Scientific
Research
Publishing

# Research on Telecom Customer Churn Prediction Based on GA-XGBoost and SHAP

**Ke Peng, Yan Peng**

College of Computer Science and Engineering, Sichuan University of Science and Engineering, Yibin, China
Email: 695078611@qq.com, 2634932795@qq.com

## Abstract

To address the prominent problems faced by customer churn in telecom enterprise management, a telecom customer churn prediction model integrating GA-XGBoost and SHAP is proposed. By using the ADASYN algorithm for data processing on the unbalanced sample set; based on the GA-XGBoost model, the XGBoost algorithm is used to construct the telecom customer churn prediction model, and the hyperparameters of the model are optimized by using the genetic algorithm. The experimental results show that compared with traditional machine learning methods such as GBDT, decision tree, KNN and single XGBoost model, the improved XGBoost model has better performance in recall, F1 value and AUC value; the GA-XGBoost model is integrated with SHAP framework to analyze and explain the important features affecting telecom customer churn, which is more in line with the telecom industry to predict customer the actual situation of churn.

## Keywords

GA-XGBoost Algorithm, SHAP, Genetic Algorithm, Customer Churn, Prediction, Machine Learning

## 1. Introduction

With the booming development of information technology and mobile networks, the competition in the telecommunication industry is becoming increasingly severe. It is known that the cost of developing a new subscriber is 5 - 6 times higher than the cost of retaining an existing customer [1], and the choice of the subscriber determines the development of the company, therefore, the ability to successfully predict churn and effectively reduce it has become an important concern for the telecommunication industry. It is necessary for telecommunication companies to build a model that can accurately predict the tendency of cus-

tomers to churn in a certain period of time in the future and take appropriate measures to solve it.

In recent years, for the telecommunication customer churn prediction problem, scholars at home and abroad have been using data mining techniques to analyze and establish customer churn prediction models and apply classification algorithms to the field of telecommunication customer churn, which is of great practical significance for telecommunication companies to tap effective customers. In China, Qian *et al.* [2] improved the support vector machine by introducing a cost-sensitive function and used different penalty coefficients on the objective function to incorporate different misclassification costs into the modeling process for the study of telecommunication customer churn prediction. The literature [3] introduced kernel principal component analysis (KPCA) to customer churn prediction and proposes a corresponding feature extraction algorithm. The literature [4] gives a pruned random forest approach and proposes an effective random forest similarity matrix-based variance estimation based on the degree of variance of important factors affecting the stability of the combined classifier. Wang [5] used a real dataset of a provincial telecom operator as a research object, preprocessed the data and used PSO-BP neural network for customer churn prediction by data equalization based on AdaCost algorithm and feature selection based on Relief filtering method, and proved the feasibility of the algorithm for operator customer churn management work. In foreign countries, Veronikha Effendy *et al.* [6] used combined sampling and weighted random forest to deal with imbalanced data in customer churn prediction. Ammar A.Q. Ahmed [7] used a hybrid firefly algorithm based approach for churn prediction of large telecommunication data, using firefly algorithm for identifying optimal solutions and combining simulated annealing algorithm with firefly algorithm for optimization. Awnag *et al.* [8] proposed a regression based churn prediction model to identify customer churn by using multiple regression analysis, this technique uses customer feature data for analysis and provides good performance. Sara Tavassoli *et al.* [9] proposed three hybrid integrated classifiers based on bagging and boosting, the proposed method can be applied not only for customer churn prediction but also for any other binary classification algorithm application.

Although the above studies have contributed to telecommunication customer churn prediction, due to the complexity of machine learning models, there is no rational explanation of the model, and only the importance of a feature can be judged in terms of the importance of the features affecting customer churn, and it cannot explain how the feature affects the prediction results. In order to solve the above problems, this paper tries to establish a telecom customer churn prediction model that integrates SHAP and the improved XGBoost algorithm, and hyper-parameter tuning of the parameters of XGBoost model by genetic algorithm. SHAP can explain and analyze various factors affecting telecom customer churn and provide corresponding information for telecom companies to adopt corresponding policies.

