ICIC Express Letters
Volume **18**, Number **1**, January **2024**

ICIC International ©2024 ISSN 1881-803X
pp. **87–95**

CUSTOMER CHURN PREDICTION USING HISTOGRAM AUGMENTATION TECHNIQUE AND XGBOOST MODEL WITH BAYESIAN OPTIMIZATION

Doni Pradana and Prima Dewi Purnamasari*

Department of Electrical Engineering
Universitas Indonesia
Jl. Prof. DR. Ir R Roosseno, Kukusan, Beji, Depok City 16425, West Java, Indonesia
doni.pradana@ui.ac.id; *Corresponding author: prima.dp@ui.ac.id

Received January 2023; accepted April 2023

Abstract. *Customer churn is a significant issue in many sectors, such as the telecommunication sector. Therefore, telecommunication companies need to recognize churn risk as early as possible. Data from the IBM telco customer churn dataset was selected as a case study. One of the common challenges in classification problems is an imbalanced dataset, which will likely fail to predict the minority class. Oversampling with Histogram Augmentation Technique (HAT) is proposed in this study for handling the imbalanced class data. An ensemble learning of gradient boost machine learning techniques, namely XGBoost, was used in this study. In addition, we used Bayesian Optimization (BO) to find the best hyperparameter of the model. The experimental result shows that the accuracy of HAT-XGBoost-BO is 0.88 and the F1-score is 0.85, outperforming the XGBoost, HAT-XGBoost, and SMOTE-XGBoost models.*

Keywords: Customer churn, HAT, XGBoost, Augmentation, Bayesian optimization, Imbalanced dataset

1. **Introduction.** Competition among telecommunications companies is high these days. These companies compete to provide their best services and promotions to retain customers, so they stay connected and avoid customer churn. Customer churn is defined as clients who may be considering switching providers and could be migrated at any time, regardless of time commitments to service providers. One of the methods used by churn-prone telecommunication companies is to maintain Customer Relationship Management (CRM) [1]. However, discovering knowledge in large CRM databases, typically containing thousands or millions of customer information, is challenging and complicated [2]. Therefore, several telecommunication companies have implemented various statistical and Machine Learning (ML) algorithms on CRM database that is widely used to predict customer churn, such as logistic regression, Naïve Bayes [3], and further using Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) [4]. The current study focuses on ensemble learning in this regard. Ensemble learning is a progressive machine learning paradigm that applies multiple models in a process known as bagging, boosting, and stacking [5]. Previous studies have shown that ensemble learning suits predictive models with classification and regression problems for churn prediction, such as random forest, decision tree, random tree [6], and XGBoost [7].

This study aimed at customer churn prediction using a development model based on ensemble learning, namely XGBoost. The most crucial factor behind the success of ensemble learning using XGBoost is scalability in all scenarios [8] due to its multiple decision trees, linear regression problem [9], and gradient descent optimization [10]. Then, to overcome the imbalanced class challenge [11,12], we propose the Histogram Augmentation

DOI: 10.24507/icicel.18.01.87

Technique (HAT) algorithms [14] that use the distribution of the original tabular data for augmenting the new data. This augmentation technique will be compared to the more common Synthetic Minority Oversampling Technique (SMOTE) [13]. Due to the algorithm's intricacy and many parameters, adjusting the hyperparameters in XGBoost can be challenging. Thus, we proposed a new type of evolutionary algorithm, namely Bayesian Optimization (BO) [15], which uses the probability model of the Bayesian network to assemble the direct relationship between the data set and the learning result [16]. The contributions of this study are summarized as follows: 1) We implemented histogram augmentation technique for handling imbalanced dataset, and this is the first time it has been used in a customer churn dataset; 2) We implemented Bayesian optimization on the XGBoost classifier model to find out the best learning_rate, max_depth, and n_estimator hyperparameter for reaching out a higher accuracy. From the experimental results, our proposed model outperforms other classification models. This paper is structured as follows. The methodology is listed in Section 2 of this paper. Section 3 describes the experimental result, and the discussion is explained. Then, the conclusion is presented in Section 4.

