International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 07 July 2024
Page No: 1909- 1914
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0261

Enhancement of XG-Boost Using Custom Hyper Parameter Tuning for Bank Churning

S.P. Valli[1], Sharmila Sankar[2], C. Hema[3], Mohammad Munzir[4]

[1,3]Associate Professor, Computer Science and Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India.

[2]Professor, Computer Science and Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India.

[4]UG-Computer Science and Engineering, B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India.

Email ID: *vallisp@crescent.education[1], sharmilasankar@crescent.education[2], hemac@crescent.education[3], mohammadmunzir2k@gmail.com[4]*

Abstract

Bank is an important component of our society which deals with money transaction i.e., lending and deposit of money. Customer churn is termination of business of a customer with the company. Bank customer churn creates an impact on revenue and operational efficiencies of banks, where a customer switches or leaves availing the services of bank. Bank is an important part of our society since it makes money by lending money to others. To understand customer churning behavior it is necessary to retain customers and increase the number of customers. In order to predict the bank customer churning behavior a few algorithms such as XGBoost, CatBoost, AdaBoost, Random Forest, K Near Neighbor, Decision Tree, and Logistic Regression are analyzed. Finally, the best model has been recommended by analyzing the above-mentioned algorithms

Keywords: *AdaBoost; CatBoost; Customer bank churn; hyper parameter tuning model; predictive model.*

1. Introduction

Banking sector occupies an important position in the modern societies. The retail banking caters to the financial need of the people like deposits, withdrawals, and loans. Customers are the backbone of the banking system. Therefore, banks always try to retain and even increase the number of customers to increase their profitability and efficiency. But there is always the danger of customers shifting to other banks and terminating availing the services of the bank, which is known as customer churn. Customer churn is defined as the customer terminates its business with the company. Banks often deal with customer churn to retain their customers. As banks spend most of their energy and money attracting new customers to increase profit, banks tend to put in less effort to concentrate on the wellbeing of existing customers. It is found in earlier research that cost of acquiring new customers are higher than retaining existing customers [1]. Bank faces customer churn on a regular basis. Therefore, it becomes necessary for the bank to analyze the customer's behavior, predict the customer churn, and formulate strategies to reduce it. In the bank dataset there are many customers, credit card users and non-credit card users. This study primarily focuses on credit card users who can churn. Hence churning of credit card users is dealt with in this paper. The primary focus of this paper is on finding optimum method for predicting customer churn, as well as finding alternative methods to optimize the algorithm. To achieve this, a few machine learning algorithms like XGBoost, CatBoost, AdaBoost, Random Forest, K-Near Neighbor, Decision Tree, and Logistic Regression have been analyzed. After analysis the algorithm that produces the best performance is selected and optimized for higher accuracy and R^2 score.

2. Related Work

A short summary of prior studies and works on customer churning behavior proposed by researchers is presented in the following section. Adbelrahim et al. [2] used various algorithms such as decision tree, GBM (Gradient Boosting Machine) tree algorithm, random forest and XGBoost for customer churn. It is found that XGBoost performed superior compared to others in comparative analysis. Although, it was suggested the model can be improved by optimization algorithm. Praveen et al. [3] in their research work havediscussed customer churn and delivered comparative analysis on customer churn. In order to carry out this research the authors have used Support Vector machine, decision tree, naïve bayes, and logistic regression. Further they used SVM-POLY using AdaBoost performed better than others. It has been suggested that the algorithm can be improved by implementing feature selection strategies. There are few main important actions involved in reversal of churn and retention of clients:

- Retention of existing customers reduces the need to acquire new customers which allows them to improve their relationships with the existing customers.
- Older customers tend to purchase more than the new customers, as they are familiar with the bank.
- Maintenance for older customers is less than new customers.

