International Journal of Electronics Communication and Computer Engineering
Volume 10, Issue 6, ISSN (Online): 2249–071X

Performance Evaluation of Credit Card Fraud Transactions using Boosting Algorithms

Kavya Divakar [1*] and Chitharanjan K. [2]
[1] Department of Computer Science, SCT College of Engineering, Kerala, India.
[2] Assistant Professor, Department of Computer Science, SCT College of Engineering, Kerala, India.
*Corresponding author email id: kavyadivakar95@gmail.com

Date of publication (dd/mm/yyyy): 19/12/2019

***Abstract* – In the era of digital world, internet has reached a global connectivity. The whole world has transformed into digital now. All the firms whether it is educational organizations, governmental organizations, shopping, businesses etc have turned into a digital format. With the increasing trend in online marketing, the credit card companies have rapidly expanded. Due to this makeover, fraudulent cases have started to grow up. Analyzing fraudulent transactions manually is time consuming and tedious. Hence, with the advent of internet technology like artificial intelligence, machine learning etc, it is possible to detect and predict the chances of fraudulent actions. To evaluate the model, a publicly available credit card dataset is used. By implementing traditional machine learning algorithms like naive bayes classifier, decision tree classifier, etc, the classifier which performs poorly is found out. Boosting algorithms AdaBoost, Gradient Boost and XGBoost are implemented to find out the one which performs more accurately and precisely to predict the fraudulent cases. By comparing the results, it was found out that XGBoost performs better.**

***Keywords* – AdaBoost, Decision Tree Classifier, Gradient Boost, XGBoost.**

I. INTRODUCTION

Credit card fraud is considered to be one of the serious issues in financial domain [1]. Rapid growth in online marketing has increased day by day. People mostly prefer to use online mode of payment now [3]. As a result, fraudsters too have started to increase in number. Credit cards can be typically classified into physical and virtual. In the physical credit card, the user has to show up the details while making payment and hence the fraudster has to steal the card in order to access their data. In virtual credit card, the fraudsters can access the cardholder's providential data through various methods and hence is prone to attack. Fraudsters keep their identity and location hidden. Hence, there is a need for a proper fraud detection mechanism to reduce or eliminate fraud cases.

Fraud detection in online marketing is tedious and creates a huge overhead. As technology develops, more and more new inventions increase and the whole world get transformed to a digitalized format [2]. Even though we live in a digital world where everything is at our hand tip, many illiterate people are still unknown to these technologies. These illiterate people are the great victims to these fraud attacks. Due to demonetization, everyone has started to depend upon various financial cards like credit card, debit card etc [9]. The crucial information of every individual is being getting dispersed across the globe in digital languages and hence they are threatened to various attacks. As a result, it is a necessity to develop technologies which could prevent the fraud attacks happening thus by preventing the leak of big data to the underworld.

In the current scenario, technologies like machine learning, deep learning and artificial intelligence is at a peak. Machine learning is an application of artificial intelligence which includes various means of learning. The learning can be supervised learning, unsupervised learning and semi-supervised learning. In supervised learning, the labels are assigned to each data and hence easy to predict them [9]. In unsupervised learning, no labels are predefined and hence require computations to properly assign labels for each data. In semi- supervised learning, few data are

assigned labels and others are not assigned and hence require some complex computations. Most of the real time applications are based on unsupervised learning models. Commonly used machine learning algorithms include Naive Bayesian classifier, Decision Tree Classifier, Linear Regression, Logistic Regression, etc.

In this paper, publicly available credit card dataset is taken. Initially some machine learning algorithms like Naïve Bayesian Classifier, Decision Tree Classifier etc. are used to model the dataset to identify the weak learner. On identifying it, boosting algorithms like AdaBoost, Gradient Boost and XGBoost are used to model the data by a number of iterations of weak learner, thus by producing a strong learner to identify the fraud attacks.

II. Background Study

Hybrid model based credit card fraud detection is proposed in [1]. Initially standard machine learning algorithms like Naive Bayes, Decision Tree, and Random Forest etc. were used to model the dataset. The one which showed the least accuracy and precision was chosen as the weak learner. Later hybrid models like AdaBoost and Majority voting were used to model the dataset by iterating the weak learner a definite number of times to become a strong learner. Thus the hybrid model is used to accurately predict the fraud transactions.