2. **Methodology.** In this section, we present the methodology used in this study. Machine learning techniques are applied for customer churn prediction, as illustrated in Figure 1. These experiments started with data collection, data preprocessing, data augmentation, data split, training the data, which contained hyperparameter optimization, and evaluating each model.

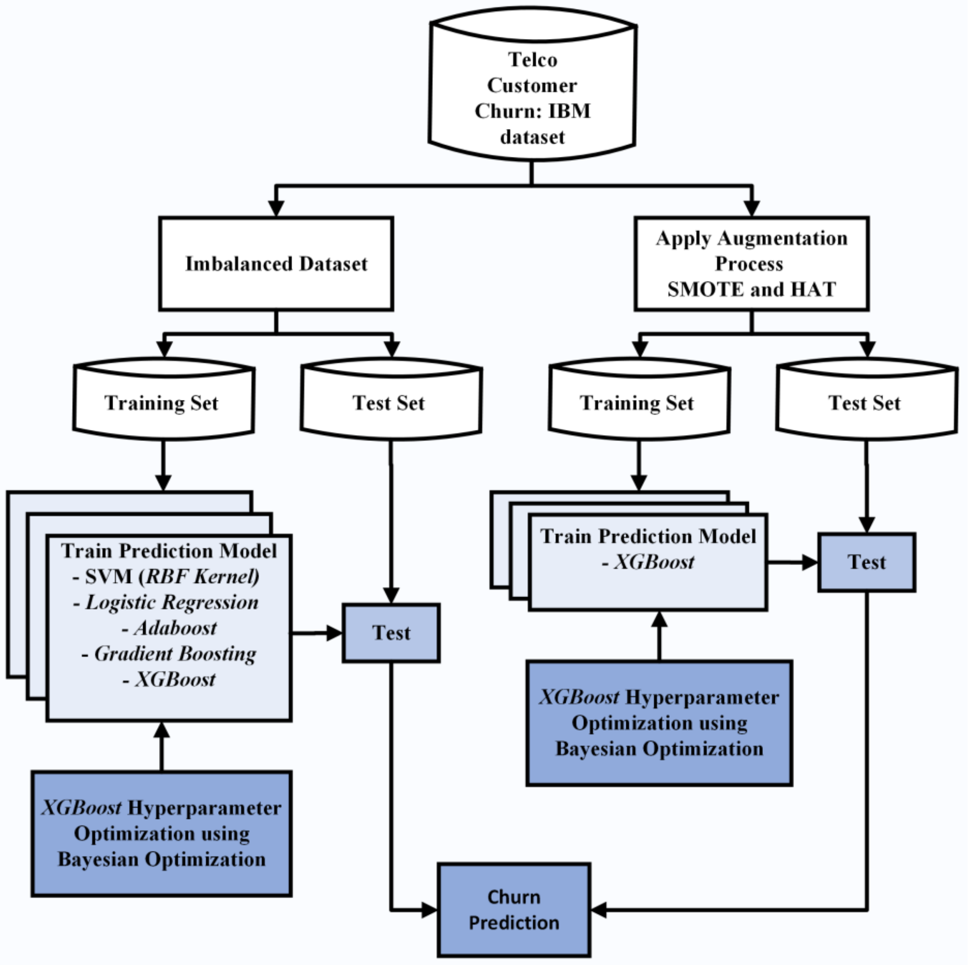


Figure 1. The study methodology

2.1. **Data collection.** This study is conducted on the publicly available dataset from the web of Kaggle data science competition, namely the Telco Costumer Churn: IBM dataset. The dataset has 7043 instances, 18 features, and target values. The dataset feature represented each customer's behavior and then the target value was flagged as 1 = churn and 0 = non-churn, of which 26.5% are churn customers, and there is a data imbalance. The dataset contains the following feature $F = \{$*Gender*, *Senior Citizen*, *Partner*, *Dependents*, *Tenure Months*, *Phone Service*, *Multiple Lines*, *Internet Service*, *Online Security*, *Online Backup*, *Device Protection*, *Tech Support*, *Streaming TV*, *Streaming Movies*, *Contract*, *Paperless Billing*, *Payment Method*, *Monthly Charges*, *Total Charges*$\}$, where *Churn Label* represents the class.

