Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220083

Improve Accuracy in Prediction of Credit Card Approval Using Novel XGboost Compared with Random Forest

Pathipati Yasasvi[a] and S. Magesh Kumar[b,1]
[a] *Research Scholar, Department of CSE, Saveetha School of Engineering,*
[b]*Professor, Department of CSE, Saveetha School of Engineering,*
[a,b]*Saveetha Institute of Medical and Technical Sciences,*
[a,b]*Saveetha University, Chennai, Tamilnadu. India*

Abstract: The aim of this work is to conclude the credit card approval using XGBoost algorithm and compare it with Random Forest (RF) to improve accuracy. Prediction of credit card approval using XGboost Classifier with sample size of N =10 and logistic regression with sample size of N =10, and dataset size of 48678. The dataset contains 19 attributes that help to determine whether a person gets approval for a credit card or not. The **a**ccuracy of the Xgboost Classifier is 87.97% and loss is 12.03%, which appears to be better than Random Forest (RF), which is 82.86% and loss is 17.14 %, with a significant value $p = 0.001$ ($p<0.05$, 2-tailed) in SPSS statistical analysis. The results show that the Novel Xgboost Classifier seems to perform significantly better than Random Forest (RF) for credit card approval prediction in terms of accuracy.

Keywords. Prediction, Credit Card Approval, Machine Learning, Novel Xgboost Classifier, Random Forest (RF), Accuracy.

1. Introduction

A credit card is a payment card issued to users, which enables the cardholder to pay a merchant for goods and services based on the cardholder's promise to the card issuer to pay them for the amounts plus other agreed charges. Credit card approval dataset is used to analyze and predict the approval of credit cards by using ML algorithms (Sugiyarto, Sudarsono, and Fadillah 2019)[1]. By using the previous customer credit_score machine learning model, evaluates customers as valid for loan purposes. [2]There is an order to issue credit cards that have various applications like (Yu 2020) usage in banking sectors, financial purpose of individuals, [3] and ease of payments for merchants and payers (Duan 2020). Bank employees easily access good customer applicants for providing credit cards and loans (Zarnaz, Biswas, and Hasan 2021)[4].

According to the research credit card approval, 65 related research articles were evaluated and 98 articles were published in IEEE and Science Direct, 147 articles were published in Google Scholar, and 12,970 articles were published in Springerlink[5]. [6]In this work (Huang 2020) they used a decision tree algorithm with an improved

[1]S. Magesh Kumar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu; E-mail: mmce6450@gmail.com.

accuracy of 89%.[6-10][4] (Warghade, Desai, and Patil 2020) implemented a logistic regression algorithm with an accuracy of 79%. (Azhan and Meraj 2020) used ANN algorithm with an accuracy of 67%. (Gupta et al. 2020) implemented a random forest algorithm with an accuracy of 75%. (SriLaxmi 2020) used an isolation forest algorithm with an accuracy of 88%. (Popat and Chaudhary 2018) ANN was used and accuracy was 85%. (Huang 2020) In this paper, it is the best model for detecting credit cards using machine learning techniques.

In this application, accuracy is 89%, which is comparatively more than the remaining models. Our wide portfolio of research has translated into publications in numerous interdisciplinary projects. The existing literature shows a lack of accuracy in prediction of credit card approval. Objective of the current study is to predict credit card approval using xgboost classifier and random forest for improved accuracy by removing outliers in the dataset, which is a major reason for the defective results of the model.

2. Material and Methods

The research and required work are carried out in the Image Processing Lab, Department of Computer Science and Engineering, Saveetha School of Engineering. The total groups are splitted into 2 and they are defined as Group 1 is XgboostClassifier and Group 2 is Random Forest Algorithm. Each group contains 10 samples, for a total of 20 samples. The computation is performed using G-power 0.8, alpha - 0.05, and beta 0.2 with a confidence interval of 95% (Changwat Poll 2017). In this study, accuracy of two algorithms, i.e, Xgboost classifier and Random Forest are compared. Credit Card Approval dataset is used to determine whether a person is approved for a credit card or not. The data set was collected from Kaggle.com. Dataset contains 19 attributes such as id, gender_code, Amt_Income_Total and so on. Dataset is divided into two parts, i.e., testing data and training data.