A survey on credit card fraud detection using various machine learning models is proposed in [2]. A study on six data mining approaches like classification, clustering, prediction, outlier detection, regression and visualization is done. There are many existing techniques for fraud detection based on Artificial Immune System, Logistic Regression, Decision Tree Classifier, Genetic Algorithm, Bayesian Belief Network, Neural Network, Support Vector Machine etc. Research on each of these methods is still on a progress.

An ensemble learning framework is proposed in [3] based on training set partitioning and clustering. They have ensured the integrity of sample features. The dataset is partitioned into training and testing set and the training set is further divided into majority and minority samples. Using random sampling technique they have balanced the dataset by selecting equal number of minority and majority class samples in the training set. Now each base classifier is trained with each balanced training set and every estimator votes to get the final ensemble model based on majority voting rule.

Credit card fraud detection using collating machine learning models is proposed in [4]. Credit card dataset is performed data preprocessing by cleaning the data by removing redundancy, filling empty spaces etc. K-Fold Cross validation is performed to split the dataset into training and testing one. Models are created for Logistic Regression, Decision Tree Classifier and SVM and also metrics like accuracy, precision etc is computed and finally a comparison is made.

Machine Learning based approach to detect financial fraud in mobile payment system is proposed in [5]. An actual mobile payment data is taken for the study. The preprocessed data is performed feature selection by filter based techniques. SMOTE oversampling technique is used to balance the dataset and then both supervised and unsupervised machine learning algorithms are used to model the dataset.

III. Materials and Methods

A. *Data Collection*

A real world credit card dataset is collected from Kaggle website. The dataset contains transactions of European cardholders. There are a total of 2, 84,807 transactions, out of which 492 are fraud ones. It contains 31 features

out of which 28 numerical input variables are obtained by the transformation of PCA method and two features Time and Amount that is not transformed. The last feature Class is the response variable and it takes value 1 as fraud and value 0 as non-fraud.

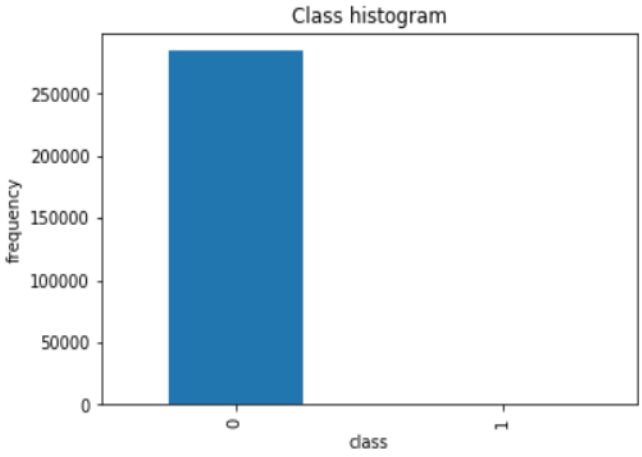

Fig. 1. Imbalanced credit card dataset.

From the fig.1, it is understood that the dataset is highly imbalanced because majority class instances (non-fraud) is high in number than minority class (fraud). Hence the dataset is needed to be balanced.

B. *Data Preprocessing*

In preprocessing phase, to find out the feature importance, we used Extra Trees Classifier (ET) technique. ET is an extremely randomized tree classifier. For each feature, gini index is computed. After computation of gini index of all features, they are sorted in descending order and the top 11 features are selected out of 30 ones (31st is the Class feature).