2.2. **Data processing.** After getting the dataset, the next step is to process the data to suit our needs, including the handling of missing values and the data labeling process. This dataset has 11 missing values, particularly in the total charges attribute. A drop data value and feature labeling process [17] will be carried out, so it becomes 7032 (5163 data = non-churn customers and 1869 data = churn customers) behavior customer data and categorical data has changed into numerical data (i.e., 1, 2, 3). This method to avoid machine learning will go through a faltering process.

2.3. **Data augmentation.** Imbalanced datasets refer to datasets whose numbers of samples in each class are not even [18]. This study uses the SMOTE and HAT to manage the class imbalanced dataset. SMOTE algorithm increases the number of data instances by generating random data of minority class feature space from its nearest neighbors using Euclidean distance. The default value of k is 5. For each randomly selected neighbor $\tilde{X}$, a new sample is constructed using Equation (1), where x stands for the original sample, $\tilde{X}$ represents the neighbor sample and X_{new} is the synthetic sample.

$$X_{new} = x + random(0,1) \times \left(\tilde{X} - x\right) \tag{1}$$

On the other hand, HAT generates data based on the distribution of the original tabular data with analysis of particular features and the feature's type [14], continuous or discrete. In the actual case, a continuous feature takes any value between a specific interval, and a discrete feature consists of a particular set of values; therefore, two different histogram augmentation algorithms are obtained: continuous feature augmentation and discrete feature augmentation. The histogram generation of continuous features gives the frequencies of the corresponding bins calculation using Equation (2), which is called Doane's rule [19]. The calculation is applied because each histogram bin must be of equal width, and adjacent classes have a typical limits range of values. Each bin's mid-point is considered and mapped to its corresponding frequency. Given that n is the number of elements, X_i is an element in the set and $\bar{X}$ is the mean of the set, k is the number of the bins and then, Doane proposes the number of k as

$$k = \log_2(n) + 1 + \log_2\left(1 + \frac{\sqrt{b}}{\sigma\sqrt{b}}\right); \quad \sqrt{b} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^{(3/2)}\right]};$$
$$\sigma\sqrt{b} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}} \tag{2}$$

Next, the bin values and the intermediate values between each bin are selected for the augmentation process. The frequency of intermediate value between two pre-existing bins, is calculated as arithmetic mean frequencies of the pre-existing bins. Figure 2 represents the values selected for further processing and their corresponding calculated frequencies with the 'Monthly Charges' continuous data example. The green point with a solid line