2.1. Xgboost Classifier

The Xgboost Classifier helps to create the classifier models to allow the wrapper classes. This wrapper class provides a regression framework for scikit-learn. Scikit-learn is a machine learning library providing various tools which lead to regression, clustering, and classification models. By using scikit-learn library, we can create multiple wrapper classes for the data models. Xgboost classifier library provides a feature to convert a dataset into a subset by allowing various features. While converting the dataset into a subset, the pre-trained model will work as the entire training is loaded into a single training model. Because of this feature, the threshold can be low, accuracy can go high, and loss can become low (Wade 2020). The pseudocode of the Xgboost Classifier is shown in Table 1.

Table 1. Pseudocode of the Xgboost Classifier

```
Input:
X:premodel dataset
X1: No.of classifiers present in xgboost for given input
For loop starts
X:preprocessing the data and test participle.
End loop
While loop starts
Extracting text subject characteristics
End loop
Output:
X:X1 using xgboost discretize features
```

2.2 Random Forest

Random forest is a bagging technique, and it is not a boosting technique. Random trees in random forests are run in parallel. While building trees, there is no interaction between these trees. The basic idea behind Random Forest is that it combines multiple decision trees to determine the final output and also classifies data into subsets. Those subsets are said to be trees. A decision tree gives only one decision, whereas a random forest gives multiple decision trees. These algorithm takes results from all decision trees and predicts the output based upon the voting majority of each decision tree. If data has more decision trees, then prediction is more accurate with less overfitting. That is, it builds multiple decision trees and merges their expectations with each other to get a more accurate and consistent estimate. The random forest algorithm consists of two stages: one is random forest creation, and the other is estimation from the random forest regulator created in the first stage. The pseudocode of Random Forest is shown in Table-2. A system with Windows 10th Gen and a 64-bit OS is used. The RAM of 8GB and the language used in Python are implemented in Jupyter (Anaconda). The processor used is an Intel i7, 10Th Gen. The accuracy results are shown in Table 3 and Table 4 for both the algorithms.

Table 2. Pseudocode for Random Forest.

```
Input:
X:pre model dataset
X1: No .of classifiers present in random forest for given input
For loop starts
X:X1→randomly selected features h,h1
For loop ends
While loop
Calculate each node in h features to extract best split points
Repeat with h1
End while loop
Output
H:h1→d
Where d=no of tree are created
```

Table 3. Predicting Accuracy of credit card approval using XGBoost Classifier (mean accuracy=87.97, mean Loss=12.04)

Sl.No	Sample_Size	XG Boost Accuracy in percentage	Loss
1	100	86.00	14.00
2	200	85.50	14.50
3	300	88.66	11.34
4	400	89.50	10.50
5	500	88.60	11.40
6	600	89.33	10.67
7	700	88.00	12.00
8	800	87.75	12.25
9	900	88.11	11.89
10	1000	88.20	11.80

Table 4. Predicting the accuracy of credit card approval using Random Forest (mean accuracy= 82.86, mean Loss= 17.14)

Sl.No	Sample_Size	Random Forest Accuracy in Percentage	Loss
1	100	82.00	18.00
2	200	82.00	18.00
3	300	83.00	17.00
4	400	83.25	16.75
5	500	83.20	16.80
6	600	83.66	16.34
7	700	83.57	16.43
8	800	82.62	17.38
9	900	83.00	17.00
10	1000	83.30	17.70

2.3 Statistical Analysis

Statistical Software used for our study is IBM SPSS version 23. The independent variables are credit card_score and annual _income and the dependent variable is improved accuracy values. An independent t-test analysis has been carried out in this analysis.

3. Results

Table 1 and Table 2 show the pseudocode for the Xgboost and Random Forest algorithms. Table 3 and Table 4 represent data collection from samples of datasets for credit card approval prediction using the novel Xgboost Classifier and Random Forest algorithm to gain accuracy (%) as given in equation (1)..

$$\text{Accuracy}=\text{TP}+\text{TN} / \text{TP}+\text{TN}+\text{FP}+\text{FN} \qquad (1)$$

Where TP stands for True Positive. TN = True Negative, FP = False Negative Loss. The IBM SPSS version 23 statistical software was used for our study. The independent variables are Credit card_score, Annual_income and the dependent variable is improved accuracy values (%).

