Computing, Information Systems, Development Informatics & Allied Research Journal
Vol. 13 No. 2, June, 2022 - www.cisdijournal.net

Credit Card Fraud Detection Using Voter Ensemble Technique (VET)

[1]Adepegba, O. A., [2]Akinboro, S.A., [3]Olabiyisi, S.O., [1]Lala, O.G. & [1]Onamade, A.A. & [1]Aroyehun, A.A
[1]Department of Computer Science, Adeleke University, Ede, Osun State, Nigeria
[2]Department of Computer Science, University of Lagos, Akoka, Lagos, Nigeria
[3]Department of Computer Science, Ladoke Akintola University, Ogbomoso Nigeria
Corresponding Author's E-mail: adepegbafunmilola01@gmail.com

ABSTRACT

This paper proposes to design a majority vote ensemble classifier for accurate detection of credit card frauds. There is a lack of research studies on analyzing real-world credit card data owing to the issues of confidentiality, some features were hidden. Credit Card Fraud Dataset from kaggle.com was used for this study. The dataset is made up of 284,807 the number of legitimate transactions was found to be 284,315 while the number of fraudulent transactions was found to be 492, this shows that dataset is highly imbalanced, skewed data set like this may lead to an unreliable prediction performance, which is a major classification problem. Hence this study review literatures on methods of addressing the problem imbalance in dataset, and profile a methods of using the Voter Ensemble approach to address the classification problem. The real-world credit card dataset is analyzed to check for missing values and results show that there were no missing or duplicate values in the dataset. In this study, machine learning algorithms are used to detect credit card fraud. The base learners (Logistic regression, Bagging and Naïve Bayes) and the voter ensemble methods are combined. In this experiment, data are splits into 70% training and 30 % testing, the base (learners) classifiers were trained from the training sets, and predictions evaluations were made on data using 10-fold cross-validation. Afterward, the predictions of the learned base classifiers are combined into an ensemble using a stacked generalization strategy. The experimental results positively indicate that the Voter Ensemble method achieves good performances based on Precision, Recall, F1-score and Accuracy rates in detecting fraud cases in credit cards with 95.77%, 99.99%, 97.83%, 99.99% respectively. The result is evaluated and compared with other existing works.

Keywords: Voter Ensemble, Logistic Regression (LR), Naïve Bayse, Bagging, imbalance ratio (IR), CCFD.

CISDI Journal Reference Format
Adepegba, O. A., Akinboro, S.A., Olabiyisi, S.O., Lala, O.G. & Onamade, A.A. & Aroyehun, A.A. (2022): Credit Card Fraud Detection Using Voter Ensemble Technique (VET). Computing, Information Systems, Development Informatics & Allied Research Journal. Vol 13 No 2, Pp 15-.20. Available online at www.computing-infosystemsjournal.info

1. INTRODUCTION

Credit cards have become the common mode of payment for both offline and online purchases as a result of new development in information communication systems and electronic commerce (E-commerce) system causing a substantial rise in transactions related frauds. Every year, fraudulent credit card transactions cost budget companies and customers a considerable amount of money and fraudsters are actively attempting to use new technology and methods to commit fraud (Akshaya, 2022)

In avoiding loss from fraud, two mechanisms can be used: fraud prevention and fraud detection. Fraud prevention is a proactive method, where it stops fraud from happening in the first place. On the other hand, fraud detection is needed when a fraudulent transaction is attempted by a fraudster. (Shim et al 2013) Credit card fraud is concerned with the illegal use of credit card information for purchases. Credit card transactions can be accomplished either physically or digitally. In physical transactions, the credit card is involved during the transactions. In digital transactions, this can happen over the telephone or the internet. Cardholders typically provide the card number, expiry date, and card verification number through telephone or website. With the rise of ecommerce in the past decade, the use of credit cards has increased dramatically (Adewunmi & Akinyelu 2017). Class imbalance in credit card transaction data is a primary factor affecting the classification performance of current detection models. However, prior approaches are aimed at improving the prediction accuracy of the minority class samples (fraudulent transactions), but this usually leads to a significant drop in the model's predictive performance for the majority class samples (legal transactions), which greatly increases the investigation cost for banks.