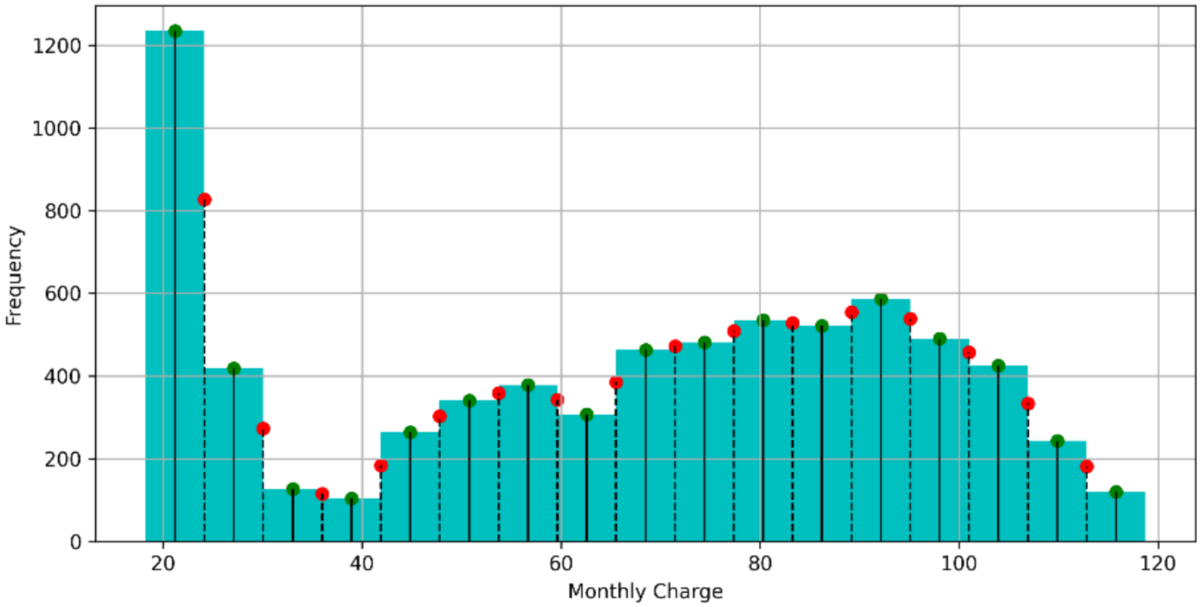


FIGURE 2. The selected values on mid-point bins and pre-existing bins

represents mid-point values, and the green point with a dashed line represents intermediate values between pre-existing bins. After that, the step of validity check of selected values occurred. The values obtained in the previous step are just an interpolation of values; hence, selecting values that follow the original data should be checked to determine whether they are eligible to form a new sample, and a validity check is performed. A uniform random distribution is used to create the random number if the number of new samples to be generated is less than twice the size of the original data. Then, if each sample passes the validity check mechanism, it is appended to the newly augmented list. Finally, the above systematic process will take place until the number of samples required to balance the dataset is attained. The same rule does not apply to discrete feature augmentation because individual column properties can only take a specific set of values. Hence, proportional sampling is applied. The continuous features in this dataset are "monthly charges", "tenure months", and "total charges" and then the discrete feature in this dataset includes all features except three features, as mentioned previously.

2.4. **Data splitting.** After the class data has been balanced in the previous augmentation process, the data will be divided into 80% train set and 20% test set for training and testing a prediction model. This method was also applied in the imbalance class dataset. The result of the augmentation sampling method and splitting data set has been detailed in Table 1.

TABLE 1. Augmentation process and data splitting result

Data	Train set		Test set	
	Non-Churn	Churn	Non-Churn	Churn
Imbalance	4151	1474	1012	395
SMOTE	4130	4130	1033	1033
HAT	4130	4130	1033	1033

2.5. **Bayesian optimization.** The Bayesian Optimization (BO) can be mentioned as a parametric optimization algorithm based on the Gaussian process and Bayes theorem [20] to find the global optimal of the machine learning function. The global optimal is obtained by iterating, constructing a probabilistic model for the optimized machine learning function, and then using this model's score to select the next collection point. Two core modules of Bayesian optimization are performed: prior function (in this study, implemented by Gaussian process) and acquisition function. Based on the idea of Bayesian optimization, this study first roughly determines the range of optional parameters. After that, it builds the XGBoost prediction model and continuously trains the model using the Bayesian optimization strategy. The Gaussian process establishes distributions over the XGBoost prediction model. Thus, the acquisition function is used to obtain the global optimal solution of the machine learning function. In the XGBoost prediction model, hyperparameters n_estimetors and learning_rate influence the prediction accuracy and the overall complexity of the model, and max_depth configuration changes the characters' weak evaluator. Table 2 explains Bayesian optimization parameter settings for the XGBoost prediction model.

