International Journal on Emerging Technologies **11**(2): 536-540(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Implementation of Random Forest and Proposal of Borda Count in Credit Card Fraud Detection

Nileena Thomas[1], Jayalakshmi J.[1], Sreelakshmi E.S.[1] and Leena Vishnu Namboothiri[2]
[1]PG Student, Department of Computer Science and IT,
Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India.
[2]Assistant Professor, Department of Computer Science and IT,
Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India.

(Corresponding author: Nileena Thomas)
(Received 11 December 2019, Revised 28 February 2020, Accepted 29 February 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Credit-Card frauds are expanding and it is becoming more astute with passage of time apparently. We need good fraud identification techniques to find these. An extortion discovery strategy should be applied to lessen the pace of Credit-Card cheats. Any supervised learning classification algorithms can be used to predict fraud however, the algorithm that gives the most elevated accuracy among others will predict the distortion better. We use Logistic Regression and K Nearest Neighbour, along with Random Forest, as part of supervised learning classification algorithms and demonstrate that Random Forest has improved accuracy rate. The other two algorithms give low precision rate contrasted with Random Forest. Therefore, to measure the accuracy in fraud detection we use Random Forest. Likewise, to cause this algorithm progressively precise it is reasonable for us to utilize Borda Count rather than Majority Voting, that is, it picks broadly favourable decisions, instead of the one with the majority. So here, we propose that Borda Count gives preferable accuracy score over the commonly used Majority Voting. Random forest constructs numerous Decision Trees and consolidates them to produce an increasingly precise and stable forest.

Keywords: Borda Count, Credit Card Fraud, Fraud Detection, Machine Learning, Majority Voting, Random Forest.

Abbreviations: CCF, Credit Card Fraud; RF, Random Forest; DT, Decision Tree; CaR, Classification and Regression; MV, Majority Voting; BC, Borda Count; PCA, Principal Component Analysis; LR, Logistic Regression; ML, Machine Learning; AUC, Area Under Curve; ROC, Receiver Operating Characteristics; KNN, K Nearest Neighbor; SVM, Support Vector Machine; MLP, Multilayer Perceptron.

I. INTRODUCTION

Credit card utilization has been radically expanded over the world, and so are frauds. These frauds are getting more astute with the progression of time. CCF is maybe the most serious hazard to business establishments today. The attacker requires next to no measure of data for leading any fraudulent activity in the online transaction. Credit card extortion acts done by any individual who, with the expectation to cheat, utilizes a card which is disavowed, called off, stated lost, or taken to get anything of significant worth. Utilizing the unique numeral of the Credit card without ownership of the real card is additionally a category of card fraud. Taking an individual's identity to obtain a Credit card is another additionally undermining type of CCF since it works related to fraud. A lot of information is moved during the online transactions, bringing about a twofold outcome: certifiable or fake. Thus, a considerable amount of money lost is induced by credit card frauds.

There is relatively small number of fraudulent transactions in the overall number of transactions. That is you have, in the first place, a very short period for which to determine whether to describe the transaction as fraud or legitimate and review a vast range of criteria during training and decision-making. So detecting fraudulent transaction from the whole group is strenuous.

RF is one of the Supervised Learning Classification algorithms. It is a propelled rendition of DT. RF develops numerous DT and combines them to call for a progressively exact and stable forecast. We can think about a decision tree as something of a continuum of no/yes analyses presented about our information, which ultimately prompts a foreseen category. RF has the following advantages:

– In Random Forest calculation, an individual tree is raised on a subclass of information. Fundamentally, the RF depends on the intensity of "the group"; in this way, the general biasedness of the algorithm is diminished.

– RF is entirely steady. Regardless of whether another information point is obtainable from the dataset, the general algorithm isn't influenced much since new data may affect one tree, however, it is challenging for it to affect all the trees.

– The RF algorithm functions admirably when information has missing qualities or it has not been scaled well.

– It is very well may be utilized equally in CaR jobs.

– Overfitting is one basic issue that may aggravate the outcomes, however for the RF algorithm, if trees are ample in the forest, the model is not over fitted by the classifier.

– The RF can be utilized for recognizing the most noteworthy highlights from the training dataset, as it were, including designing.

For classification issues, we have to guarantee that the two classes - for our situation, fraudulent and genuine transaction, are available in the training set and test set. Since one class is essentially less happening than the other is, stratified sampling is prompted here as opposed to random sampling. In fact, while random sampling may miss the examples from the least various classes, stratified sampling guarantees that the two classes spoken to in the last subset as per the initial allocation.