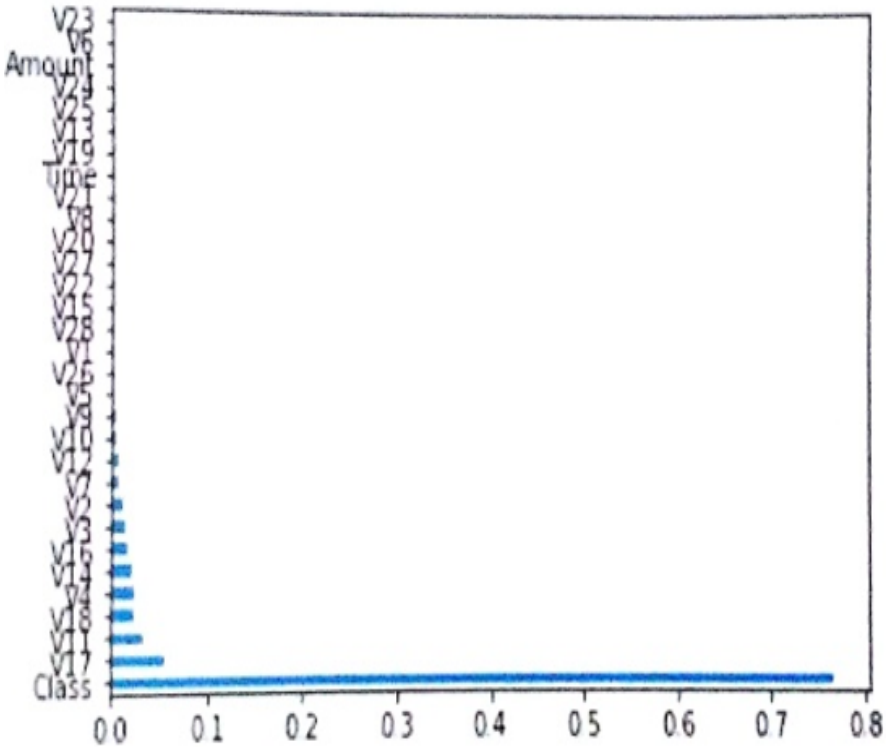

Fig. 2. Feature selection.

From the fig. 2, it is found out that only 11 features actually contribute to the target variable Class. Hence all the remaining features are dropped.

C. *Dataset Partitioning*

K-Fold Cross Validation (CV) technique is used to partition the dataset into training and testing ones. The parameter k specifies the number of groups that the given data sample has to be split. CV shuffles the dataset randomly and split the dataset into k groups. For each unique group it compute the mean squared error and the group which shows the least mean squared error is chosen.

D. *Data Sampling*

The dataset is highly imbalanced as already mentioned in fig. 1. One of the simplest strategies to handle the imbalanced dataset with majority class instances more is to under-sample the dataset. Random Under-sampling technique is used here. Randomly eliminating instances from the majority class and assigning it to the minority class is known as random under-sampling. The void that is getting created is called random.