2. LITERATURE REVIEW

A number of researches are going on in the field of data mining and machine learning such as keyword extraction (Showrov & Sobhan, 2019), summarization (Abulaish *et al.,* 2018; Showrov *et al.,* 2019), breast cancer detection (Showrov *et al.,* 2019b) and so on. Yet, the most challenging problem nowadays in this field is still class imbalance. Several scholars have suggested various types of approaches for dealing with problems with class imbalances. Methods of data level, methods of levelling algorithms and methods of the ensemble are categorized methods. Hybrid methods are another group type for dealing with the problem of class imbalances.

i. Data Level Methods

This approach is geared towards matching the class distributions. The class distribution are being balanced using the sampling methods by resizing the training datasets. The sampling methods can be categorized into techniques for under-sampling and over-sampling. Sampling of existing resampling techniques are SMOTE (Synthetic Minority Oversampling Technique), RUS (Random Undersampling Technique) and so on.

ii. Algorithm Level Methods

Algorithm level approaches, also known as internal approaches, focus on improving the ability of current classifier algorithms to learn from minority classes. For example, adjusting the estimation of probability or modifying the cost per class may benefit the minority class. Support Vector Machine (SVM) Support Vector Machine was introduced in the mid1990s, this technique discriminates over input spaces in a finite area. It is necessary for classifications to be obtained by learning from the training sample (Rahman *et al.,* 2011).

iii Ensemble Methods

Ensemble approaches involve the synthesis of various methods. Ensembles methods build a series of N classifiers and combine them to produce the final classifier C^*. They aim to obtain a high precision classifier. . Applications of ensemble methods for improving performance of imbalanced target variables have been thoroughly studied in classification literature. Several classification studies have demonstrated how ensemble techniques can improve prediction accuracy for imbalanced classes (Galar *et al.,* 2012).

2.1 Related works

Barahim *et al.,* (2019) worked on Enhancing the Credit Card Fraud Detection through Ensemble Techniques. Decision tree, Naives Bayes, support vector machines, boosting and Adaboost were the classifiers used for this study. Boosting and Decision Tree resulted with the highest accuracy of 98.37% and F -Measure of 94.49% while Decision Tree only resulted with accuracy of 98.27% and F -Measure of 93.98%. Use of accuracy and F- measure score as the only performance evaluation metrics. Ishan, *et al.,* (2018) in this study on Ensemble Learning for Credit Card Fraud Detection an ensemble of Random Forests and Neural Networks was used. The experimental results point to this being a superior method than other popular approaches. An open direction of the work is to improve the accuracy parameters of the classifier, nevertheless they were faced with the issue of dataset having numerical values, Unavailability of large Datasets, highly skewed dataset.

3. METHODOLOGY

Data description: The European Credit Card dataset as obtained from the Kaggle database contains only numerical input variables which are the result of a Principal Components Analysis (PCA) transformation of the original data. Due to confidentiality issues and sensitive nature of the data, the original features and more background information about the data were not provided. The only two features that were not transformed with Principal Component Analysis are 'Time' and 'Amount'. The feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset.

Software and Hardware Requirements: Matlab programming language was used for implementation in this work. The reason for the choice of Matlab is because it provides much more favorable data visualization than any other platform and it is a much better language for mathematical computing and math-based algorithms. For effective implementation, models were trained using MATLAB, on Processor of Core i5-5200U, CPU @2.2GH specifications, Memory (RAM) of 4gigabytes, operating system 64-bits Windows OS, while learners run on MATLAB functions. The workings of the ensemble algorithm comprise of first of all pre-defining the k-fold splits of the dataset, then evaluating all the base-learning algorithms on the same splits of the dataset. All out-of-fold predictions are then kept and voting was done how to best combine the predictions from the base-classifiers. The Ensemble Learner is basically a specific configuration of stacking specifically to k-fold cross-validation with all the base-classifiers considered for the predictive modelling. The algorithm adopted here operates as follows. (As derived from Pintelas & Livieris 2020), the majority voting method was used, where the class for unseen instance X can be predicted as thus:

$$M_v(\mathrm{Z}) = \mathrm{argmax}\sum_{i=1}^{K}[H(X) = y]$$

The procedure for the ensemble classifier as used for predicting sample dataset(s) in this study is as represented in Algorithm 1.

