Turkish Journal of Computer and Mathematics Education Vol.12 No.12 (2021), 3271-3278

Research Article

Recognizing Credit Card Fraud Using Machine Learning Methods

AppalaSrinuvasuMuttipati[a], SangeetaViswanadham[b], RadhikaSenapathi[c], K. N. BrahmajiRao[d]

[a,d]Associate Professor, Raghu Institute of Technology, Vishakhapatnam, Andhra Pradesh, India.
[b]Professor, Raghu Institute of Technology, Vishakhapatnam, Andhra Pradesh, India.
[c]Assistant Professor, Raghu Institute of Technology, Vishakhapatnam, Andhra Pradesh, India.

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract: The rapid growth of e-commerce and online-based payment possibilities carries along with an empirical universe of economic fraud in which credit card fraud is more preventing for many years; several researchers have developed many data mining base methods to overcome this problem. Recently there has been a major interest in applying machine learning algorithms in place of data mining techniques to discover credit card fraud. Continuous work is done to bring in a conceptual difference between fraud recognition and forecasting probable non genuine opportunities in the digital space of financial transactions. This research shows several algorithms that can be used to solve credit card fraud, however, a number of challenges appear, such as dataset imbalance, variant fraudulent behaviour, etc. In this paper, we are going to address the problem of an imbalanced dataset. SMOTE sampling technique is used to convert the imbalanced dataset to a balanced binary dataset. There are now various machine learning algorithms that are tested by using the European credit card dataset and are compared by using evaluation metrics like accuracy, precession, AUC value, and ROC curve, etc. The results are very encouraging and graphs have been plotted which identify the best suitable algorithm.

Keywords: Machine Learning, Credit card fraud, XGBoost, Random forest classifier

1. Introduction

Now a days, society is growing globally in all areas one of the areas is e-commerce. Due to the increase in e-commerce possibilities in making online payments and they are easier to use, e-commerce business gained user confidence increased the number of users and online transactions have given a drastic rise in revenue generation. Increase in the user's revenue generation has placed a path to be vulnerable to fraudulent behavior.

Credit card fraud is one of the ad-aimed difficulties in the present world. In 2016, there happened to be a benchmark increase in credit card fraud up to 92% compared to the 2012 count. The credit card may happen in one of the following ways. 1) Application fraud, 2) Stolen or lost cards 3) Account taken over 4) the counterfeit card. The stolen or lost card and account takeover are major problems and are named as Card Not Present (CNP) fraud. In CNP, the cardholder is cheated by theft the card sensitive information like CVV, card No, etc, and using it remotely. It leads to the transfer of a large amount or the purchase of costly items before the cardholder discovers.

As the availability of Internet is increasing in the world, most of the people are willing to purchasing things online rather than offline. Due to this, the growth of e-commerce sites is increasing, and thereby the chance of credit card fraud (Sailusha, R. et al., 2020). To solve credit card fraud, we have to find out algorithms that may either avoid or reduce credit card fraud.

This paper introduces a technique for preprocessing an imbalanced dataset to produce balance using SMOTE technique. Thinking about given realities, the Authors of this paper obvious to compare the correctness of Random forest, AdaBoost classifier, CatBoost, and XGboost for recognition of credit card fraud. To accomplish that, an examination was showed.

The remaining paper is organized as introduction in a section I, where section II discusses related work. In section III, presents proposed methodology in three steps. In section IV, the experiments are led among the various machine learning

models and Section V discuss the results and discussion. The metrics correctly classified and AUC ROC value shows the dominance of our model. Finally, section VI gives a conclusion of this work.

2.Related Work

(Hema Gonaboina & Appala Srinuvasu Muttipati, 2021) have suggested some ensemble models for detecting credit card fraud. Models like random forest, logistic regression, Catboost have shown better results. The results when compared, Random Forest and Catboost have outperformed and could create ROC curve and Area under curve.

(Varmedja et al., 2019)(Sailusha et al., 2020)(Awoyemi et al., 2017)have done performance comparison of Naive Bayes, K-nearest neighbor and Logistic regression models in the binary classification of imbalanced credit card fraud. KNN has outperformed the competition based on all of the evaluation metrics.

To identify fraudulent transactions in European credit card data, traditional algorithms such as Decision Tree, Support Vector Machine (SVM) (Godi et al., 2020)**,** Least Square Regression, Naive Bayes Classifier, K-Nearest Neighbors (KNN), and Gradient Boosting (GB) have proven useful. KNN and outlier detection approaches were suggested by(Malini& Pushpa, 2017)**,** are effective in fraud detection. They can help reduce false alarm rates and improve fraud detection rates. In an experiment the author has tested and compared the KNN algorithm with other classical algorithms and KNN performed well.