Table 2. Bayesian optimization parameter settings for the XGBoost prediction model

Optimization	Learning rate	Max_depth	N_estimators
Bayesian search	$[0.23, 0.25, 0.33, 0.35]$	$[2, 4, 6, 8]$	$[60, 80, 100, 120]$

2.6. **Prediction model and evaluation.** Then the final stage is the training process and the evaluation of the model. This study performs seven training prediction models using default parameters: SVM (RBF kernel), logistic regression, AdaBoost, Gradient Boosting, XGBoost, SMOTE-XGBoost, and HAT-XGBoost. Then, we tried training the XGBoost, SMOTE-XGBoost, and HAT-XGBoost using the Bayesian optimization process. The (XGBoost) prediction model is a highly scalable boosting algorithm. This supervised learning algorithm attempts to predict a target variable accurately by combining an ensemble of estimates from a set of more superficial and weak models. The algorithm performs well and has high scalability due to the robust handling of various data types, relationships, distributions, and the variety of hyperparameters that can be fine-tuned.

After the model training process is conducted, the results will be evaluated using testing data split previously in the original dataset and each augmentation data process result (SMOTE and HAT). Four standard performance metrics will be used: accuracy, precision, recall, and F1-score. Besides that, we also compare the result using the Area under the Curve (AUC), which is a prevalent metric used in the prediction model, representing the area under Receiver Operating Characteristic (ROC) or the Precision-Recall (PR) curve. The PR curve will be illustrated the trade-off between precision and recall. However, the PR curve focuses on a positive class, so AUC must be calculated from the ROC curve to measure both the class detection. The ROC curve plots the TPR (True Positive Rate) against the FPR (False Positive Rate) at various thresholds of the model.

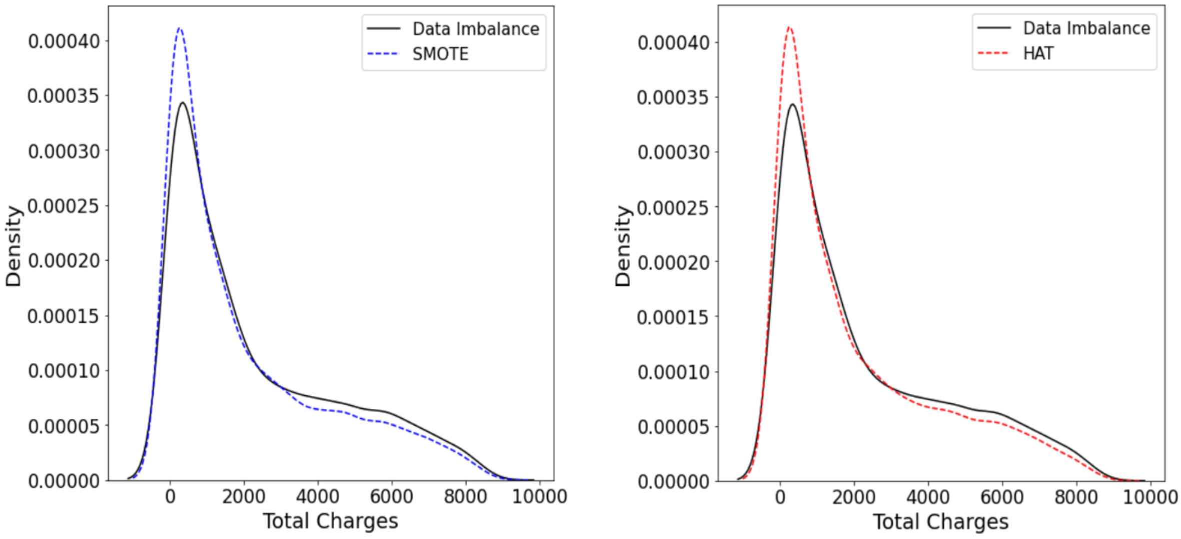
3. **Experimental Results and Discussion.** This section will present the results obtained for customer churn prediction. The experimental result of this study will be divided into three evaluation sections according to the uses of the imbalanced dataset, SMOTE, and HAT. The first prediction evaluation used SVM (RBF kernel), logistic regression, AdaBoost, Gradient Boosting, XGBoost, and XGBoost-BO. In the first prediction evaluation, the imbalanced original dataset was performed. Table 3 shows that the first prediction evaluation has a training accuracy of around 0.81 to 0.84, except for the SVM (RBF kernel), which has a training accuracy of only 0.73. For the test accuracy metric, the results are around 0.79 to 0.80. Then, the SVM (RBF kernel) test accuracy is still the lowest result, only 0.73. Other metrics, such as recall, precision, and F1-score, show that SVM (RBF kernel) has the lowest effects with 0.50, 0.37, and 0.42. Since the data

