gives better results [8]. An intelligent and efficient approach for credit card frauds used an optimized light gradient boosting machine. Two real world datasets were used for checking performance. The results showed that ACU is 92.88%, Precision 97.34% and F1-scoe 56.95%. The author claimed that proposed machine learning algorithm outperformed and achieved the highest results in terms of accuracy, precision, AUC and F1-score [1]. Another study showed fraud detection model creation to detect fraud from imbalance and anonymous datasets. The study shows that 40% frauds from all financial frauds are related to credit cards and loss is about 5.55 billion. Methods used by the author are K-Nearest Neighbor and Support Vector Machines (SVM). K-Nearest Neighbour stores all available instances and it checks nearest neighbor. If the transaction is about nearest to fraudulent then it is considered as fraud and if the act of neighbor is genuine, it is considered as genuine transaction. The KNN gives better performance results. Matching Algorithms have used in this paper. Fraud detection model is used to find the behavioral changes. Frauds detection had done by changing the small database system rather larger [13]. Another research solved the problem how meta learning strategy can be used as a tool for detecting frauds? The methods used in this paper are meta learning techniques, selecting base classifiers using diversity metric. Meta-learning stages, ranking evaluation and the meta-classifier can catch more fraudulent transaction at the earlier time. By ranking transactions of both NN scores and MC probability to determine the transactions having greatest fraud risk of 28% improvement to the FI's NN of score ranking method. The saving improvements will help to improve the transactions [4]. A reserch scholar Mohadavesh tried to solve a problem that the Existence credit card fraud detection methods are not efficient enough to handle the frauds. There should be a mechanism to detect frauds. The methods used in this paper by the author are neural networks, genetic algorithm, decision Tree, HMM and K-Means clustering. The results showed that low medium and high rates transactions are used to detect the frauds [5].

By summarizing the studies of different authors, we found that different authors used different methods for credit card fraud detection. The previous studies used KNN, SVM, Genetic Algorithm, ANN, BNN, Neural Networks and many more for credit card fraud detection. But now in this study we used KPCA as feature selection method. Our main objective is to detect credit card frauds using nonlinear feature selection method KPCA with classification methods decision tree and boost. We will try to prove how these methods help to improve results as compared to previous ones. Our research problem will focus on Using KPCA as feature selection method with machine learning techniques to detect credit card frauds. To increase precision rate data mining and classification methods are used.

II. LITERATURE REVIEW

Genetic algorithm are the evolutionary algorithms that are used to check which transaction is fraudulent. 7 steps of Genetic algorithm have used to make a clear transaction without fraud. These steps include one-time password (OTP) in two steps to authenticate that transaction is legal or fraudulent. This will result in reducing loss to the card holder and to the banks as well [17]. Data mining techniques for credit card fraud detection are having better results. These researchers used PCA as feature selection method with classification methods, isolation forest and local outlier factor. The overall accuracy of their study remained 99.6% and precision 33% [14]. Geneting algorithms are evolutionary algorithms. Geneting algorithm help to detect credit card frauds using selected features. An effort was made to detect credit card frauds employing Geneting Algorithm. It is claimed that Geneting Algorithm works better for detecting credit card frauds at different steps rather for secure transactions [3], [5], [11], [10]. Web frauds increasing rate can be detected using Artificial Neural Network and Stimulated Anealing. The results showed that ANN provides better results when used with Annealing algorithms. In future a combination of Genting algorithms along with Simulated Annealing will be used for best configuration [2]. An experiment was employed on artificial neural networks and geneting algorithm collectively. As ANN works like human brain, so a combination of ANN and GA togather help to detect credit card frauds. But the use of ANN is subjected on the ground that it needs some initial data to detect credit card frauds [9]. A model was proposed that allows shopping to valid customers . The methods used by author are Genetic Algorithms using a formula and Neural Network. The results showed that Lost or Stolen card contributes the 48%, identity theft 15%, cloning 14%, counterfeit card 12%, mail intercept fraud 6% and others 5% in total transactions and that cause a great loss [12].