## 2. Construction of Telecom Customer Churn Prediction Model

### 2.1. Overall Process of Customer Churn Prediction Model Construction

The overall flowchart of telecom customer churn prediction model construction is shown in Figure 1. In this paper, we use the common dataset of customer churn of a telecom company on Kaggle website, analyze and process the abnormal values and missing values in the dataset, extract features from the dataset, and use ADASYN algorithm to process the data imbalance problem, and then construct the telecom customer churn prediction model by XGBoost, decision tree, K-nearest neighbor, GBDT and LightBGM, extreme forest algorithm to construct the telecom customer churn prediction model, and obtain the optimal model by comparing the corresponding evaluation indexes, then use genetic algorithm to intelligently optimize the optimal model to obtain the final model, and finally use SHAP framework to interpret and analyze this customer churn prediction model.

### 2.2. XGBoost Algorithm

XGBoost (Extreme Gradient Boosting) is an algorithm based on GBDT, which was proposed by Chen in 2016 [10]. It is optimized for GBDT and can effectively handle the relationship between data. For example, it optimizes the loss function by using the second-order Taylor formula expansion to improve the computational accuracy, simplifies the model by using regular terms to avoid overfitting, and uses Blocks storage structure for parallel computation [11]. The XGBoost algorithm is shown in Equation (1).

$$\widehat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{1}$$

where $\widehat{y}_i$ denotes the predicted value of XGBoost output, $x_i$ denotes the $i$-th input sample, taking values in the range [1, K], $f_k$ denotes the $k$-th classification
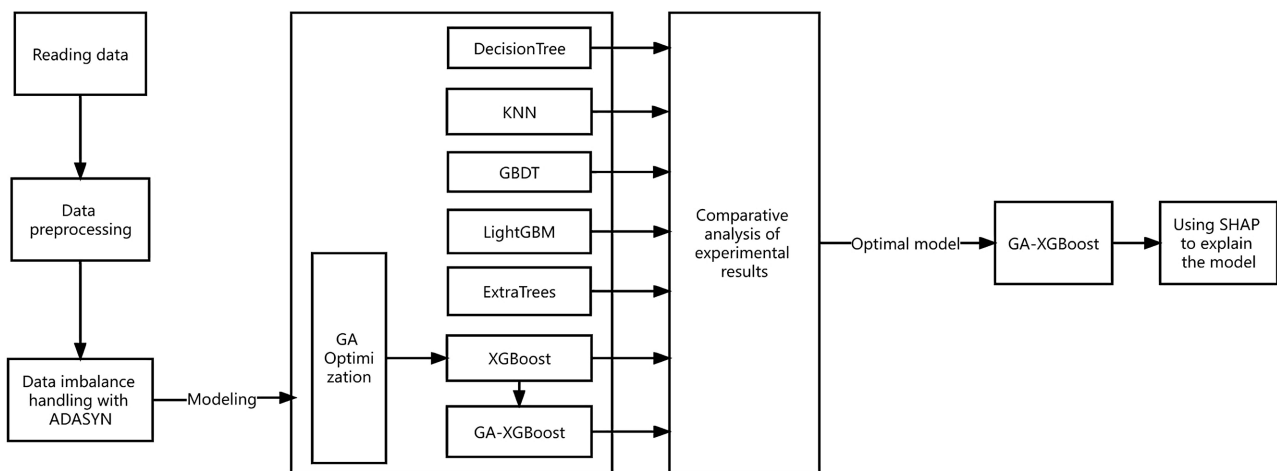


**Figure 1.** General flow chart of telecom customer churn prediction model.

regression tree, and $F$ denotes the classification regression tree space. The XGBoost objective function can be composed of a loss function and a regularization term. The specific objective function is shown in Equation (2), where $L(\varphi)$ is the loss function of the predicted value and the true value; $\Omega$ is the penalty to the model complexity, *i.e.*, the regularization term of the objective function, and the specific regularization term is shown in Equation (3); $Y$ and $\lambda$ denote the term coefficients of the regularization, and $T$ denotes the number of leaf nodes of the $k$-th tree.