Table 3. Classifiers train and test results

Prediction model	Train accuracy	Test accuracy	Recall	Precision	F1-score
SVM (RBF kernel)	0.73	0.73	0.50	0.37	0.42
Logistic Regression	0.81	0.80	0.73	0.74	0.74
AdaBoost	0.82	0.80	0.74	0.74	0.74
Gradient Boosting	0.83	0.80	0.72	0.75	0.73
XGBoost	0.83	0.79	0.71	0.73	0.72
XGBoost-BO	0.84	0.80	0.72	0.74	0.73
SMOTE-XGBoost	0.85	0.82	0.82	0.82	0.82
SMOTE-XGBoost-BO	0.86	0.82	0.82	0.83	0.82
HAT-XGBoost	0.86	0.83	0.83	0.83	0.83
HAT-XGBoost-BO	**0.88**	**0.85**	**0.85**	**0.85**	**0.85**

set is imbalanced, we prefer to use the F1-score metrics rather than accuracy because the accuracy metrics will be biased in imbalance classification. Logistic regression and AdaBoost prediction models are much better, with an F1-score of 0.74, although still below the expected score. At that point, the first prediction evaluation shows that SVM (RBF kernel) has the lowest metrics score. Next, the second prediction evaluation based on Table 3 shows a better metrics score than all the prediction models in the first prediction evaluation. For train accuracy, the SMOTE-XGBoost prediction model scored 0.85, which is lower than the SMOTE-XGBoost-BO prediction model, which has a score of 0.86. In both prediction models, other metrics score (test accuracy, recall, precision, and F1-score), the results are around 0.82 to 0.83. The third prediction evaluation shows that the training accuracy of the HAT-XGBoost prediction model scored 0.86, which is lower than the HAT-XGBoost-BO prediction model, which scored 0.88. Then other metrics scored results in the test accuracy, recall, precision, and F1-score with the same results, scoring 0.83 (HAT-XGBoost) and 0.85 (HAT-XGBoost-BO).

Figure 3 represents the distribution of the “Total Charges” feature from the dataset after applying the SMOTE (Figure 3(a)) and HAT (Figure 3(b)). This distribution shows that both augmentation methods give the same results, almost imitating the original dataset distribution after augmentation. It can be mentioned that no augmentation technique can replicate a dataset distribution exactly [21]. However, ensemble machine learning models, such as XGBoost, are insensitive to distribution data because they are based on the decision tree-boost principle that uses a sampling technique [22]. These reasons do not affect the prediction results of the SMOTE-XGBoost and HAT-XGBoost models. Figure 4 of ROC-AUC results shows that the XGBoost model using an imbalanced dataset has a lower score than using augmented data. The HAT-XGBoost prediction model provides outperformed results when compared to SMOTE-XGBoost, with an AUC score of 0.9281. In addition, Bayesian optimization improves the AUC score of HAT-XGBoost, with an AUC score of 0.9329. Then, the HAT-XGBoost-BO prediction model outperformed all prediction models in this study. Bayesian optimization parameter input model can effectively improve the accuracy, F1-score, and ROC-AUC value of all XGBoost prediction models. It is because the objective function values of all XGBoost prediction models fitted by the Bayesian optimization algorithm are closer to the actual values. Another critical factor in augmentation is the execution time. Therefore, it is necessary to estimate the complexity of each model to provide exemplary performance in terms of time. SMOTE has the least time complexity of HAT because this augmented model is based on N-dimensional vector space, where N is the number of features. As a result, this augmentation model can produce results instantaneously. The HAT employs continuous or discrete augmentation depending on the feature type. Whereas the discrete distribution