III. METHODOLOGY

Methodology checks out reliability and validity of a particular study. Methodology is a logical way to study things. In this study a hybrid approach is used, that is nonlinear feature selection method KPCA with classification methods decision tree and boost. This study will help to handle the gap in feature selection using two data sets.

Framework: It forms the basis of theoretical work. It is a defined way for doing a task. It helps to complete a task within a specific time.

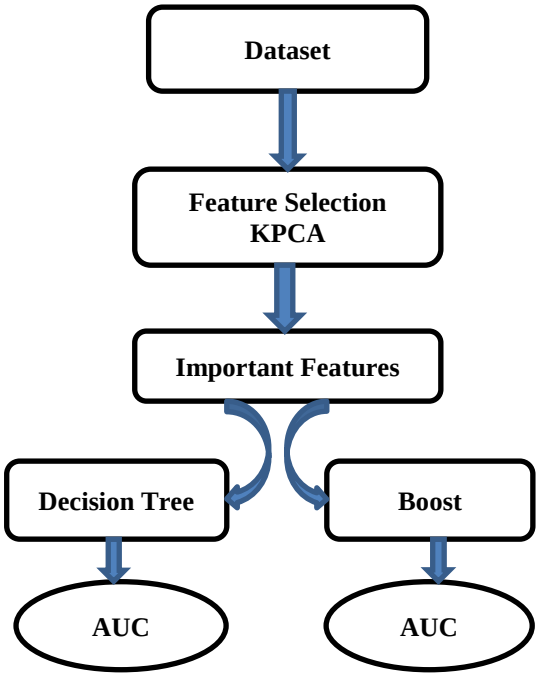

Fig. 1 Framework

Data Description: Collecting data for credit card fraud detection is very difficult because we must collect it from different financial institutes and banks. The first dataset used in this study is taken from Kaggle [16]. Kaggle is a community for data analysis and machine learning. It contains 284807 transactions with total variables 31. Class is target variable 0 denotes for genuine transactions and 1 for fraudulent. Our second dataset is also taken from Kaggle community, with 30000 transactions and 25 variables. Our Target variable is listed below that is default.Payment.next.month [15]. 1 shows fraudulent and 0 genuine transactions. Total no of fraudulent transactions are 6636.

Feature Selection: Feature selection is a method to reduce the no of input variables for machine learning modeling. Feature selection methods evaluate each input variables with target variable. Feature selection evaluates variables and selects important features. It selects only those variables that have more strong relationship with the target variable

KPCA: kernel principal component analysis is an improved and extended form of PCA. PCA is linear form of feature selection method. PCA is used for exploratory data analysis and dimnsionality reduction. But it should be kept in mind that PCA is useable for linear datasets [19]. Sometimes it happens that data is non linear. To make data in linear form is difficult task. KPCA is used for nonlinear datasets because data is not always in linear form. Therefor for nonlinear datasets KPCA is more valid method and provides better results.

Machine Learning: Machine learning methods provide a facillity to work in a natural environment. Machine Learning methods directly use data for computation. Deep learning is a special form of machine learning methods. Machine learning methods deal with more data, do more computation and provide better result. It helps in solving complex situations and large amount of data. Decision Tree and Boost are efficient, simple and better methods in machine learning environment.

Decision Tree: One of the most common data mining models is decision tree. Decision tree is most popular and powerful model for prediction. It is popular because the resulting models are easy to understand. It is flowchart like structure. It automatically handles decision-making and can be trained on a small data set [21].

Boost: The basic idea behind the boosting method is to manage a weighted data with each observation of the given data set. A series of models can be built, and the weights are accelerated if a model is classified incorrectly. The approach of boost ensemble tries to develop a model that shows less variance and bias as compared to decision tree related observation.

ROC: Receiver Operating Characteristic Curve is basically a graph. That depicts classification model performance. Two parameters are plotted with ROC curve. That are;

i. TPR = TP / FP + TN

ii. FPR = FP / FP + TN

Area under the ROC curve (AUC) provides and aggregate measure of performance. The range of AUC curve lies between 0 and 1. Use of AUC is for two purposes. i Scale invarient ii. Classification threshhold invariance. But both are not always required [17].