Some of the experiments have been carried out by comparing their accuracy, efficiency and working to find an effective algorithm from a group of algorithms. Although some other, publications have taken RF in the CCF detection or prediction. We wanted to demonstrate here that RF works best from a random algorithm collection. We also recommend BC instead of MV as BC produces the best result with RF.

RF uses MV in predicting the result. However, we propose the RF & BC combination instead of RF & MV. Every elector has one vote in majority voting which can be cast on behalf of any one candidate. The candidate, who earned the majority (greater than 50%) of the votes, wins the election. The actual benefits of this approach are its low mistake count and easiness. The technique solely chooses a winner in the event of a majority contestant, so the majority of the classifiers must be wrong in order to create an error. The odds of this happening are small, particularly with lots of classifiers. We propose to replace it with BC. The advantages of this RF & BC combination in contrast with the MV is that, the BC is aimed to choose broadly agreeable choices or candidates, as opposed to those favored by a dominant component, as is frequently portrayed as a consensus-grounded voting method instead of a one with the majority. Despite the point that the BC has the necessity of a total ranking, it is not as much of demanding as the confidence strategies.

II. BACKGROUND STUDY

Lakshmi *et al*., (2018) in their studies have shown that the accuracy of LR, DT and RF is 90.0, 94.3 and 95.5 respectively. So it is better only for RF than the LR and DT [1]. Tsymba *et al*., (2006) resolved that one way for improving RF is to replace the MV with a more sophisticated combination function [2]. Niveditha *et al*., (2019) concluded that RF attains good results on a small data set. The accuracy level compared to other algorithms gives more [3]. Devi Meenakshi *et al.,* (2019) adopt the two algorithms SVM and RF and compared results. Although its speed is affected, the Random Forest Algorithm works best with a large training data. Although SVM requires good preprocessing to achieve better results given its imbalanced data set problems [4]. Khare and Sait (2018) the authors stated that RF obtains the best outcome with clear-cut accuracy of 98.6%. Different methods like LR, SVM, and DT have accuracy of 97.7%, 97.5% and 95.5% respectively. The results got along these lines presume that RF gives the most exact and great accuracy score of 98.6% in the issue of CCF detection. Even though in RF speed during testing and application will suffer, it execute improved with grander size of training data [5]. The accurate CCF detection value was obtained, i.e. 99.93 per cent using RF algorithm by Jonnalagadda ET AL. The RF algorithm will increase performance with many training data, but speed will continue to suffer during testing and implementation. Utilizing more pre-processing techniques will help this [6]. Monika *et al.,* studied that over a larger number of training data, the Random Forest Algorithm can do better and the result is 99.9%. Alternatively, SVM can be used, but still suffers from an imbalanced collection of data and needs further pre-processing in order to achieve better performance [7].

Van Erp *et al*., (2002) discusses the Borda Count. It says that the strategy which works very great on the small MLP mix is the BC. It is less difficult as well. The BC performs great on bigger ensemble sizes, so it consequently turns into a supplement to the product rule and sum rule. The MV effecting is low conversely with the various voting techniques. In this way, it is desirable over training the product rule, sum rule or the BC rather than plurality voting [8]. Randhawa *et al*., (2018) a technique in ML is utilized for CCF detection. At first, standard models were utilized later hybrid classic appeared which made use of AdaBoost and MV methods. The data-sets that are publically reachable have been used to survey the model viability and included dataset set utilized from the financial sector. Numerous voting techniques attained a decent score of 0.942 for 30% included noise [9]. Shirgave *et al.,* in their paper has analyzed numerous algorithms to detect fraud during the credit card purchase, such as RF, LR, SVM, DT, KNN, and Naive Baye. Many factors act as the foundation of the success of all these techniques. The supervised RF learning technique was selected to identify the warning as fraud or legal [10]. Hasan *et al.*, (2014) authors technologically advanced two prototypes for IDS using SVM and RF. Random Forest takes a shorter period of time to train the classifier than SVM [11]. Maniraj *et al.,* explained in detail, how to apply machine learning to achieve better results in fraud detection along with the algorithm, pseudo code, description and experimental findings. Based on the machine learning algorithms, this system improves its performance even when more data is put in it over time. This large percentage of accuracy is predicted by the vast difference between the successful and real transactions [12]. Thennakoon *et al.,* in their paper they suggest a new detection method for credit card fraud by detecting four different models of fraudulent transactions using the best suited algorithms and by resolving the related problems found by past credit card analysts. The machine learning models with the highest accuracy of four frauds are LR, NB, LR and SVM. In addition, the models show an accuracy of 74%, 83% 72%, and 91% respectively [13]. Varmedja *et al.,* in their paper the main objective was to compare several algorithms for machine learning like LR, RF, Naive Bayes and MLP, and random Forest shows the best results, namely to decide whether transactions are legal or fraud. This was calculated by various tests, such as recall, accuracy and precision [14].