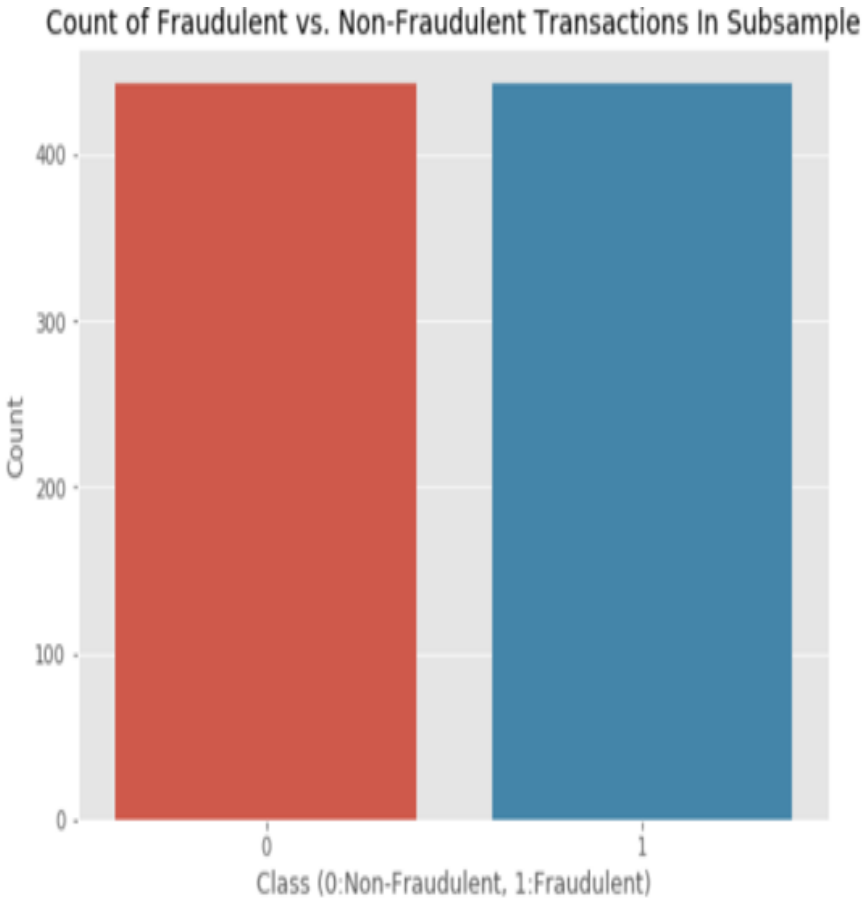

Fig. 3. Count of transactions after sampling.

From fig. 3, it is understood that both the classes, class 0 (non-fraud) and class 1 (fraud) are equally sampled. Hence the dataset is now balanced.

E. *Standard Machine Learning Models*

Initially, four standard machine learning models Random Forest Classifier, Naïve Bayesian Classifier, Logistic Regression and Decision Tree Classifier is used to model the dataset.

E.1 Naive Bayes Classifier

Naive Bayes uses Bayes theorem for classification. It is assumed that certain features are not correlated to others. Only a small training dataset is needed for classification.

===== Naive Baiye Classifier =====

Cross Validation Mean Score: 99.8%

Model Accuracy: 95.6%

Confusion Matrix:

[[190378 8642]

[60 284]]

Classification Report:

	Precision	Recall	f_1 - Score	Support
0	1.00	0.96	0.98	199020
1	0.03	0.83	0.06	344

Fig. 4. Naive Bayes Classifier Performance.

From fig.4, it can be referred that the accuracy of Naive Bayes Classifier is 95.6%.

E.2 Decision Tree Classifier

Decision Tree is a collection of nodes that creates decision on features. Every node represents a split rule for a feature. New features are established until the stopping criterion is met. The leaf nodes specify the class to which a particular feature belongs to.

===== Decision Tree Classifier =====

Cross Validation Mean Score: 99.9%

Model Accuracy: 90.0%

Confusion Matrix:

[[179160 19860]

[0 344]]

Classification Report:

	Precision	Recall	f_1 - Score	Support
0	1.00	0.90	0.95	199020
1	0.02	1.00	0.03	344

Fig. 5. Decision Tree Classifier Performance.

From fig. 5, it can be found out that the Decision Tree Classifier is having an accuracy of 90%.

E.3 Random Forest Classifier

Random Tree operates as a Decision Tree operator with the exception that only a random subset of feature is available. RF creates an ensemble of random trees. The user can set the number of trees. The resulting model employs voting of all created trees to determine the final classification outcome.

===== Random Forest Classifier =====

Cross Validation Mean Score: 99.9%

Model Accuracy: 97.7%

Confusion Matrix:

[[194452 4568]

[10 334]]

Classification Report:

	Precision	Recall	f_1 - Score	Support
0	1.00	0.98	0.99	199020
1	0.07	0.97	0.13	344

Fig. 6. Random Forest Classifier Performance

From fig.6, it can be referred that the Random Forest Classifier is having an accuracy of 97.7%.

E.4 Logistic Regression

LR can handle data with both nominal and numerical features. It estimates the probability based on one or more predictor features.

===== Logistic Regression =====

Cross Validation Mean Score: 99.9%

Model Accuracy: 98.3%

Confusion Matrix:

[[195750 3270]

[40 304]]

Classification Report:

	Precision	Recall	f_1 - Score	Support
0	1.00	0.98	0.99	199020
1	0.09	0.88	0.16	344

Fig. 7. Logistic Regression Classifier Performance.

From fig. 7, it is understood that the Logistic Regression classifier is having the highest accuracy of 98.3% compared to other standard models. Hence Decision Tree Classifier is chosen as the weak learner since it has the least accuracy of 90%.