Precision: Precision is used as a factor of retrieving the documents that are query relevant. Formula to calculate precision is TP/TP+FP.

IV. DATA ANALYSIS & RESULTS

In this study KPCA has used as a feature selection. Selected features have tested using decision tree and boost as sown in figure 1. In this paper three sampling size has been tested. First 70% training 15% validation and 15% testing. Second 70% training 30% testing and third 80% training 20% testing.

Results:

Results Dataset 1

First dataset is taken from Kaggle. We applied KPCA as feature selection method. After selecting important features, we used these features as input in R. Using decision tree and boost as machine learning methods with first dataset the results are shown in table below.

TABLE 1

First Dataset Results

Model/Sampling	AUC	Precision
Decision Tree Testing 70/15/15	0.89	0.99
Decision Tree validation 70/15/15	0.92	0.99
Boost Validation 70/15/15	0.99	0.99
Boost Testing 70/15/15	0.96	1
Decision Tree Testing 70/0/30	0.91	0.99
Boost Testing 70/0/30	0.98	0.99
Decision Tree Testing 80/0/20	0.90	0.99
Boost Testing 80/0/20	0.98	0.99

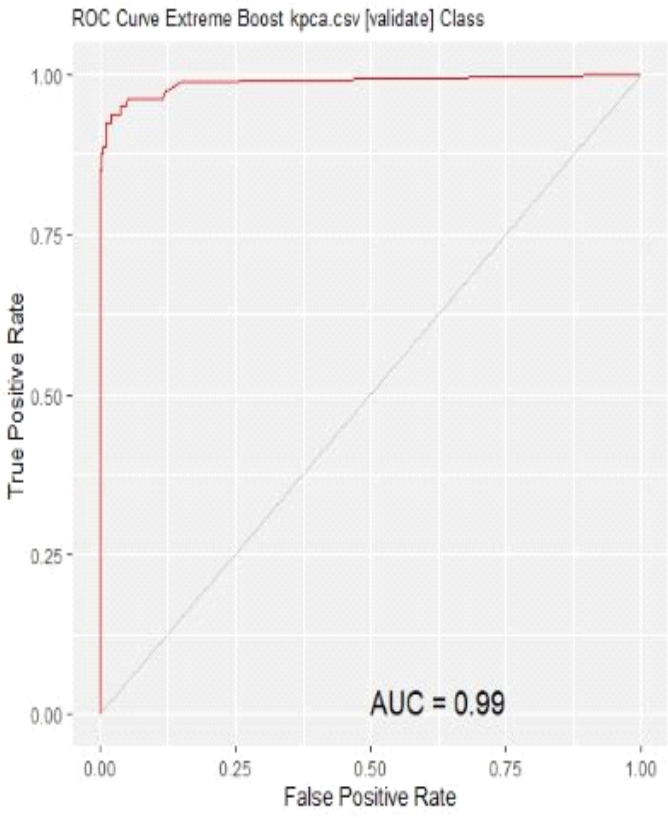

Fig. 2 AUC Boost Validation

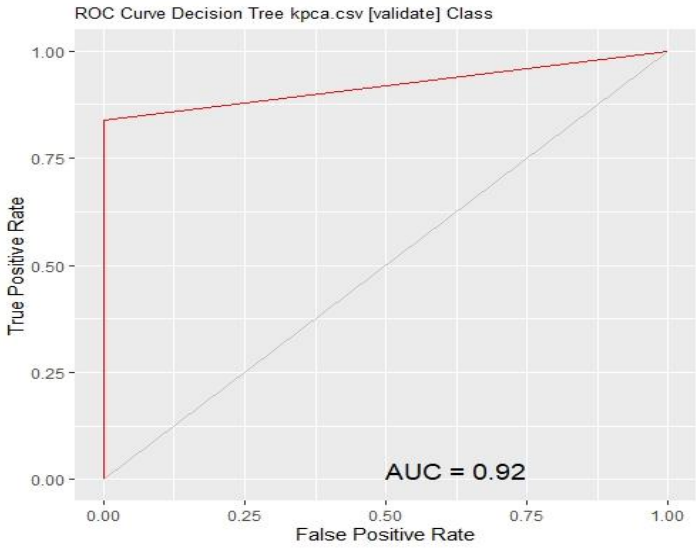

Fig. 3 AUC Decision Tree Validation

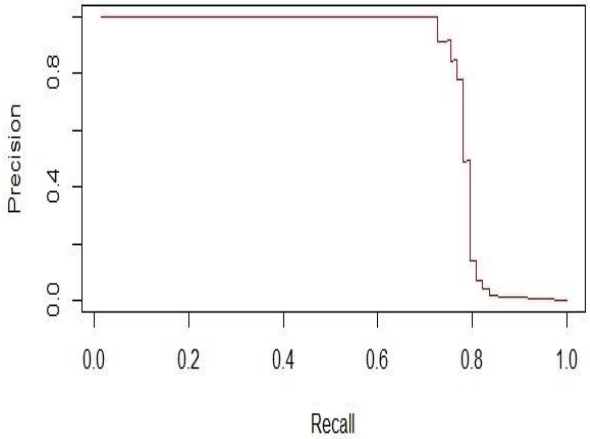

Fig. 4 Precision Boost Testing

Results Dataset 2

Second dataset is also taken from Kaggle. We applied KPCA as feature selection method. After selecting important features, we used these features as input in R. Using decision tree and boost

as machine learning methods with second dataset, the result shows that Decision Tree gives better AUC as compared to Boost that is 0.96 with different sampling size. However AUC is high with Boost validation that is 0.78. The results are shown in table below:

TABLE 2

Second dataset results

Model/Sampling	AUC	Precision
Decision Tree Testing 70/15/15	0.65	0.95
Decision Tree validation 70/15/15	0.65	0.96
Boost Testing 70/15/15	0.77	0.94
Boost validation 70/15/15	0.78	0.95
Decision Tree Testing 70/0/30	0.65	0.96
Boost Testing 70/0/30	0.77	0.94
Decision Tree Testing 80/0/20	0.64	0.96
Boost Testing 80/0/20	0.75	0.92