$$L(\varphi) = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i\right) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

$$\Omega(f_k) = YT + \frac{1}{2}\lambda \|\omega\|^2 \tag{3}$$

Minimize the loss function of the optimized XGBoost with the regularization term, as shown in Equation (4).

$$L(\varphi)^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{4}$$

For the objective function use the second-order expansion of Taylor's formula and simplify it as shown in Equation (5), where $g_i$ is the first-order derivative of $x_i$ and $h_i$ is the second-order derivative of $x_i$.

$$
\begin{aligned}
L(\varphi)^{(t)} &\cong \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + YT + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \\
&\cong \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\omega_j^2\right] + YT
\end{aligned}
\tag{5}
$$

Derivative of $\omega$ and making the derivative zero, the $\omega$ minimizes the objective function can be found as shown in Equation (6).

$$\omega_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{6}$$

$L(\varphi)_{\min}$ is the minimum value of the objective function, the smaller the value, the better the tree model, corresponding to the minimum value of the objective function as shown in Equation (7).

$$L(\varphi)_{\min} = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + YT \tag{7}$$

Calculate the score of the split node of the tree model as shown in Equation (8).

$$\text{Gain} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - Y \tag{8}$$

## 2.3. Genetic Algorithm

Genetic Algorithm (GA) is a stochastic global optimization and search method that obtains the output optimal solution by inputting an objective function and constraints, drawing on Darwin's biological evolution and Mendel's genetic mechanism [12]. The genetic algorithm simulates the replication, crossover, and mutation occurring in natural selection and heredity according to the law of evolution of organisms in nature, and starts from any initial population, and through random selection, crossover, and mutation operations, produces a group of individuals better suited to the environment, evolves the population to a better region in the search space, and so on from one generation to another, and finally converges to a group of individuals best adapted to the environment, and to obtain the optimal solution of the problem [13]. As a relatively mature algorithm, genetic algorithm can be applied in the fields of function optimization, combinatorial optimization, and shop floor scheduling.

## 2.4. GA-XGBoost Algorithm

The XGBoost model suffers from many parameters, slow convergence, and large influence of parameters on model prediction results, while the traditional grid search for hyperparameter tuning suffers from low accuracy and long running time. Therefore, this paper proposes a GA-XGBoost model combining genetic algorithm and XGBoost model, and uses the global search capability of genetic algorithm to optimize the tuning parameter selection for XGBoost, and uses AUC as the fitness function to adjust the index.

The GA-XGBoost optimal hyperparameter combination is the optimal number of chromosomes output by the genetic algorithm when the number of iterations meets the termination requirement. For the telecom customer churn prediction model, three parameters, n_estimators, learning_rate, and max_depth, are optimized and initialized by the genetic algorithm. The optimal parameters are obtained by using the genetic algorithm to obtain a new generation of population, for which the parent population is replicated, crossed, and mutated, and then the best individuals are replaced by the worst individuals by calculating the fitness values of the offspring population. The GA-XGBoost algorithm is shown below:

---

GA-XGBoost algorithm

---

Input: Iterations per Time Step T, parameter number N, population size P, Number of excellent individuals M

Output: optimal paramrters

1. Initialize( $\theta_{i1}, \theta_{i2}, \cdots, \theta_{iN}$ )

2. while termination condition is not met do

3. AUC on the XGBoost model

4. Calculate fitness

5. Select the optimal parameters according to fitness value

6. GA(T, N, M, P)

7. Produce new parameters

8. end while

---

## 2.5. Interpretability of the Integrated SHAP Model

SHAP (Shapley additive explanations), an explanatory framework proposed by Lundberg, is an important approach from game theory [14]. For each prediction sample, the model produces a prediction value, and SHAP can calculate the Shapley value of each feature to reflect the contribution of each feature to the overall model prediction. SHAP interprets the prediction value of the model as the sum of the Shapley values of each input feature on the magnitude and direction of the role of individual features under the full sample, thus explaining the results of the model [15].