Algorithm 1: Voter Ensemble Learning Model Algorithm for Prediction

Input: Sample dataset
Output: Ensemble Learner Model
Data Preparation:
1. Input CCFD dataset
2. Perform PCA on the resampled dataset
3. Split total dataset into Training dataset (70%) and Testing dataset (30%)

Model Training:

4. Select *m* base-classifiers. (In this study, *m* = 3: Logistic Regression, Naïve Bayes, and Bagging)
5. For each base-classifier Select a k-fold split of the training dataset (for k = 10):
   (a) Evaluate using k-fold cross-validation.
   (b) Store all out-of-fold predictions.
   (c) Train the model on the full training dataset and store.
6. Perform Voting using
(M_v(Z) = argmax$\sum_{i=1}^{K}[H(X) = y]$) on the out-of-fold predictions.

Model Testing:

7. Evaluate the Classifier on a holdout dataset (i.e., use the model to make predictions)

Models Training and Evaluation of Credit Card Dataset

Figure 1 represents the architectural design of the voter ensemble approach.

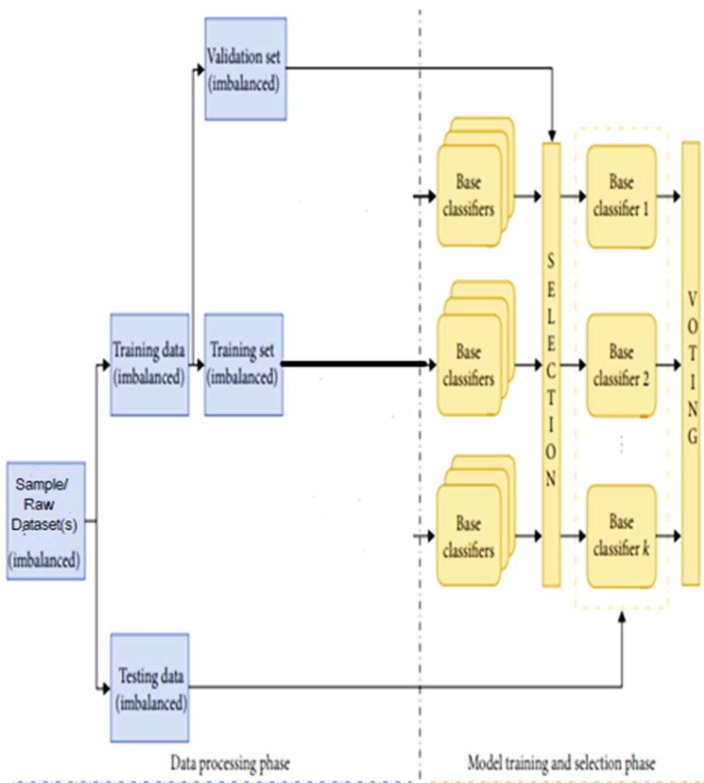

Figure 1: Architectural Diagram of the Voter Ensemble Learner Approach

The resampled dataset was split into 70% training set (consisting of 199,365 records) and 30% test set (consisting of 85,442 records). The entire 70% training set was then presented to the ensemble learner approach (which has three different base learners) for 10-fold cross validation process (in which 10% of the training dataset is held out for an out-of-fold prediction and the remaining used for training).

Computing, Information Systems, Development Informatics & Allied Research Journal
Vol. 13 No. 2, June, 2022 - www.cisdijournal.net

Each of the base models are then trained on the training set and evaluated to obtain their out-of-fold predictions. Then voting was done on all the out-of-fold predictions and evaluated on the held-out test set. All the features were considered for the training process. The performances of each of the base models were measured based on the predictions made and the performance of the super learner ensemble model was also obtained. The performances of the base-models and super learner ensemble were measured using the following performance metrics; accuracy, precision, recall, F1 score and accuracy. Analysis of the samples used for both training and testing.

4. RESULTS AND DISCUSSION

A study on credit card fraud detection using machine learning algorithms has been presented in this paper. A number of standard models which include LR, Bagging, and Naïve Bayes have been used in the empirical evaluation. Available credit card dataset has been used for evaluation using individual (standard) models and hybrid models using majority voting combination methods. At the experiment, base (learners) classifiers were trained from the training sets, and predictions evaluations were made on data using 10-fold cross-validation. Afterward, the predictions of the learned base classifiers are combined into an ensemble using a stacked generalization strategy. Table 1 shows the evaluation of results for Credit card dataset. The majority voting method has yielded the best for all the performance. This shows that the majority voter method is stable in performance even in the presence of noise which is a machine learning algorithm that combines all of the models and model configurations that have been investigated for a predictive modeling problem and then using them to make a prediction as-good-as or better than any single model that were investigated) is employed by using stacking method in combining models together to achieve a more efficient and better result.