(Dhankhad et al., 2018)(Dornadula & Geetha, 2019)Random forest uses random tree-based and CART-based methods to train the behavioral features of standard and nonstandard transactions. Despite the fact that random forest obtained results on a small data set, it faces the issue of unbalanced data. The focus of future work will be on resolving datasets that are unbalanced.

3. Proposed Methodology

The proposed approach uses a three step procedure which are stated below-

Step 1: Attaining the dataset from repository

Step 2: Sampling technique is used to convert from imbalanced to balance the dataset

Step 3: The Creation of Machine Learning Models

The flowchart in the figure describes the approach

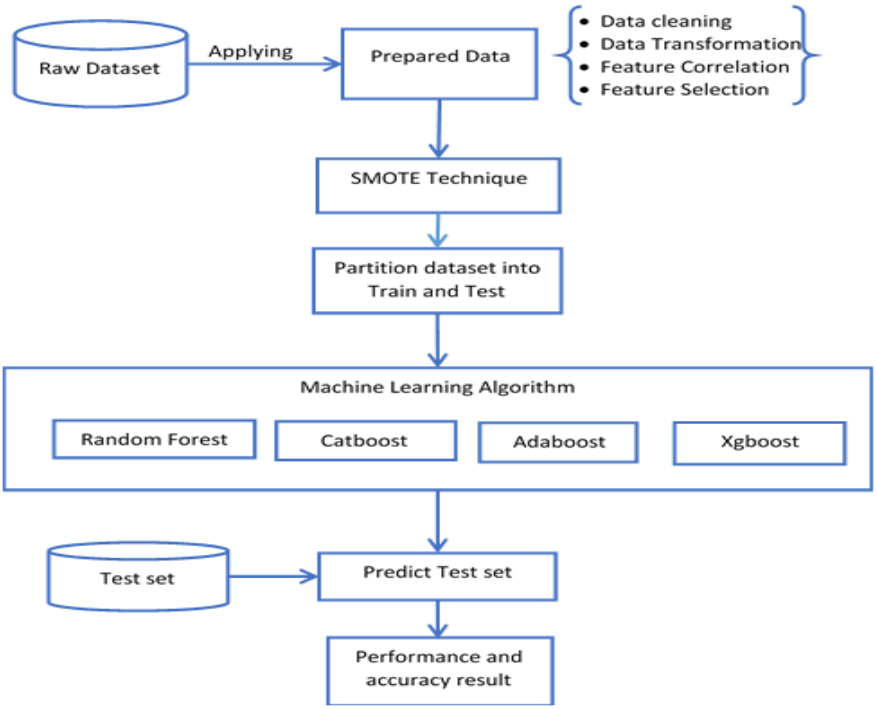

Figure 1: System Architecture

1. Attaining the dataset from repository:

The dataset is organized based on the sequential transactions executed using credit card by European credit cardholder. The dataset encloses a total of 284315 transactions and is a complex dataset which containing 30 variables like, difference between transaction times, transaction amount etc. It also contains 28 other attributes which are kept anonymous in order to protect the identity of the customer. It also contains a column with binary values '0' directs non-fraudulent transaction and '1' directs fraudulent transactions.

One thing we can observe in the dataset is, it is highly skewed. It is because the dataset is sway towards the genuine class. We can observe this, as out of the 284315 transactions only 492 are not genuine. So, only 0.172% fraudulent transactions are present when compared to whole number of transactions.

2. Balancing the dataset using a Sampling technique

3. The credit card dataset which we have chosen for reference dataset consists of an error of 0.172 percent. It gives a meaning that; the referenced dataset contains 0.172 percentages of no genuine transactions. This infers that, the dataset is uneven and is prejudiced towards Attaining the dataset from repository:

The dataset is organized based on the sequential transactions executed using credit card by European credit cardholder. The dataset encloses a total of 284315 transactions and is a complex dataset which containing 30 variables like, difference between transaction times, transaction amount etc. It also contains 28 other attributes which are kept anonymous in order to protect the identity of the customer. It also contains a column with binary values '0' directs non-fraudulent transaction and '1' directs fraudulent transactions.

One thing we can observe in the dataset is, it is highly skewed. It is because the dataset is sway towards the genuine class. We can observe this, as out of the 284315 transactions only 492 are not genuine. So, only 0.172% fraudulent transactions are present when compared to whole number of transactions.