Sample sizes are statistically tested by using the SPSS tool with the GroupID-1, i.e., Xgboost classifier and the GroupID-2 Random Forest (RF) algorithms. This study observed that the Xgboost classification proved to have significantly better results than the random forest algorithm and 87.97% better accuracy. The mean accuracy values and mean loss values for Xgboost and Random Forest are shown in Table 3 and Table 4. The Independent paired T-Test values in SPSS results are shown in Table 5 and Table 6. The Statistical significance of graphical representation states that Xgboost gives more accuracy and less loss value in comparison with Random Forest (RF) classifiers, which are depicted in Figure 1.

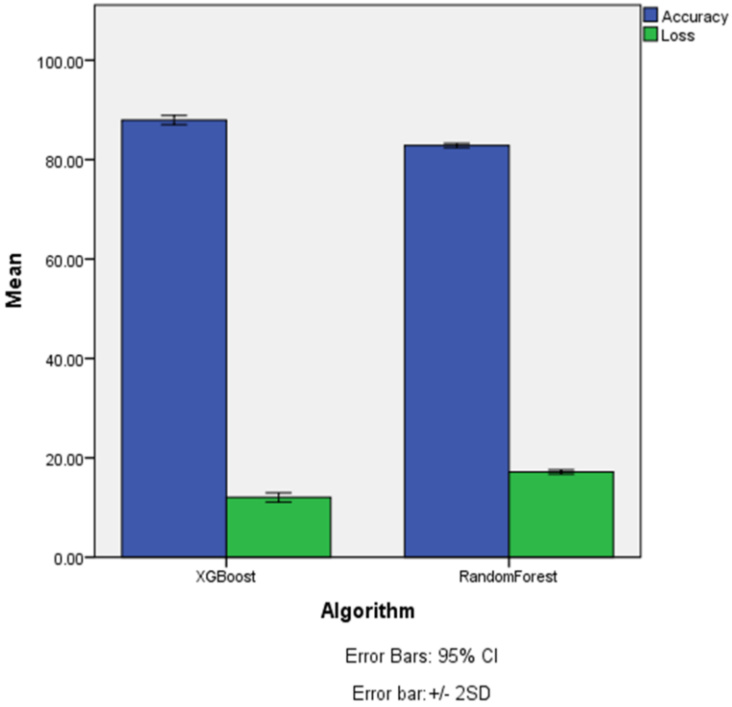

Figure 1. Xgboost and Random Forest accuracy and loss comparison.

Table 5. The **mean** accuracy and loss for Xgboost in a group statistical analysis are 87.96 and 12.03, respectively. For the Random Forest algorithm, the mean accuracy and loss are 82.86 and 17.14, respectively.

Algorithm		N	Mean	Std.Deviation	Std.Error Mean
Accuracy	XGBoost	10	87.9650	1.29815	0.41051
	Random Forest	10	82.8600	0.60566	0.19153
Loss	XGBoost	10	12.0350	1.29815	0.41051
	Random Forest	10	17.1400	0.60566	0.19153

Table 6. Independent Sample T-test(Xgboost Classifier better than Random Forest Algorithm)

		Levene's Test for Equality of Variances		T-test for equality of means						
		F	sig.	t	df	Sig. (2 - tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the difference T	
									Lower	Upper
Accuracy	Equal Variances assumed	2.228	.153	11.269	18	.000	5.105	.452	4.153	6.056
	Equal variances not assumed			11.269	12.741	.000	5.105	.452	4.124	6.085

In the above figure 1. The Xgboost classifier has more accuracy in comparison with random forest. The Random forest has more losses in comparison to Xgboost. It will lead to a decrease in the prediction of the given input. X-Axis: Xgboost Classifier vs. Random Forest, Y-Axis: Average credit card approval accuracy of ± 2 SD with 95% confidence interval (CI).