F. *Boosting Algorithms*

Boosting is a method of converting a set of weak learners into strong learners. A weak learner has an error rate lesser than 0.5 and strong learner has an error rate closer to 0. A family of weak learners combined together to form a strong learner. Three boosting algorithms are used: AdaBoost, Gradient Boost and XGBoost.

F.1 AdaBoost

In AdaBoost, the base learner is a weak learner upon which the boosting method is applied. The training data is randomly sampled and decision stump algorithm is applied to classify the points. After the classification, the

decision tree stump is fit to the complete training data. This iterates until the training data fits without any error or a specified number of estimators is reached.

===== Adaboost =====

Cross Validation Mean Score: 99.9%

Model Accuracy: 99.9%

Confusion Matrix:

[[198979 41]

[106 238]]

Classification Report:

	Precision	Recall	f_1 - Score	Support
0	1.00	1.00	1.00	199020
1	0.85	0.69	0.76	344

Fig. 8. AdaBoost algorithm performance.

From fig. 5, it is understood that AdaBoost algorithm is having an accuracy of 99.9%.

F.2 Gradient Boost

In gradient boosting, many models are trained sequentially. Each new model gradually minimizes the loss function using gradient descent method. It consecutively fit new models to provide a more accurate estimate of the response variable.

===== Gradient Boost =====

Cross Validation Mean Score: 99.9%

Model Accuracy: 99.9%

Confusion Matrix:

[[198942 78]

[96 248]]

Classification Report:

	Precision	Recall	f_1 - Score	Support
0	1.00	1.00	1.00	199020
1	0.76	0.72	0.74	344

Fig. 9. Gradient Boost performance.

From fig.6, it is understood that Gradient Boost algorithm is having an accuracy of 99.9%.

F.3 XGBoost

XGBoost classifier is an Extreme Gradient Boosting. In XGBoost, the trees have a varying number of terminal nodes and left weights of the trees shrunk more heavily. The extra randomization parameter is used to reduce the correlation between the trees. Lesser the correlation among classifiers, better the ensemble model.

===== XGBoost =====

Cross Validation Mean Score: 99.9%

Model Accuracy: 100.0%

Confusion Matrix:

[[199001 19]

[57 287]]

Classification Report:

	Precision	Recall	f_1-Score	Support
0	1.00	1.00	1.00	199020
1	0.94	0.83	0.88	344

Fig. 10. XGBoost performance.

From fig.7, it is understood that XGBoost is having an accuracy of 100%.

IV. Results and Discussion

The study used standard machine learning algorithms Naïve Bayes classifier, Decision Tree classifier, Random Forest Classifier and Logistic Regression **to** find out the weak learner. From fig.5, it is found out that decision tree classifier is having the least accuracy and hence is chosen as the weak learner. By selecting decision tree classifier as the weak learner, three boosting algorithms AdaBoost, Gradient Boost and XGBoost modeled the dataset accordingly as referred in fig. 8, fig. 9 and fig. 10. By comparing the accuracy of these boosting algorithms, it can be found out that XGBoost algorithm is the most accurate and precise in predicting the fraud transactions.

V. Conclusion

In this paper, the study proved the role of boosting algorithms in machine learning prediction models. AdaBoost, Gradient Boost and XGBoost algorithms are the hybrid models which help to improve the accuracy of weak machine learning algorithms. Initially standard machine learning models were used to find out the weak learner. From fig.4, it can be found out that decision tree classifier is having the least accuracy of 90% and hence it was selected as the weak learner. Then hybrid models were created out of three boosting algorithms. From fig.5, AdaBoost classifier shows an accuracy of 99.9% with f1-score of 0.76, from fig.6, Gradient Boost classifier shows an accuracy of 99.9% with f1-score of 0.74(fraud) and in fig.7, XGBoost classifier shows an accuracy of 100% with an f1-score of 0.88(fraud). Out of this, Xgboost algorithm performed better in accurately and precisely predicting the model. Hence XGBoost algorithm is considered to be the best boosting algorithm in predicting models.