There are several algorithms to churn predictions. Those algorithms are used, where data used have classification type for the result. One such algorithm is Decision tree, which is widely used for solving classification type problems [4]. This algorithm helps to determine which data attributes have the greatest influence on prediction. The model can be improved by applying feature selection algorithms. K-nearest neighbor is defined as simple and efficient non-parametric classification by [5]. For classification, the data is retrieved and the k closest neighbor for that data point, from its neighborhood by similarity or distance metric is calculated. After identification certain functions (e.g., Majority or minority) are applied to classify the data points. By this we can conclude, k-nearest neighbor is highly sensitive to k parameter, which is recommended to be defined using cross validation. Linear regression does not have the ability to predict classification problems, therefore Logistic regression has been adapted to solve that problem. Therefore, logistic regression is an extension of linear regression. Since bank churn is a binary classification, [6] [7] the authors have opted for logistic regression to handle classification attributes. To classify, there are three following steps: (i) map the result of linear regression in [0,1], using sigmoid function. (ii) interpret the result and classify as 0 or 1. (iii) and predict 1 if probability is higher than threshold value (i.e. 0.5); else 0. In [8] authors have introduced the Random Forest method. Random Forest selects subset of attributes randomly and cultivates sample of training sets for each tree [9]. The method’s concerning disadvantage is computational time for constructing trees, which is directly proportional to number of trees as explained by [6]. Although the mentioned disadvantage, a greater number of trees are required better the predictions [10]. In [11] the authors have applied various classification techniques to predict customer churn and concluded that Random Forest outperformed other algorithms. But the algorithm can further be improved for feature extraction. [12] uses deep learning models and 10 overlap cross validation methods for prediction. They yielded AUC (Area Under the Curve) score of 0.89. In [13-19] the authors have presented a comparative analysis by using popular classification approaches. They selected algorithms namely: Light GBM, XGBoost, Random Forest and Decision Tree and to obtain a high AUC score they used soft voting technique.

3. Methodology

In the algorithm, some of the features are used such as: Credit score, Balance, Tenure, Credit card, Age, Country. Once the features are selected, the data is normalized and some of the data is converted to binary data for better prediction. The Prediction model consists of three parts:

- Data cleaning and normalization (Smote-Enn).
- Implementation of Standard Scalar

- Implementation of XGBoost followed by implementation of Bayesian Optimization for hyperparameter tuning.

3.1. XGBoost

XGBoost is an extension of gradient-boosting decision trees. It is used to optimize the machine learning model. In XG Boost there are two types of Machine learning models

- Regressor
- Classification

3.2. Classification

The data first constructs a decision tree, and with the formula the results are in leaf nodes. The similarity score defines how close the score is to leaf node. As the tree is divided into branches, those branch's leaf node is close to the condition the tree is divided.

$$\text{Similarity Score} = \frac{\left(\sum \text{Residuals}\right)^2}{\sum^{N}[P\ (1 - P)] + \lambda}$$

P = Probability

The above equation is the Simple XGBoost similarity score formula. With this formula, the whole tree is generated. And with the image below the tree's result is taken for next iteration until the tree has achieved an optimal accuracy. The formula for Regressor is similar to Classifier. The only difference is in the calculation of the loss function. [20-25]

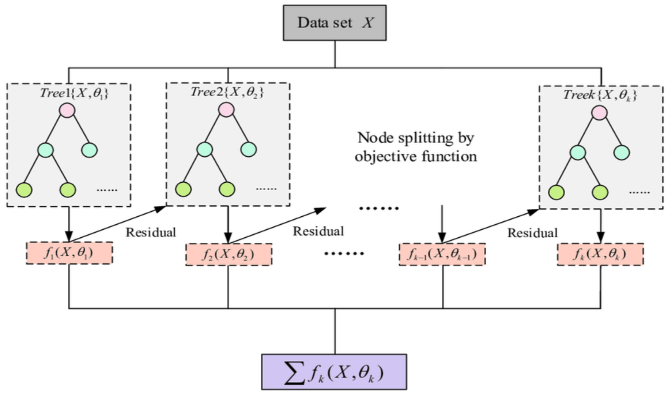

Figure 1 **XGBoost Architecture**

In order to find Residual value, the first decision tree decision tree must be generated and final output value must be calculated. The output of previous value will act as Residual value for next decision tree. The Similarity Score is a measurement parameter for decision tree nodes. Due to Similarity Score, we can determine the placement of nodes in a decision tree as shown in Figure 1. When this process is done 'n' times then, a tree is generated with maximum accuracy. Bayesian Optimization: Bayesian Optimization is an algorithm that build a probability model on Objective function which is used for selecting hyperparameter to validate with true predicted results. This algorithm comprises of Objective function, surrogate model, and acquisition function. [26]