Table 1: Result Performance of the Base Learners and the Ensemble Learner on Credit Card Dataset

Performance Metrics	Logistic regression (LR)	Bagging (Bag)	Naive Bayes (NB)	Voter (LR+Bag+NB)
Accuracy (%)	99.986	99.9707	99.3025	99.9813
Precision (%)	98.5507	96.1832	18.8366	95.7746
Recall (%)	99.9977	99.9941	99.313	99.993
F1-score (%)	99.2689	98.0517	31.6669	97.8384

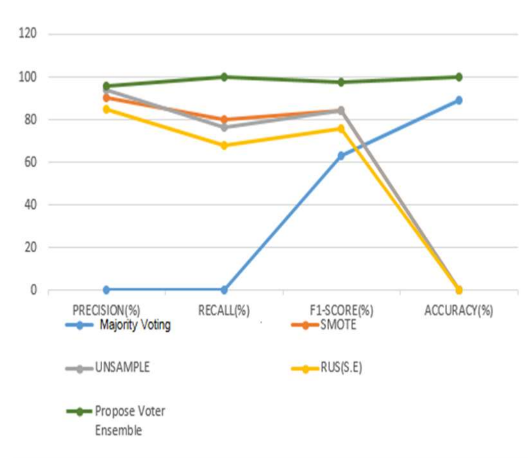

Figure 2: Performance Comparison Of The Proposed System With Other Existing Works

Figure 2. shows the performance comparison of the proposed system with other existing works using same sample CCFD dataset. Meanwhile, the majority voting does not consider precision and recall as part of the metrics used in the study. From all indications, proposed Voter ensemble out performs the existing works in all the performance metrics.

5. CONCLUSION

Khan *et al.,* 2020 explain that ensemble approaches involve the synthesis of various methods and are efficient solutions for the class issue with imbalances. The main hypothesis in this study is the correctly combined base learner models in other to obtain more accurate and/or robust model. The use of voter learner ensemble approach adopted helps to provide a solution to avoid the problem of data balancing experienced in the existing works. The results of this model was not worse than the best performing model evaluated during k-fold cross-validation and has better performance better than any single model.

REFERENCES

1. **Adewumi A. & Akinyelu A. (2017)** A survey of machine-learning and nature-inspired based credit card fraud detection techniques. *International Journal of System Assurance Engineering and Management* volume 8, pages937–953 (2017)
2. Akshaya V., Sathyapriya M., Ranjini D., &
3. Sivanantham S., (2022) Detecting Credit Card Fraud Using Majority Voting-Based Machine Learning Approach, Intelligent Systems and Sustainable Conference paper Computing pp 327–334
4. Abulaish, M., Showrov, M., & Fazil, M. (2018). A layered approach for summarization and context learning from microblogging data. In Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services (pp. 70-78).
5. Printed Edition of the Special Issue Published in Algorithms www.mdpi.com/journal/algorithms
6. Barahim A., Alhajiri A., Alasaibia N., Altanimi N., Aslam N., Khan I., (2019) Enhancing the Credit Card Fraud Detection through Ensemble Technique **Source:** Journal of Computational and Theoretical Nanoscience, Volume 16, Number 11, November 2019, pp. 4461-4468(8) **Publisher:** American Scientific Publishers**DOI:** https://doi.org/10.1166/jctn.2019.8619
7. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,https://doi.org/10.1109/TSMCC.2011.2161285,http://www.keel.es/dataset.php, 2012.
8. Yalong, X., Aiping,, L., Liqun, G., & Ziniu L. (2021) "*A Heterogeneous Ensemble Learning Model Based on Data Distribution for Credit Card Fraud* " College of Computer, National University of Defense Technology, Changsha, Hunan 410073, China Correspondence Aiping Li; liaiping@nudt.edu.cn Received 11 June 2021; Revised 4 July 2021; Accepted 12 July 2021; Published 22 July 2021