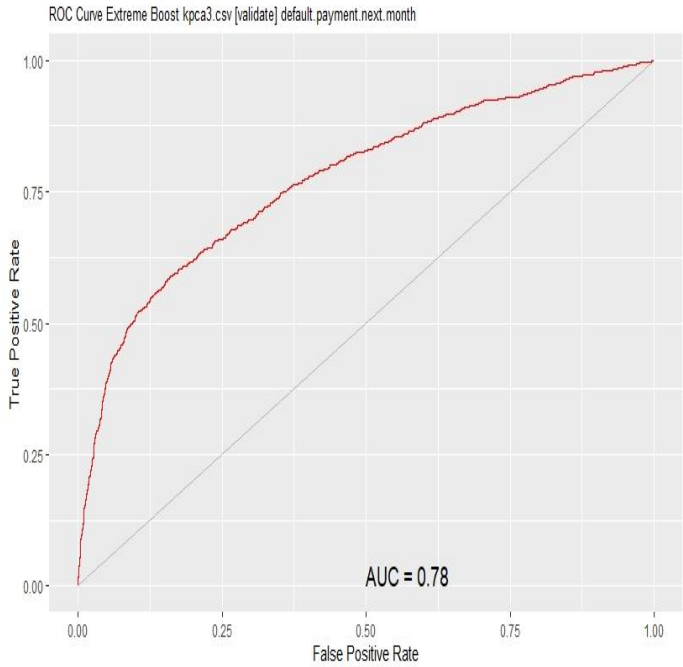

Fig. 5 AUC Boost Validation

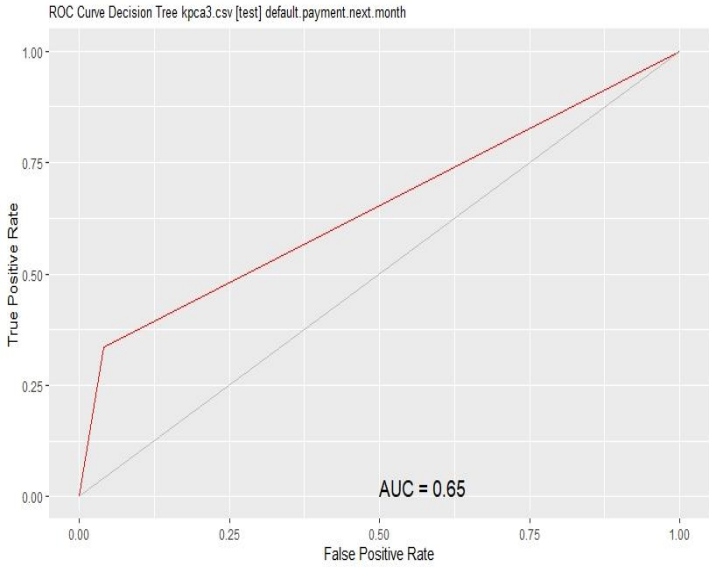

Fig. 6 AUC Decision Tree Test

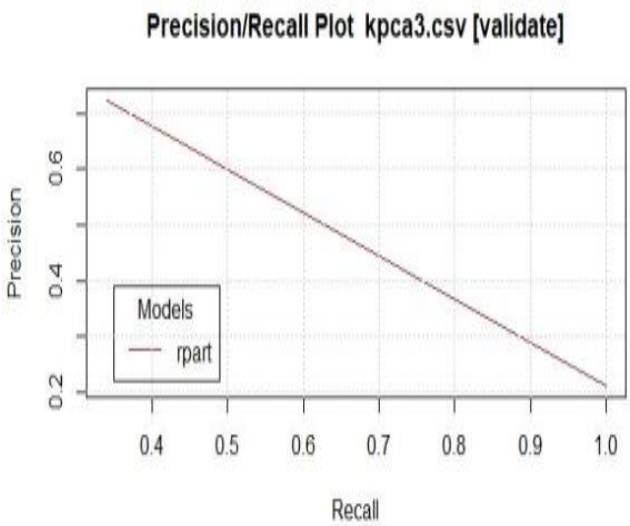

Fig. 7 Precision Decision tree validation

The results with second dataset depicts that decision tree validation gives more precision with 70% training 15% validation 15% testing, 70%

training 30% testing and 80% training 20% testing as compared to boost. However the accuracy is high with boost validation.

The results comparison of both data sets shows that our calculated accuracy is 99% and in previous research it is 99.6%. However, our calculated precision level is high as it is 1. This comparison also shows that Boost method with R gives better result as compared to Decision Tree.

V. DISCUSSION

Frauds in the field of credit card causing financial loss to customers and to the financial institutes as well. We studied how different methods are used by different authors to detect credit card frauds. Different types of frauds are associated with credit card industry like Skimming, stealing card, stealing card owner personal information for online purchases. Our main objective of this study was to use a hybrid model of feature selection and machining learning methods to detect credit card frauds. We tried to employ different machine learning methods to reduce the credit card fraud rates. Different authors till now have used different methods for feature selection and machine learning to detect credit card frauds. In this study we used kernel principal