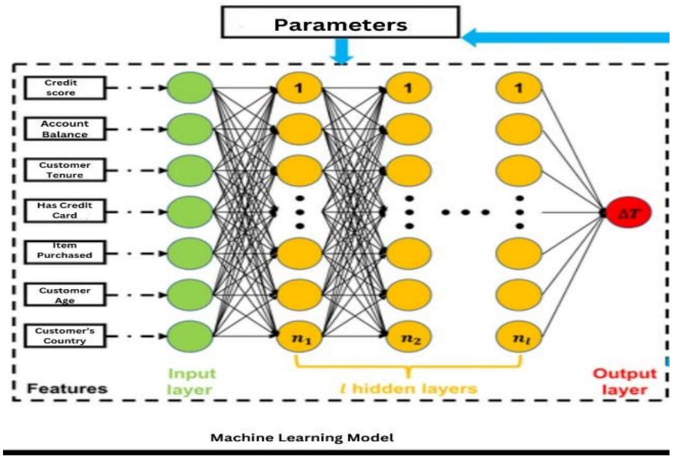

Figure 2 **Proposed Work**

- Objective function: It is a function to compute Root Mean Square Deviation (RMSE). Since Bayesian Optimization is the black-box method, it is assumed that information about RMSE is not available. Hence this function is responsible for finding it.
- Surrogate model: This function is the probability representation of the objective function. For this Gaussian Process is generally used. Instead of directly optimizing, Bayesian Optimization uses probabilistic surrogate model to approximates objective function.
- Acquisition function: Acquisition function is a utility function based on surrogate model. The acquisition function guides the selection of the next evaluation point by balancing exploration and exploitation.

4. Results and Discussion

While tuning XGBoost hyperparameter, two set of results are obtained. One is RandomSearchCV and another is modified Bayesian Optimization. In this section, comparison of RandomSearchCV and modified Bayesian Optimization is done. The accuracy and the time taken for tuning are considered as the primary parameters for comparison.

4.1. Comparison between RandomSearchCV and Modified Bayesian Optimization

The comparison between RandomSearchCV and modified Bayesian Optimization is as shown in Table 1. The RandomSearchCV uses random hyperparameter for tuning rather than learning from past. Bayesian Optimization's core function is surrogate function i.e., Gaussian Process. So, enhancing the surrogate function can impact the function of algorithm. As shown in Table 1, Modified Bayesian Optimization outperforms RandomSearchCV in terms of accuracy.

Table 1 Comparison with Random Search CV

Results	**Modified Bayesian Optimization**	**RandomSearchCV**
Accuracy	87.73%	85.9%
Time taken to complete tuning	4-5 minutes	1 minute 30 seconds

Table 2 Comparison with Bayesian Optimization

Results	Modified Bayesian Optimization	Bayesian Optimization
Accuracy	87%	87.73%
Time taken to complete tuning	6-7 minutes	4-5 minutes

4.2. Comparison between Bayesian Optimization and Modified Bayesian Optimization

The comparison between Bayesian Optimization and Modified Bayesian Optimization is carried out. As shown in Table 2, Modified Bayesian Optimization outperforms Bayesian Optimization in terms of accuracy. The proposed work is shown in figure 2.

Conclusion

To tune hyperparameter RandomSearchCV and Bayesian Optimization has been used by modifying Matern and alpha value. It has been observed from the results that the runtime is also reduced in addition to the increase in accuracy. Results have proved that modified Bayesian Optimization's accuracy increases and runtime decreases. In conclusion, the need for building a robust framework for bank churn predictor is revealed. The superior accuracy and efficiency gain of XGBoost with Bayesian Optimization provide a foundation for building a robust system. Equipping banks with this powerful tool empowers them to proactively address churn risk. Suggesting strategies can be further added to retain customer using Generative AI. The accuracy and Root Mean Square error can be improved for better precision.

References

[1] Roberts, "Developing new rules for new markets [J]," Journal of the Academy of Marketing Science, p. 31, 2000.

[2] Ahmad AK, Jafar A and A. K, "Customer churn prediction in telecom using machine learning in big data platform," Journal of Big Data, p. :28, 2019.