Suppose $x_i$ is the $i$-th sample, where the $j$-th feature of $x_i$ is $x_{ij}$, $y_i$ is the predicted value of the model for this sample, the mean value of the target variable of all samples, *i.e.*, the baseline value of SHAP is $y_{base}$, and $f\left(x_{ij}\right)$ is the SHAP value corresponding to $x_{ij}$, then the SHAP value is:

$$y_i = y_{base} + f\left(x_{i1}\right) + f\left(x_{i2}\right) + \cdots + f\left(x_{ik}\right) \tag{9}$$

The interpretability of machine learning algorithms is currently a hot topic in artificial intelligence research, and the GA-XGBoost model is poorly interpretable as a black-box model due to the high complexity of the integrated learning model. To solve the problem of poor interpretability of the GA-XGBoost model, the SHAP framework is introduced to interpret the results reliably, and the SHAP framework has powerful visualization functions and possesses the ability to display the interpreted results of model predictions, which is widely applied to interpret more complex classification and regression models [16]. Meanwhile, the traditional feature importance ranking can only determine the importance of a feature and does not explain how the feature affects the prediction results, while the greatest advantage of SHAP value is the ability to explain and analyze the degree of influence of each feature, and also to reflect the positive and negative influence of each feature.

## 3. Experimental Analysis

### 3.1. Experimental Data and Its Preprocessing

The experimental data in this paper originates from the data of a telecom company on the Kaggle platform. There are a total of 4025 customer samples in the data training set, of which each sample includes 20 feature attributes, consisting of several dimensions of label information affecting customer churn characteristics and whether the user is finally lost, and the basic information of the dataset is shown in Table 1. The analysis reveals that the dataset is seriously unbalanced, in which there are 589 pieces of data of customer samples that have been lost and 3652 pieces of data of customer samples that have not been lost, as shown in Figure 2.

For continuous features in the dataset, the features are normalized by removing the mean and scaling variance; for discrete features in the dataset, one-hot is used for solo-hot coding when there is no size relationship, and numerical mapping is used when there is a size correlation between attributes.

Table 1. Basic information of the dataset.

| Symbols | Description |
| --- | --- |
| *State* | 2-letter code of the US state of customer residence |
| *Account_length* | Number of months the customer has been with the current telco provider |
| *Area_code* | string = "area_code_AAA" where AAA = 3 digit area code |
| *International_plan* | The customer has international plan |
| *Voice_mail_plan* | The customer has voice mail plan |
| *Number_vmail_messages* | Number of voice-mail messages |
| *Total_day_minutes* | Total minutes of day calls |
| *Total_day_calls* | Total number of day calls |
| *Total_day_charge* | Total charge of day calls |
| *Total_eve_minutes* | Total minutes of evening calls |
| *Total_eve_calls* | Total number of evening calls |
| *Total_eve_charge* | Total charge of evening calls |
| *Total_night_minutes* | Total minutes of night calls |
| *Total_night_calls* | Total number of night calls |
| *Tal_night_charge* | Total charge of night calls |
| *Total_intl_minutes* | Total minutes of international calls |
| *Total_intl_calls* | Total number of international calls |
| *Total_intl_charge* | Total charge of international calls |
| *Number_customer_service_calls* | Number of calls to customer service |

## Label Distribution Map
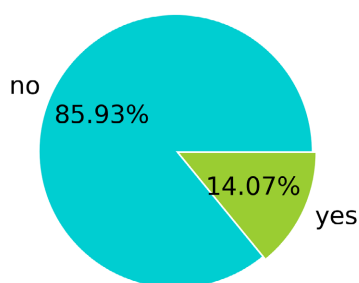


Figure 2. Distribution diagram of customer churn labels.