component analysis (KPCA) as feature selection method with classification methods. KPCA is a nonlinear method used for feature selection. On selected features decision tree and boost methods are employed with ROC as evaluation method. The sampling size used are 70% training 15 % validation 15 % testing,70% training 30 % testing and 80% training 20 % testing.

With dataset 1 using sampling 70 % Training 15 % Validation 15% Testing with parameters min split 28, min bucket 8 Decision Tree testing values of AUC is 0.89 and Precision is 0.99. Total calculated time for model is 1.54 mins. Decision tree validation with same sampling, value of AUC is 0.92 and precision is 0.99. Using Same Sampling size with Boost parameters max. Depth 9, learning rate 0.1, values of boost validation are AUC 0.99, Precision 0.99 and the total calculated time is 41.84 secs. Results with boost testing are AUC 0.96 and Precision 1. Using sampling size 70% training and 30% testing with Decision Tree parameters min split 28, min bucket 12 and Boost parameters max. Depth 12, learning rate 0.1.The values of AUC is 0.91 and 0.98, precision 0.99 and 0.99 respectively for both models. Using sampling size 80% training 20% testing with Decision Tree parameters min split 24, min bucket 8 Decision Tree testing values for AUC 0.90, Precision 0.99 and the total calculated time for model is 1.84 mins. Using Same Sampling with parameters max. Depth 17, learning rate 0.5 values of Boost validation are AUC 0.98, Precision 0.99 and the total calculated time is 51.47 secs.