III. PROBLEM STATEMENT

The most exceptionally dreadful matter is that frauds are yet expanding un-defensive and un-criminologist way. Credit card extortion is not uncommon, and it could occur to anyone. It is one among the quickest developing sorts of frauds and one among the hardest to forestall. A misrepresentation identification technique should be applied to decrease the pace of fruitful credit card cheats. Consequently, a powerful and inventive technique should be created which will develop in like manner to the need. In our proposed paper, we assembled the CCF recognition utilizing Machine Learning. In this way, we move for the succeeding phase for identifying false and genuine cases in credit cards. We utilize supervised learning calculation, for example, Random forest calculation to characterize the CCF exchange in on the web or by disconnected. Therefore, we apply the RF calculation to group the dataset containing the credit card transaction.

Random Forest is adopted in predicting the fraudulent and the genuine transaction by training it in this. It uses Majority Voting to create a prediction grounded on the majority decision. However, the drawback is that when no majority candidate is available, no outcome is created and this voting technique dismisses the sample.

So here, we bring Borda Count. Majority vote and Borda Count and are compared for their accuracy the one with the better results can be combined with RF for more accurate prediction of results.

IV. IMPLEMENTATION

LR, KNN, etcetera are ML algorithms. Here the above-mentioned algorithms are utilized to compare and show that RF gives the best accuracy in contrast to the other algorithms.

Logistic Regression: LR a classification algorithm, appoints observations to a discrete system of classes. Logistic regression changes its yield using the logistic sigmoid to reestablish probability value which then have the option to be mapped to at any two discrete classes.

K Nearest Neighbor: An Algorithm which is utilized to take care of both CaR problems. The KNN algorithm supposes that relative things be in nearness.

Dataset: To foretell the CCF, the dataset we use here has 284,807 transactions in total and 492 of them are frauds. European cardholders made these transactions in September 2013. Input variables are numerical because of PCA. PCA was done on some features of this dataset due to some confidentiality concerns. Principal components attained because of this transformation are features from V1 to V28. Time and Amount are the other features. 'Time' covers the seconds slipped by between every exchange and the primary exchange. 'Amount' is the transaction value. 'Class' takes 1 for fraud and 0 otherwise. Fig. 1 shows the distribution of legal and fraud transaction in the dataset.

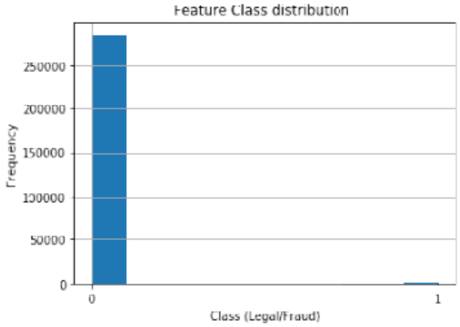


Fig. 1. Feature Class Distribution.

AUC Graph: A ROC curve is utilized to assess the accuracy of a classification prediction. The bigger the zone underneath the ROC curve, the higher the accuracy is. In the event, that is increasingly centred on the accuracy, we tried some algorithms for taking care of the issue.

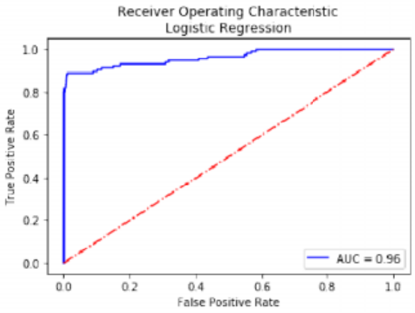


Fig. 2. ROC-LR.

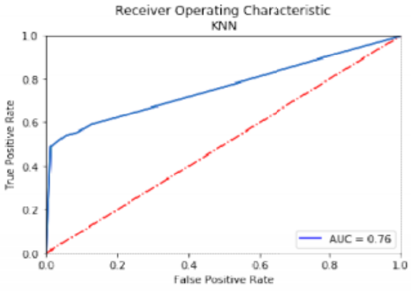


Fig. 3. ROC–KNN.