(a) SMOTE augmentation (b) HAT augmentation

Figure 3. Distribution of the "Total Charges" feature from the dataset after applying the augmentation process

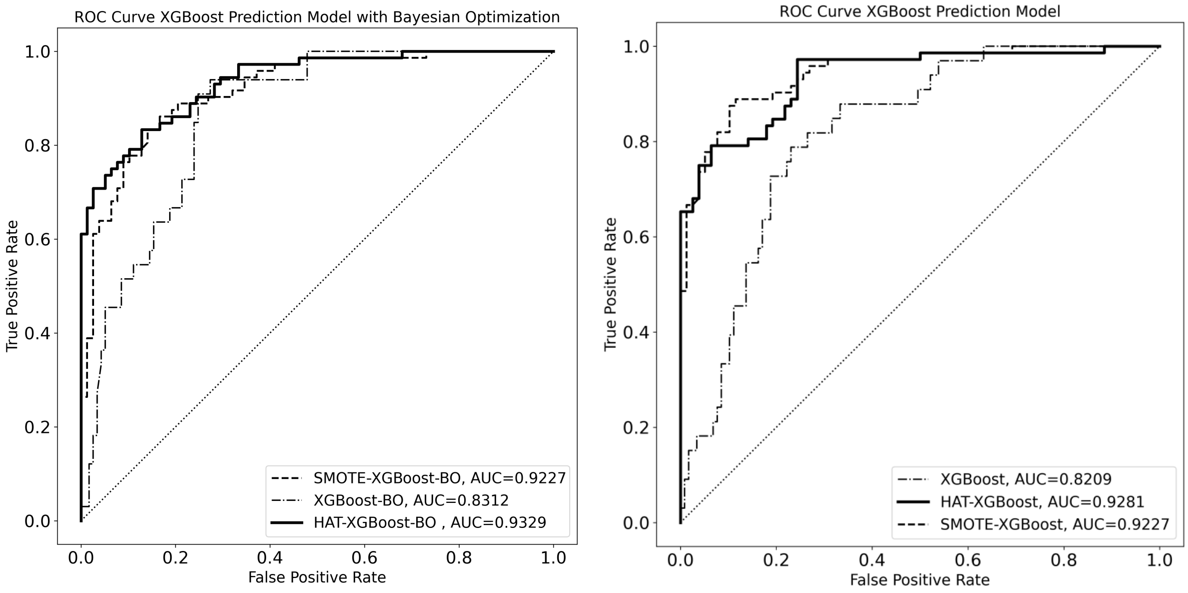


Figure 4. XGBoost prediction model ROC curve

feature uses proportional sampling, it has augmented instantaneously. On the other hand, the continuous distribution takes longer to augment data since applying the calculations of bin size uses Doane's rules, data selected process, validity check, and iteration process until the number of samples required to balance the dataset is attained.

4. **Conclusions.** In this paper, we have applied the proposed customer churn prediction to the IBM customer churn dataset using the XGBoost prediction model with data imbalanced class problems. This study applies two augmentation processes, i.e., SMOTE and HAT. We also proposed several ML for comparing churn prediction models: SVM (RBF kernel), logistic regression, AdaBoost, and Gradient Boosting. The result shows that SVM (RBF kernel) prediction has the lowest metrics score, and commonly XGBoost prediction model gives the best result in this study. In other results, the HAT-XGBoost model outperforms XGBoost and SMOTE-XGBoost. The Bayesian optimization gives the

optimal solution for increasing accuracy, F1-score, and ROC-AUC metrics. After that, data distribution after applying HAT and SMOTE has the same result, almost imitating the distribution of the original dataset. This augmentation technique can be used to imbalance class dataset problems. However, the fact is that the distribution of data on the XGBoost model is insensitive because this model uses a sampling technique. From the time complexity perspective, the HAT augmentation technique gives a longer time than SMOTE. In the future, it is interesting to learn how this method performs in multiclass classification and regression problems.