With second dataset using sampling size 70 % training 15 % validation 15% testing with

parameters min split 28, min bucket 8 Decision Tree validation values for AUC 0.65 and Precision 0.96, the total calculated time for model is 3.67 secs. Decision tree testing values for AUC 0.65 and the precision value is 0.95. Using Same Sampling with Boost parameters max. depth 9, learning rate 0.1 values of boost testing are AUC 0.77 and Precision 0.94, total calculated time is 3.71 secs. Results with Boost validation are AUC 0.78 and Precision 0.95. Using sampling size 70% training and 30% testing with parameters values, Decision Tree min split 28, min bucket 12 and Boost parameters max. depth 12, learning rate 0.1.The values of AUC 0.65 and 0.77 , precision 0.96 and 0.94, time calculated is 3.57 secs and 4.83 secs respectively for both models. Using sampling size 80% training and 20% testing with parameters min split 24, min bucket 8, Decision Tree testing values for AUC 0.64 and Precision 0.96 the total calculated time for model is 4.27 secs. Using Same Sampling with Boost parameters max. Depth 17, learning rate 0.5 the values of boost validation are AUC 0.75, Precision 0.92 and the total calculated time is 7.54 secs.

VI. CONCLUSION

It is concluded that Boost testing with sampling size 70% training 15 % validation 15 % testing gives highest accuracy 0.99. Boost testing with sampling size 70% training 15 % validation 15 % testing gives highest precision that is 1 with dataset 1. Overall Performance of Boost is better than Decision Tree. The comparison of our results with past work showed that our precision level is high that is 1. In future the results can be improved and credit card fraud can be detected with other feature selection methods that have not tested earlier using different machine learning evaluation methods and tools.

REFERENCES

[1] MALEBRY, A. A. (2020). An Intelligent approach to Credit Card Fraud *IEEE, 9.*

[2] Azeem Ush Shan Khan, N. A. (2014). Real-Time Credit-Card Fraud Detection using Artificial Neural Network Tuned by Simulated Annealing Algorithm. *Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing, ITC*, 9.

[3] Ishu Trivedi1, M. M. (2016). Credit Card Fraud Detectionz. *International Journal of Advanced Research in Computer and Communication Engineering*, 4.

[4] Joseph Pun, Y. L. (2012). Improving Credit Card Fraud Detection using a MetaClassification Strategy. *nternational Journal of Computer Applications (0975 – 8887)*, 6.

[5] K.RamaKalyani, D. (2012). Fraud Detection of Credit Card Payment System by Genetic Algorithm. *International Journal of Scientific & Engineering Research Volume 3*, 6.

[6] MR.K.KATHIRVEL. (2013). CREDIT CARD FRAUDS AND MEASURES TO DETECT. *International Journal of Marketing, Financial Services & Management ResearchInternational Journal of Marketing, Financial Services & Management ResearchInternational Journal of Marketing, Financial Services & Management Research*, 8.

[7] Oberoi, R. (2017). Credit – Card Fraud Detection System: Using Genetic Algorithm. *International Journal of Computer & Mathematical Sciences*, 5.

[8] A. Singh, A. J. (2019). an Emprical study Of AML Approach For Credit Card Fraud Detection-Financial Transactins. *INTERNATIONAL JOURNAL OF COMPUTERSCOMMUNICATIONS & CONTROL*, 21.

[9] Raghavendra Patidar, L. S. (2011). Credit Card Fraud Detection. *International Journal of Soft Computing and Engineering (IJSCE)v*, 7.