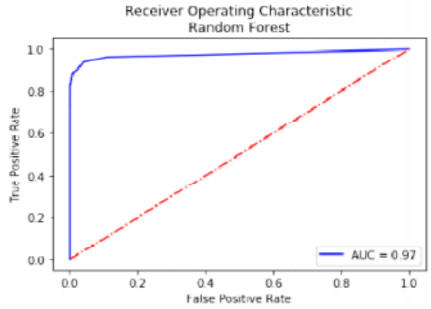


Fig. 4. ROC–RF.

From AUC Graphs in Fig. 2, 3 and 4, the more area beneath the ROC curve is for Random Forest (Fig. 4.), AUC = 0.97. Also below, the metrics evaluation scores for these algorithms. RF has improved scores compared to other algorithms.

Logistic Regression:

Accuracy:	0.9804780764585265
Precision:	0.07152496626180836
Recall:	0.8833333333333333
AUC:	0.964672608161466

KNN:

Accuracy:	0.9408724474031628
Precision:	0.015632401705352912
Recall:	0.55
AUC:	0.7616222841058754

Random Forest:

Accuracy:	0.999522485323446
Precision:	0.8706896551724138
Recall:	0.8416666666666667
AUC:	0.9657257932153476

Working of Random Forest:

Below is the pseudo-code of RF creation:

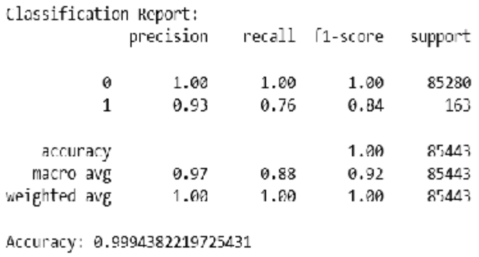
```
– Select "N" features from all out of "M" features where N << M
– Among the "N" features, compute the node "D" utilizing the best split point
– Split the node into child node utilizing the best split
– Repeat the above steps until the "K" number of nodes has been attained
– Construct forest employing reiterating above steps for "n" number of times to make "n" count of trees.
```

Pseudocode of prediction in Random Forest:

```
– Go through the test qualities and utilize the rules of every single arbitrarily made decision tree for predicting the outcome and stores that result (target)
– Compute the votes in support of each anticipated target
– Consider the highly voted target as the concluded result from the RF calculation.
```

As the features are PCA transformed, it does not require data cleaning. However, data cleaning is needed, as the actual world data is messy. Nevertheless, the dataset we used is imbalanced with the feature as "class" which has slanted distribution. The model will be marked under-fitting or overfitting if it is imbalanced. This is often faced in classification subjects. To construct an ideal model, we have to have a balanced dataset to accomplish higher accuracy.

Supervised classification algorithm’s prerequisites are training set to train the model and a test set to assess the prototypical value. In the wake of perusing, the information, accordingly, must be parceled into a training set and a test set. Basic dividing extents differ between 80-20% and 60-40%. For our model, we embraced 70-30% apportioning, where 70% of the first information is placed into the training set, remaining 30% is held as the test set for last model assessment. Now we have split the data and now will implement the RF algorithm. In Fig. 5, it shows the classification report and accuracy we get when fraud prediction is made using RF.



```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     85280
           1       0.93      0.76      0.84       163

    accuracy                           1.00     85443
   macro avg       0.97      0.88      0.92     85443
weighted avg       1.00      1.00      1.00     85443

Accuracy: 0.9994382219725431
```

Fig. 5. Classification Report of RF.

RF uses MV as the voting method. That is the RF as the conclusive outcome picks the choice of most of the trees. It goes with the majority. However, the BC is expected to choose comprehensively worthy alternatives, as opposed to those favored by a dominant part. When there is no majority available, then there will be no outcome at all. Thus, MV will dismiss the sample. In BC, voters rank choices or applicants arranged by liking. The BC decides the result of argumentation by giving every candidate, for every voting form, various focuses relating to the number of competitors positioned lower. When the total of what votes have been tallied the choice or applicant with the most focuses is the winner. Moreover, the accuracy rates of both approaches are given.

(’Accuracy with Borda Count: ’, 0.97960432014334976)
(’Accuracy with Majority Voting:’, 0.95670432014334976)

V. CONCLUSION

Fraud transactions are generally little contrasted with genuine transactions. From the experiments, the conclusion is that LR and KNN has an accuracy of 0.98 and 0.94 respectively. However, the best outcomes are gotten by RF with an exact accuracy of 0.999. The outcomes acquired shows that RF gives the most exact and great accuracy score in detecting of CCF with the given dataset. Therefore, Random-forest method built on the credit card dataset demonstrated 0.999 accuracies in fraud detection.