Due to data imbalance, this experiment uses the ADASYN method for over-sampling, which is a modified version of SMOTE, by assigning different weights to samples from different minority classes and thus obtaining different numbers of samples, while taking into account the majority of the sampled samples and setting an adjusted hyperparameter d, *i.e.*, a threshold. The distribution of the data is improved by generating synthetic samples to reduce the bias caused by

class imbalance and by adaptively adjusting the classification boundaries to difficult instances.

## 3.2. Evaluation Index

In the field of telecom customer churn, the learning metrics for assessing telecom customer churn were not limited to Accuracy for assessment, but were selected to focus on positive examples, where Accuracy, Recall, Precision, F1-score metrics were derived from the confusion matrix, as shown in Table 2.

The evaluation metrics used in the telecom customer churn prediction problem are accuracy, precision, recall, F1-score, and AUC. The specific formulas for the above evaluation metrics are shown in the following order.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{F1-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{13}$$

## 3.3. Experimental Results and Analysis

The paper selected six classification models, XGBoost, LightGBM, DecisionTree, KNN, GDBT, and ExtraTrees for comparison, and selected Accuracy, AUC, Recall, Precision, F1-score as the evaluation indexes of the models, and the ROC comparison results of each classification model are shown in Figure 3.

After the experimental comparison, it was found that XGBoost reached 0.9477 in terms of AUC, which is significantly higher than other models, and therefore is the optimal model. The iterative results of the genetic algorithm for hyperparameter tuning are shown in Figure 4. The final approximate optimal solution is n_estimators = 55, learning_rate = 0.18, max_depth = 4.

The XGBoost model with hyperparameter tuning by genetic algorithm is defined as GA-XGBoost, where the confusion matrix of GA-XGBoost model is shown in Figure 5, and the ROC comparison between GA-XGBoost and XGBoost is shown in Figure 6, from which it can be seen that GA-XGBoost improves 1% in AUC compared to XGBoost model and the AUC value is basically above 0.90.

Table 2. Confusion matrix.

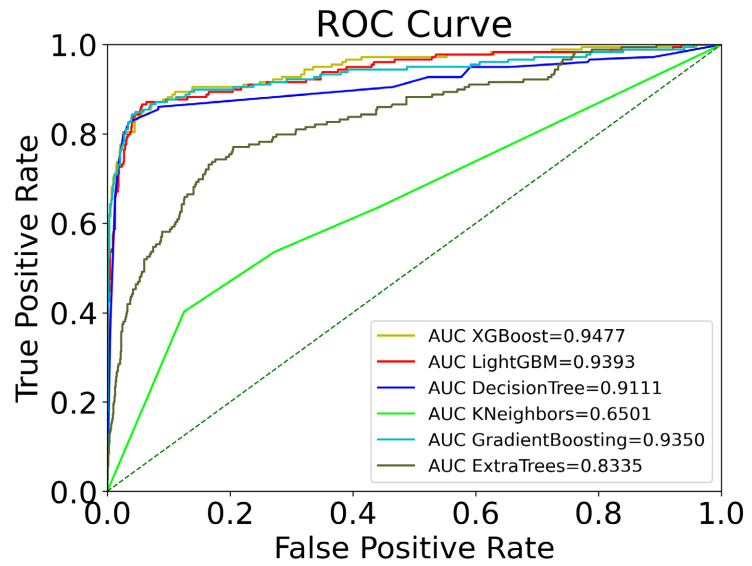| Actual value | Predictive value | |
|---|---|---|
| | Churn | Not Churn |
| Churn | TP | FN |
| Not Churn | FP | TN |

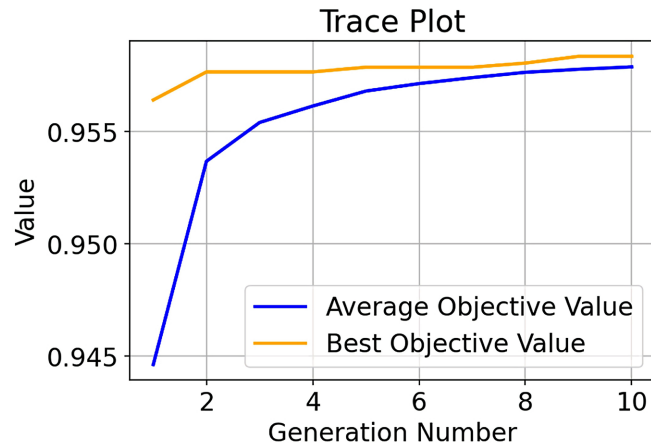**Figure 3.** Comparison results of each classification model.