4. Balancing the dataset using a Sampling technique

The credit card dataset which we have chosen for reference dataset consists of an error of 0.172 percent. It gives a meaning that, the referenced dataset contains 0.172 percentage of no genuine transactions. This infers that, the dataset is uneven and is prejudiced towards genuine transaction. Because of the bias the network is unable to identify and could give a correct prediction of the error. This problem can be solved by using 2 techniques, i) under-sampling ii) over-sampling techniques to condense the partiality for accurate results.

Under-sampling technique, balances the dataset by basing on the non-bias class i.e. fraudulent Transactions.

By adjusting, total of genuine transactions on equality with fraudulent transactions by removing the excess genuine values from the data. For example, suppose there is an 100 observations, then 7 fraudulent values give a 7% error. Similar to that we compute the total of genuine transactions for 492 fraudulent ones, by removing the excess. This produces data with 3% error, it become because easier to detection process. By utilizing this method it results to loss of information.

Over-sampling is another technique that is utilized for imbalanced to balanced data. Here, the occurring of bias is due to a replica of information in terms of the recurring rows but not for the loss of information. We need to eradicate the bias. For sample, here recurring non-genuine transitions are added from 492. A total of observations should give an error of 3%. In this context instead of removing, adding a more number of observations. Hence, by utilizing this technique, we can achieve a high-accuracy model.

5. Machine Learning Models

Machine learning is categorized into four: Supervised, Unsupervised, Semi-supervised and Reinforcement learning. The deliberated machine learning algorithms are ensemble models and gradient boosting algorithms.

4. Experimental Setup

The experiment was performed on Windows 7 operating system and the open-source software environment is jupyter notebook environment was developed to run the models. Various libraries are utilized such as NumPy, Pandas, Matplotlib, Seaborn, Sklearn and imblearm.

Formerly stated algorithms are utilized in the experiment are defined in the following.

a. Random forest

Random forest is a supervised learning algorithm and ensemble model. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The universal idea of the bagging method is that an amalgamation of learning models increases the overall result.

b. AdaBoost Classifier

Yoav Freund and Robert Schapire in 1996, introduced an ensemble boosting classifier that is named Ada-boost or adaptive Boosting. The accuracy of classifiers will be increased by combine multiple classifiers. The performance of the AdaBoost classifier becomes strong by syndicating poorly performing classifiers. The fundamental idea driving AdaBoost is to set the weights of classifiers and training the data sample in every iteration such that uncommon observations can be exactly predicted. Somewhat machine learning methods can be utilized as simple classifiers and accepts weights on the training set. Adaboost should meet two conditions. The first condition is classifier ought to be prepared intelligently on differently weighted training examples, and the second condition every emphasis, attempts to give a phenomenal fit to these models by limiting training error.

c. CatBoost

“CatBoost” term emanates from two disputes “Category” and “Boosting”. “Boost” emanates from a gradient boosting. In Machine learning utilize gradient boosting library functions. CatBoost is a powerful machine learning algorithm and easy to implement. It is extensively applied to manifold types of professional challenges like fraud discovery, recommendation things, forecasting and it performs well also. It affords very good results and runs fast with relatively less data.

d. XGBoost:

XGBoost initially shaped by Tianqi Chen and retained by the Distributed Machine Learning Community (DMLC) group. It is also known as extreme gradient boosting is one of the prominent ensemble techniques. Since it was designed and optimised for the sole purpose of model efficiency and computational speed, XGBoost has proven to ambition the edges of computing power for decision trees algorithms. For an extensive variety of regression and classification predictive modelling problems, the XGBoost algorithm is successful.

5. Results and Discussion

To handle which algorithm is most appropriate for identifying fraud transactions problem, various standards for calculation correlation have been utilized. Accuracy, recall, and precision are the most commonly used criteria for determining the outcomes of machine learning algorithms. A contingency table may be used to assess the completeness of the referenced measurements. These metrics were used to assess how well this model worked. The models were tested on SMOTE data, and the results were predicted, as well as the ROC Curve and AUC value.