[10] Rinky D. Patel, D. K. (2013). Credit Card Fraud Detection & Prevention of Fraud Using Genetic Algorithm. *International Journal of Soft Computing and Engineering (IJSCE)*, 3.

[11] SATVIK VATS*, S. K. (2013). Genetic algorithms for credit card fraud detection. *International Conference on Education and Educational Technologies*, 12.

[12] Syed Ahsan Shabbir, K. R. (2013). An Effective Fraud Detection System Using Mining Technique. *International Journal of Scientific and Research Publications, Volume 3, Issue 5, May 2013* , 4.

[13] Zareapoor, K. R. (2014). FraudMiner: A Novel Credit Card Fraud Detection Model. *Hindawi Publishing Corporation*, 116.

[14] S P Maniraj, A. S. (2019). Credit Card Fraud Detection using Machine Learning and Data Science. *International Journal of Engineering Research & Technology (IJERT)*, 6.

[15]Lichman,M.(2013).https://www.kaggle.com/murlidhar86/credit-card-clients-dataset/log?

[16] ULB, M. L. (2016). https://www.kaggle.com/mlg-ulb/creditcardfraud/metadata.

[17]https://developers.google.com/machine-learning/crash-Course/classification/roc-and-auc

[18] https://en.wikipedia.org/wiki/Diners_Club_International

[19] https://en.wikipedia.org/wiki/Principal_component_analysis

[20] Yashvi Jain, N. S. (2019). A Comparative Analysis Of Various Credit Card Fraud Detection Techniques . *International Journal of Recent Technology and Engineering (IJRTE)*, 7.

[21]https://www.geeksforgeeks.org/decision-tree-classifiers-in-r-Programming/? ref=rp

[22] C. W. (2009). Transaction aggregation as a strategy for credit card fraud detetction. 26.

[23] Akhilomen, J. (2013). Data Mining Application for Cyber Credit-card Fraud Detection System. *Proceedings of the World Congress on Engineering 2013 Vol III,*, 6.

[24] Brabazon, A., Cahill, J., Keenan, P., & Walsh, D. (2010). Identifying online credit card fraud using artificial immune systems. *IEEE Press*, 7.

[25] Chan, P. K. (n.d.). Distributed Data Mining in Credit Card F raud Detection. *Submitted to IEEE Intel ligent Systems' Sp e cial Issue on Data Mining, 1999*, 17.

[26] Deng, Y. (2007). Application of Classification Models on Credit Card Fraud Detection. 5.

[27] Duman2, Y. S. (2011). Detecting Credit Card Fraud by ANN and Logistic Regression. *Department of Industrial Eng., Dog s University, Acibadem, Istanbul, Turkey*, 9.

[28] Ekrem Duman a, ⇑. M. (2011). Detecting credit card fraud by genetic algorithm and scatter search. 8.

[29] Employing transaction aggregation strategy to detect credit card fraud. (2012). *Elsevier* , 8.

[30] Khyati Chaudhary, J. Y. (2012). A review of Fraud Detection Techniques: Credit Card. *International Journal of Computer Applications (0975 – 8887)*, 6.

[31] Manoel Fernando Alonso Gadi1, 2. X. (2008). Credit Card Fraud Detection with Artificial Immune System. 12.

[32] MohdAvesh Zubair Khan1, J. D. (2014). Credit Card Fraud Detection System Using Hidden Markov Model and K-Clustering. *International Journal of Advanced Research in Computer and Communication Engineering*, 4.

[33] MR.K.KATHIRVEL. (2013). CREDIT CARD FRAUDS AND MEASURES TO DETECT. *International Journal of Marketing, Financial Services & Management ResearchInternational Journal of Marketing, Financial Services & Management ResearchInternational Journal of Marketing, Financial Services & Management Research*, 8.

[34] N.Sivakumar#1, D. (2015). Fraud Detection in Credit Card Transactions: Classification, Risks and Prevention Techniques. *N.Sivakumar et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1379-1386*, 8.

[35] Neda Soltani Halvaiee ∗, M. K. (2014). A novel model for credit card fraud detection using Artificial Immune Systems. 10.