Acknowledgment. The authors would like to thank Universitas Indonesia for supporting the research under the Hibah PUTI Pascasarjana 2023 with grant number NKB-273/UN2.RST/HKP.05.00/2023.

REFERENCES

[1] J. Burez and D. Van den Poel, Handling class imbalance in customer churn prediction, *Expert Syst. Appl.*, vol.36, no.3, pp.4626-4636, 2009.

[2] A. Amin et al., Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study, *IEEE Access*, vol.4, pp.7940-7957, 2016.

[3] S. Wu et al., Integrated churn prediction and customer segmentation framework for telco business, *IEEE Access*, vol.9, pp.62118-62136, 2021.

[4] S. S. Raj et al., Customer attrition classification and EDA on IBM telecommunication dataset based on machine learning algorithms, *International Research Journal of Engineering and Technology*, vol.7, no.5, pp.902-911, 2020.

[5] H. Jein et al., Churn prediction in telecommunication using logistic regression and Logit Boost, *Procedia Computer Science*, vol.167, pp.101-112, 2020.

[6] I. Ullah et al., A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector, *IEEE Access*, vol.7, pp.60134-60149, 2019.

[7] P. Senthan et al., Development of churn prediction model using XGBoost – Telecommunication industry in Sri Lanka, *IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, Toronto, ON, Canada, pp.1-7, 2021.

[8] T. Chen and G. Carlos, XGBoost: A scalable tree boosting system, *Proc. of the 22nd ACM SIGKDD*, New York, USA, pp.785-794, 2016.

[9] Y. Yang, Market forecast using XGBoost and hyperparameters optimized by TPE, *2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID)*, pp.7-10, 2021.

[10] Y. Niu, Walmart sales forecasting using XGBoost algorithm and feature engineering, *International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, pp.458-461, 2020.

[11] J. Wang et al., Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding, *Proc. of the 8th Int. Conf. Signal Process.*, vol.3, pp.1-4, 2006.

[12] A. Luque et al., The impact of class imbalance in classification performance metrics based on the binary confusion matrix, *Pattern Recognition*, vol.91, pp.216-231, 2019.

[13] V. Nitesh et al., SMOTE: Synthetic minority oversampling technique, *Journal of Artificial Intelligence Research*, vol.16, pp.312-357, 2002.

[14] S. Balachander, S. Yogesh et al., Feature-based augmentation and classification for tabular data, *CAAI Transactions on Intelligence Technology*, 2022.

[15] Y. Xia et al., A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring, *Expert Systems with Applications*, vol.78, pp.225-241, 2017.

[16] L. Sun, Application and improvement of XGBoost algorithm based on multiple parameter optimization strategy, *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, pp.1822-1825, 2020.

[17] K. Ebrah and E. Selma, Churn prediction using machine learning and recommendations plans for telecoms, *Journal of Computer and Communications*, vol.7, pp.33-53, 2019.

[18] X. P. Tan, Wireless sensor networks intrusion detection based on SMOTE and the random forest algorithm, *Sensors*, vol.19, no.1, 203, DOI: 10.3390/s19010203, 2019.

[19] D. P. Doane, Aesthetic frequency classifications, *The American Statistician*, vol.30, no.4, pp.181-183, 1976.

[20] R. Ma, Q. Xing, J. Zhang, J. Wang and Y. Wang, Logging interpretation method based on Bayesian optimization XGBoost, *2022 16th IEEE International Conference on Signal Processing (ICSP)*, pp.395-400, 2022.

[21] W. Buttijak et al., Comparison of methods to tackle class imbalance in binary classification for IoT applications, *The 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, Chiang Mai, Thailand, pp.115-120, 2020.

[22] D. Chakraborty and H. Elzarka, Early detection of faults in HVAC systems using an XGBoost model with a dynamic threshold, *Energy and Buildings*, vol.185, pp.326-344, 2019.