The following are contingency table for random forest, AdaBoost, Catboost and Xgboost

Table.1. Contingency Table for Random forest

Actual Label	Predict Label		
		0	1
	0	87	14
	1	14	56847

Table.2. Contingency Table for AdaBoost

Actual Label	Predict Label		
		0	1
	0	90	11
	1	595	56266

Table.3. Contingency Table for CatBoost

Actual Label	Predict Label		
		0	1
	0	89	12
	1	509	56352

Table.4. Contingency Table for XgBoost

Actual Label	Predict Label		
		0	1
	0	87	14
	1	14	56847

The following are the ROC curve with AUC value

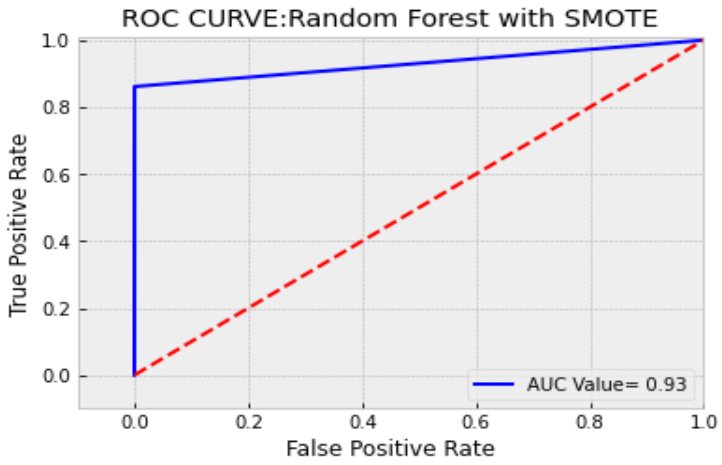

Figure.2. ROC and AUC Value for Random forest

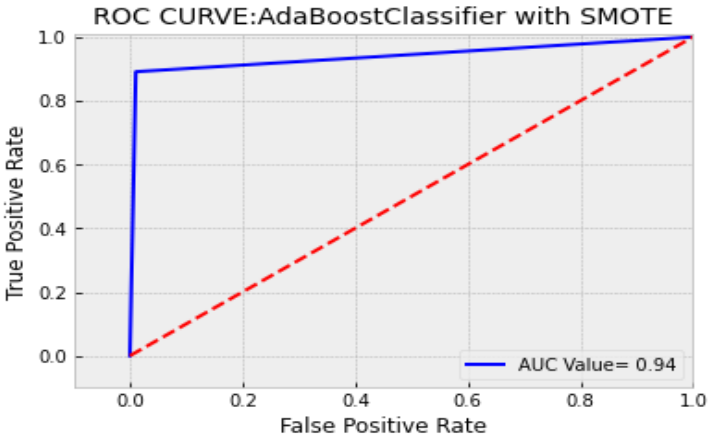

Figure.3. ROC and AUC Value for AdaBoost

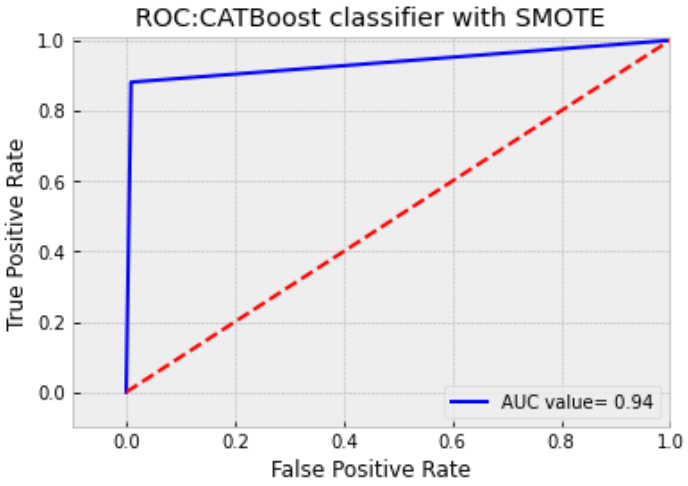

Figure.4. ROC and AUC Value for CatBoost

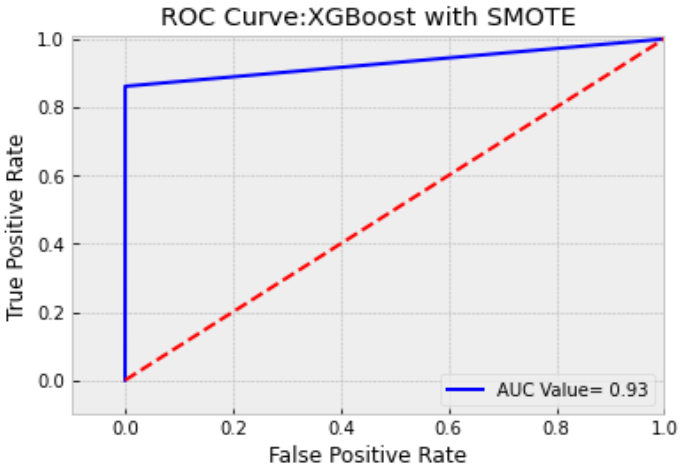

Figure.5. ROC and AUC Value for XGBoost

The metric results obtained for different models are shown in table 5

Table.5. Performance Results

Model	Precession	Recall	Accuracy	AUC value
Random Forest	86%	86%	99%	93%
AdaBoost	13%	89%	98.9%	94%
Catboost	14%	88%	98%	94%
Xgboost	86%	86%	99%	93%