References

[1] Kuldeep Randhawa, Chu Kiong, Manjeevan Seera, Chee Peng Lim, Asok K Nandi, Credit Card Fraud Detection Using AdaBoost and Majority Voting, IEEE Access, Vol. 6, Feb 2018, 77-84.
[2] Rimpal R Popat, Jayesh Chaudhary, A Survey on Credit Card Fraud Detection using Machine Learning, IEEE 2nd International Conference on Trends in Electronics and Informatics, Dec. 2018.
[3] Hongyu Wang, Ping Zhu, Xueqiang Zou, Sujuan Qin, An Ensemble Learning Framework for credit Card Fraud Detection Based on Training Set Partitioning and Clustering, IEEE Conference on SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP /SCI, Dec 2018
[4] 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, Vol.118, No. 20, 2018, 825-838.
[5] Dahee Choi, Kyungho Lee, Machine Learning Based Approach to Financial Fraud Detection Process in Mobile Payment System, IT Convergence Practice (INPRA), Vol. 5, No. 4, Dec 2017, 12-24.

[6] Hojin Lee, Dahee Choi, Habin Yim, Eunyoung Choi, Woong Go, Taejin Lee, Insuk Kim, Kyungho Lee, Feature Selection Practice for Unsupervised Learning of Credit Card Fraud Detection, Journal of Theoretical and Applied Information Technology, Vol. 96, No. 2, Jan. 2018.
[7] Joseph Gualdoni, Audrew Kurtz, Ilva Myzyri, Megan Wheeler, Syed Rizvi, Secure Online Transaction Algorithm: Securing Online Transaction using Two-Factor Authentication, Elsevier Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS, Nov. 2017, 93-99
[8] Deepak Pawar, Swapnil Rabse, Sameer Paradkar, Naina Kaushik, Detection of Fraud in Online Credit Card Transactions, International Journal of Technical Research and Applications, Vol. 4, Issue. 2, Apr. 2016, 321-323.
[9] R. Mallika, Fraud Detection using Supervised Learning Algorithms, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6, Issue 6, June 2017.
[10] Jia Song, Xianglin Huang, Sijun Qin, Qing Song, A Bi-directional Sampling based on K-Means for Imbalance Text Classification, IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 25 August 2016.
[11] Ramyashree. K, Janaki K, Keerthana. S, B.V. Harshitha, Harshitha. Y.V., A Hybrid Method for Credit Card Fraud Detection Using Machine Learning Algorithm, International Journal of Recent Technology and Engineering (IJRTE), Volume-7, Issue-6S4, April 2019.
[12] Masoumeh Zareapoor, Pourya Shamsolmoali, Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier, International Conference on Intelligent Computing, Communication & Convergence, (2015) 679 – 686.
[13] SHAILESH S. DHOK, Credit Card Fraud Detection using Hidden Markov Model, International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-1, March 2012.
[14] Ong Shu Yee, Saravanan Sagadevan, Nurul Hashimah Ahamed Hassain Malim, Credit Card Fraud Detection using Machine Learning as Data Mining Technique, Journal of Telecommunication, Electronic and Computer Engineering, Vol. 10 No. 1-4, August 2018.
[15] Lakshmi S V S S ,Selvani Deepthi Kavila, Machine Learning For Credit Card Fraud Detection System, Machine Learning for Credit Card Fraud Detection System, Volume 13, Number 24 (2018), 16819-16824.
[16] Hardik Manek, Sujai Jain, Nikhil Kataria, Chitra Bhole, Credit Card Fraud Detection using Machine Learning, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 8, Issue 4, April 2019.

AUTHOR'S PROFILE

First Author
Kavya Divakar graduated her B. Tech from Government Engineering College, Thiruvananthapuram, Kerala in the year 2018 on Information Technology. Currently pursuing her M. Tech in the department of Computer Science and Engineering from SCT College of Engineering, Thiruvananthapuram, Kerala. Her area of interest includes Machine Learning, Deep Learning and Internet of Things.

Second Author
Chitharanjan K. graduated his M. Tech from National Institute of Technology, Surathkal, Karnataka in the year 2006 on System Analysis and Computer Applications. Also got his B. Tech from Cochin University. Kerala in the year 1999 on Computer Engineering. Currently pursuing Ph.d in the Department of Information Technology, Madras Institute of Technology, Anna University, Chennai, India.. His area of interest includes Cloud Computing in particular Resource Allocation.