4. Discussion

In this study, Random Forest (RF) appears to have lesser accuracy (82.97) and more loss (12.04) than Enhanced accuracy and reduced loss in Xgboost Classifier (average accuracy = 87.97, average loss = 12.04). In this paper, [5](Huang 2020) they used a decision tree algorithm with an improved accuracy of 89%. [6] (Warghade, Desai, and Patil 2020) they implemented a logistic regression algorithm with an accuracy of 79%. (Azhan and Meraj 2020) they used ANN algorithm with an accuracy of 67%. (Gupta et al. 2020) they implemented a random forest algorithm with accuracy of 75%. (SriLaxmi 2020) they used an isolation forest algorithm with an accuracy of 88%. A (Kumar and Iqbal 2019) concept drift adaptation algorithm was implemented and the accuracy is 75%. A (Carter and Catlett 1987) decision tree algorithm was implemented and the accuracy is 65%. ANN was used (Popat and Chaudhary 2018) and accuracy is 85%. (Huang 2020)[6]. In this application, the accuracy is 89%, which is comparatively more than the remaining models. In all the above papers, novel credit card approval is done using various algorithms and gives better performance with benchmark customer data.

5. Conclusion

In this study, Random Forest (RF) appears to have less accuracy and more loss than Xgboost Classifier. The Xgboost Classifier appears to perform significantly better than

Random Forest (RF) for credit card approval prediction. Xgboost is used to predict the approval of credit cards based on customer data by varying their income amount, Good score and Bad score from real-time data of customers and achieve better accuracy than the random forest algorithm. The limitation of the proposed work is that it lacks the identification of approval of credit cards without the input of grouped data of customers in order to train the model. In the future, after the approval of a credit card, we will try to provide insights of approved credit cards to customers.

References

[1] Sugiyarto, Ipin, Bibit Sudarsono, and Umi Faddillah. 2019. "Performance Comparison of Data Mining Algorithm to Predict Approval of Credit Card." SinkrOn. https://doi.org/10.33395/sinkron.v4i1.10181

[2] Yu, Yue. 2020. "The Application of Machine Learning Algorithms in Credit Card Default Prediction." 2020 International Conference on Computing and Data Science (CDS). https://doi.org/10.1109/cds49703.2020.00050

[3] Duan, Lei. 2020. "Performance Evaluation and Practical Use of Supervised Data Mining Algorithms for Credit Card Approval." 2020 International Conference on Computing and Data Science (CDS). https://doi.org/10.1109/cds49703.2020.00057

[4] Zarnaz, Zaima, Dipannita Biswas, and K. M. Azharul Hasan. 2021. "Credit Card Approval Prediction by Non-Negative Tensor Factorization." In 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE. https://doi.org/10.1109/icrest51555.2021.9331172

[5] Carter, Chris, and Jason Catlett. 1987. "Assessing Credit Card Applications Using Machine Learning." IEEE Expert. https://doi.org/10.1109/mex.1987.4307093

[6] Huang, Jiayi. 2020. "Credit Card Transaction Fraud Using Machine Learning Algorithms." Proceedings of the 2019 International Conference on Education Science and Economic Development (ICESED 2019). https://doi.org/10.2991/icesed-19.2020.14

[7] Warghade, Swati, Shubhada Desai, and Vijay Patil. 2020. "Credit Card Fraud Detection from Imbalanced Dataset Using Machine Learning Algorithm." International Journal of Computer Trends and Technology. https://doi.org/10.14445/22312803/ijctt-v68i3p105

[8] Azhan, Mohammed, and Shazli Meraj. 2020. "Credit Card Fraud Detection Using Machine Learning and Deep Learning Techniques." 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). https://doi.org/10.1109/iciss49785.2020.9316002

[9] Gupta, Meenu Swati, Meenu, Swati Gupta, Sanjay Patel, Surender Kumar, and Goldi Chauhan. 2020. "ANOMALY DETECTION IN CREDIT CARD TRANSACTIONS USING MACHINE LEARNING." International Journal of Innovative Research in Computer Science & Technology. https://doi.org/10.21276/ijircst.2020.8.3.5

[10] Popat, Rimpal R., and Jayesh Chaudhary. 2018. "A Survey on Credit Card Fraud Detection Using Machine Learning." 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). https://doi.org/10.1109/icoei.2018.8553963.