The most significant improvement in the portrayal of RF is accomplished by changing the voting mechanism. That is Random forest, when combined with Borda Count, gives much better accurate results. The BC performs improved on bigger ensemble sizes. The MV performance is low in contrast with BC. With a bigger size of training data, RF performs better but its speed will writhe.

VI. FUTURE SCOPE

The RF works well when the MV is replaced with a much-sophisticated combination function, where here we proposed Borda Count. What is not presented here is that MV dismisses an enormous number of samples as doubtful (no majority competitor). Accordingly, the mistake pace of the MV on non-dismissed samples lowers. Random Forest gives much better scores than others but its prediction procedure is very slow in correlation with different algorithms. The total training and testing time for all the features was 240.99 seconds. Therefore, as future work, it will be better to work on its run time and to speed up the algorithm.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude and thanks to Dr. U. Krishnakumar, Director of Amrita School of Arts and Sciences, Kochi and Dr. Maya L Pai, HOD of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, for their leadership and guidance in all of the official support. To Leena Vishnu Namboothiri, Assistant Professor, Amrita School Of Arts and Sciences, Kochi, for her valuable guidance and support. Moreover, to all of our teachers and friends for their co-operation and moral support.

Conflict of Interest. The authors affirm that this paper does not include any established conflicts of interest.

REFERENCES

[1]. Lakshmi, S. V. S. S., & Kavila, S. D. (2018). Machine Learning for Credit Card Fraud Detection System. *International Journal of Applied Engineering Research*, 16819-16824

[2]. Tsymbal, A., Pechenizkiy, M., & Cunningham, P. (2006). Dynamic Integration with Random Forests. Machine Learning: ECML 2006. ECML 2006. *Lecture Notes in Computer Science*, 801-808.

[3]. Niveditha, G., Abarna, K., & Akshaya, G. V. (2019). Credit Card Fraud Detection Using Random Forest Algorithm. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *5*(2), 301-306.

[4]. Devi Meenakshi, B., Janani, B., Gayathri, S., & Indira, N. (2019). Credit Card Fraud Detection using Random Forest. *International Research Journal of Engineering and Technology, 6*(3), 6662-6666.

[5]. Khare, N., & Sait, S. Y. (2018). Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models. *International Journal of Pure and Applied Mathematics,* 825-838.

[6]. Jonnalagadda, V., Gupta, P., & Sen, E. (2019). Credit card fraud detection using Random Forest Algorithm. *International Journal of Advance Research, Ideas and Innovations in Technology, 5*(2), 1797-1801.

[7]. Monika, S., Venkataramanamma, K., Paul, P. P., & Usha, M. (2019). Credit Card Fraud Detection using Random Forest Algorithm. *International Journal of Research in Engineering, Science and Management, 2*(3), 131-133.

[8]. Van Erp, M., Vuurpijl, L., & Schomaker, L. (2002). An overview and comparison of voting methods for pattern recognition. In *Proceedings Eighth International Workshop on Frontiers in Handwriting,* 195-200. IEEE.

[9]. Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K., (2018). Credit Card Fraud Detection Using AdaBoost and Majority Voting. *IEEE Access, 6*, 14277-14284.

[10]. Shirgave, S. K., Awati, C. J., More, R., & Patil, S. S. (2019). A Review on Credit Card Fraud Detection

Using Machine Learning. *International Journal of Scientific & Technology Research, 8*(10), 1217-1220.

[11]. Hasan, M. A. M., Nasser, M., Pal, B., & Ahmad, S. (2014). Support Vector Machine and Random Forest Modeling for Intrusion Detection System (IDS). *Journal of Intelligent Learning Systems and Applications, 6*, 45-52.

[12]. Maniraj, S. P., Saini, A., Sarkar, S. D., & Ahmed, S. (2019). Credit Card Fraud Detection using Machine Learning and Data Science. *International Journal of Engineering and Technical Research, 8*(9), 110-115.

[13]. Thennakoon, A., Bhagyani, C., Premadasa, S., & Mihiranga, S., Kuuwitaarachichi, N. (2019). Credit Card Fraud Detection using Machine Learning and Data Science. *International Conference on Cloud Computing, Data Science & Engineering (Confluence), 9*, 488-493.

[14]. Varmedja, D., Karanovic, M., Sladojevic, S., & Arsenovic, M. (2019). Credit Card Fraud Detection - Machine Learning methods. *International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 18*, 1-5.

How to cite this article: Thomas, N., Jayalakshmi J., Sreelakshmi, E. S. and Namboothiri, L. V. (2020). Implementation of Random Forest and Proposal of Borda Count in Credit Card Fraud Detection. *International Journal on Emerging Technologies*, *11*(2): 536–540.