**Figure 4.** Optimization iteration results of genetic algorithm.
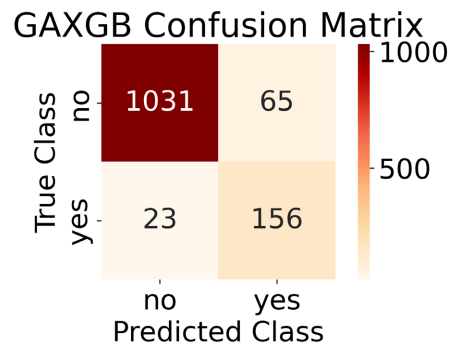


**Figure 5.** GA-XGBoost confusion matrix.

As can be seen from Table 3, the AUC values of KNN and ExtraTrees are relatively low and still need to be improved. The improved GA-XGBoost has an F1 value of 78%, an accuracy of 93.1%, and an AUC value of 95.61%, all of which
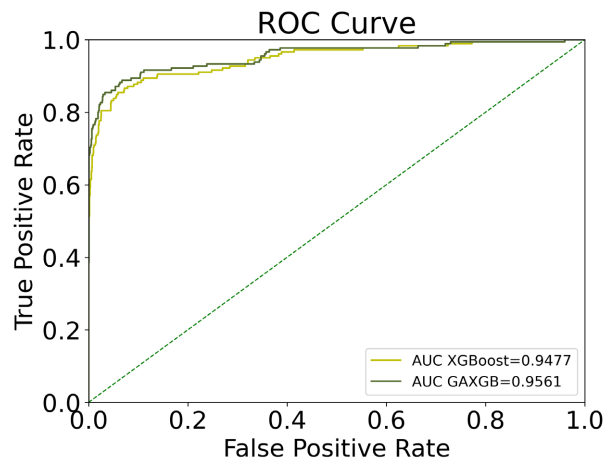
**Figure 6.** ROC comparison of GA-XGBoost and XGBoost.

**Table 3.** Comparison results of different model.

| Classifier | Precision | Recall | F1-score | Accuracy | AUC |
|---|---|---|---|---|---|
| XGBoost | 0.7037 | 0.8492 | 0.7696 | 0.9286 | 0.9477 |
| LightGBM | 0.5730 | 0.8771 | 0.6932 | 0.8910 | 0.9393 |
| DecisionTree | 0.9111 | 0.8547 | 0.7234 | 0.9082 | 0.9111 |
| KNN | 0.2430 | 0.5363 | 0.3345 | 0.7004 | 0.6501 |
| GBDT | 0.6446 | 0.8715 | 0.7411 | 0.9145 | 0.9350 |
| ExtraTrees | 0.2838 | 0.8212 | 0.4218 | 0.6839 | 0.8335 |
| **GA-XGBoost** | **0.7059** | **0.8715** | **0.7800** | **0.9310** | **0.9561** |

are slightly higher than the other six models, and the AUC values are basically above 0.90. The performance of the GA-XGBoost model proposed in this paper is obviously superior.

## 4. Explanatory Analysis of the Model Based on SHAP

SHAP feature importance, *i.e.*, the degree of contribution of each feature to enhance the overall model predictive power, provides a more direct representation of the degree of influence of the feature on the model. The focus of this chapter focuses on the explanatory analysis of model prediction results for telecommunication customer churn based on the SHAP framework [17].