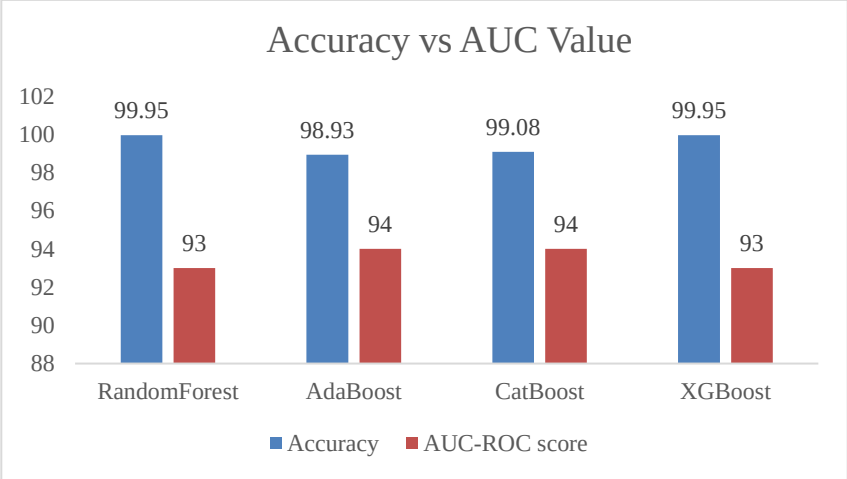

Figure.6. Comparison of Accuracy and AUC value

6. Conclusion

Fraudulent credit card purchases are a major business concern. These scams have resulted in major financial and personal losses. As a result, businesses are continually investing in the creation of new concepts and approaches that will assist in the identification and prevention of fraud.

This paper's main aim was to compare machine learning algorithms for detecting fraudulent transactions. As a result, a distinction was made. As a result of the comparison, it was discovered that the xgboost algorithm offers a best results, i.e. best classifies whether transactions are fraudulent or not. Precession, recall, precision, and area under the curve were used to determine this. It is crucial to have a high recall value for this type of problem. An impact of feature selection and dataset balancing in achieving significant results has been shown.

References

1. HemaGonaboina,&AppalaSrinuvasuMuttipati. (2021). Machine Learning methods for Discovering Credit Card Fraud. . *International Research Journal of Computer Science*, *8*(1), 1–6.
2. Dhankhad, S., Mohammed, E., & Far, B. (2018, July).Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study.*2018 IEEE International Conference on Information Reuse and Integration (IRI)*. https://doi.org/10.1109/IRI.2018.00025
3. Dornadula, V. N., &Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms.*Procedia Computer Science*, *165*. https://doi.org/10.1016/j.procs.2020.01.057
4. Godi, B., Viswanadham, S., Muttipati, A. S., PrakashSamantray, O., &Gadiraju student, S. R. (2020, March). E-Healthcare Monitoring System using IoT with Machine Learning Approaches. *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*. https://doi.org/10.1109/ICCSEA49143.2020.9132937
5. KaithekuzhicalLeenaKurien,& Dr. AjeetChikkamannur. (2019). Detection And Prediction Of Credit Card Fraud Transactions Using Machine Learning .*International Journal Of Engineering Sciences & Research Technology*, *8*(3), 199–208.

6. Malini, N., &Pushpa, M. (2017, February). Analysis on credit card fraud identification techniques based on KNN and outlier detection. *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*. https://doi.org/10.1109/AEEICB.2017.7972424
7. Sailusha, R., Gnaneswar, V., Ramesh, R., &Rao, G. R. (2020, May).Credit Card Fraud Detection Using Machine Learning.*2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. https://doi.org/10.1109/ICICCS48265.2020.9121114
8. Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., &Anderla, A. (2019, March). Credit Card Fraud Detection - Machine Learning methods. *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*. https://doi.org/10.1109/INFOTEH.2019.8717766
9. Awoyemi, J. O., Adetunmbi, A. O., &Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. *2017 International Conference on Computing Networking and Informatics (ICCNI)*. https://doi.org/10.1109/ICCNI.2017.81237