[3] Praveen and Asthana, "A comparison of machine learning techniques for customer churn prediction.," Internationa Journal of Pure and Applied Mathematics, pp. 1149-1169, 2018.

[4] Breslow LA and A. DW, "Simplifying decision trees: A survey," Knowl Eng Rev, pp. 1-40, 1997.

[5] Guo G, Wang H, Bell D, Bi Y and G. K, "Knn model-based approach in classification In OTM Confederated International

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 07 July 2024
Page No: 1909- 1914
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0261

Conferences” On the Move to Meaningful Internet Systems”," Springer, p. 986–996, 2003.

[6] Avon V, "Machine learning techniques for customer churn prediction in banking environments," Universita degli Studi di, Padova, Italy, 2016.

[7] Silva TC and Z. L, "Stochastic competitive learning in complex networks," IEEE Transactions Neural Netw Learn Syst, p. 385–398, 2012.

[8] L. Breiman, "Random forests," Mach Learn, pp. 5-32, 2001.

[9] Larivière B, V. den and P. D, "Predicting customer retention and profitability by using random forests and regression forests techniques," Expert Syst Appl, pp. 472-484, 2005.

[10] Liaw A and W. M, "Classification and regression by randomForest," R news, pp. 18-22, 2002.

[11] Y. Huang, F. Zhu, M. Yuan, K. Deng, Y. Li, B. Ni, W. Dai, Q. Yang and J. Zeng, "Telco churn prediction with big data.," Proceedings of the 2015 ACM SIGMOD international conference on management, pp. 607-618, 2015.

[12] R. Andrews, R. Zacharias, S. Antony and M. M. James, "Churn Prediction in Telecom Sector Using Machine Learning," International Journal of Information Systems and Computer Sciences, p. 8, 2019.

[13] X. Wang, K. Nguyen and B. P. Nguyen, "Churn Prediction using Ensemble Learning," In Proceedings of the 4th International Conference on Machine Learning and Soft Computing., 2020.

[14] U. P. Michele Gorgoglione, "Beyond Customer Churn," Journal of Intelligent Learning Systems and Applications, vol. 3, no. 2, pp. 90-102, 2011.

[15] A. P, "A comparison of machine learning techniques for customer churn prediction," International Journal of Pure and Applied Mathematics, vol. 119, no. 10, pp. 1149-1169, 2018.

[16] S. S. Dawes J., "Retention sans frontier: issues for retailer," International Journal of Bank Marketing, vol. 17, no. 1, pp. 36-43, 1999.

[17] K. V. d. P. D. Coussement, "Integrating the voice of customers through call center emails into a decision support system for churn prediction.," Information & Management, pp. 164-174, 2008.

[18] M. T. M. Bastan, "Customers Classification according to the Grey-Based Decision-Making Approach and Its Application to Bank Queue Systems," Asian Journal of Research in Banking and Finance, pp. 349-372, 2014.

[19] A. P, "A comparison of machine learning techniques for customer churn prediction.," International Journal of Pure and Applied Mathematics, pp. 1149-1169, 2018.

[20] A. A. and F. M. C., "E-Customization," Journal of marketing research, pp. 131-145, 2003.

[21] Abbasimehr H, Setak M and T. M, "A neuro-fuzzy classifier for customer churn prediction," International Journal of Computer Applications, vol. 19, no. 8, p. 35–41, 2011.

[22] A. Adomavicius and T. G., "Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Trans-actions on Knowledge and Data Engineering, pp. 734-749, 2005.

[23] Aarthi, G. and Karthikha, R. and Sankar, S. and Priya, S.S. and Jamal, D.N. and Banu, W.A,”Application of Machine Learning in Customer Services and E-commerce” Lecture Notes in Networks and Systems”,pp 817-832,2023

[24] B. Mobasher, R. Cooley and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," Commu-nications of the Association for Computing Machinery, pp. 142-151, 2000.

[25] Coussement K and D. B. KW, "Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning," Journal of Business Research, pp. 1629-1636, 2013.

[26] G. Adomavicius, Z. Huang and A. Tuzhilin, "Personal-ization and Recommender Systems," Tutorials in Opera-tions Research, INFORMS, Charlotte, 2008.