Figure 7 shows the SHAP summary diagram of telecom customer churn features using the SHAP model, where the horizontal coordinate is the SHAP value, each line of the vertical coordinate represents a feature variable, and each point represents a sample, red represents a large feature value and blue represents a small feature value. The SHAP value is zero as the middle dividing line, for the left side of the dividing line is blue, the middle is purple and the right side is red, it means that the feature has positive contribution to customer churn; for the left side of the dividing line is red, the middle is purple and the right side is blue, it means that the feature has negative contribution to customer churn.
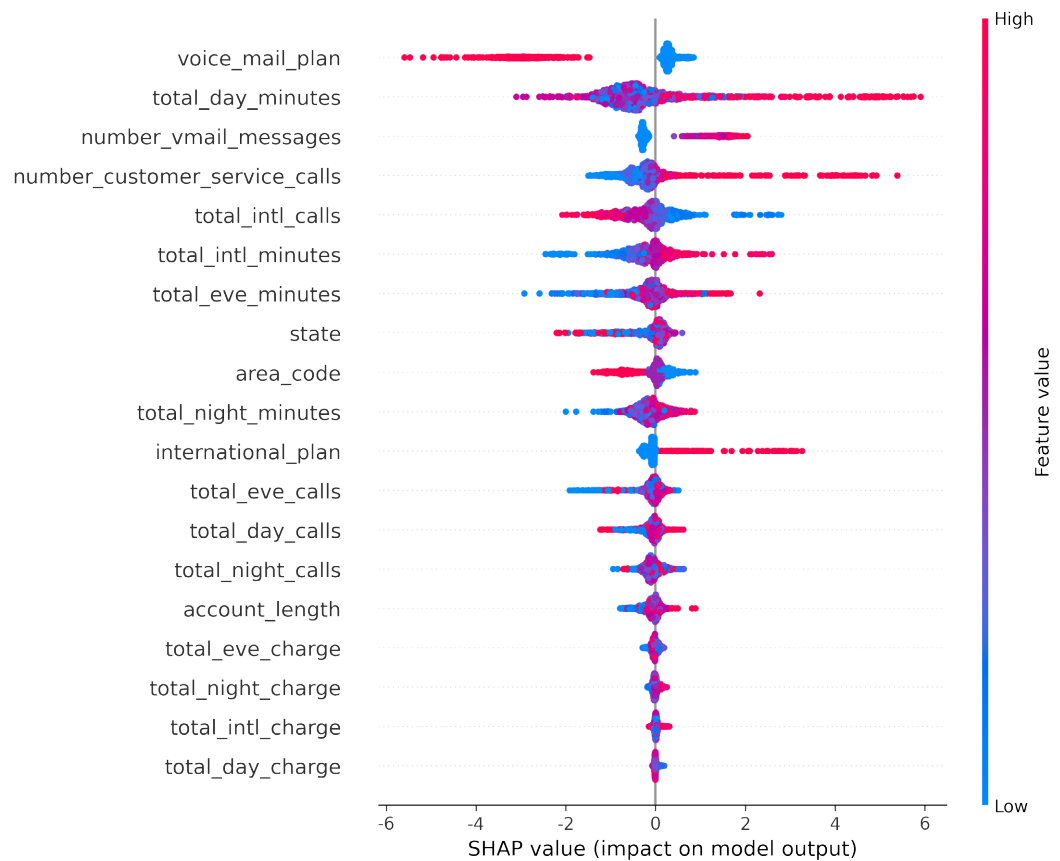
**Figure 7.** SHAP feature analysis summary diagram.

Through the SHAP values of some features in **Figure 7**, the analysis of their contribution shows that voice_mail_plan, total_intl_calls are more obviously negatively correlated, that is, the higher the value taken the lower the likelihood of model output churn. Customer voicemail plans are the most important feature of customer churn given by SHAP, and the higher number of customers subscribed to voicemail plans in the figure can determine that more effective customers can be retained if the telecom company increases the discount on voicemail plans. The higher number of international calls also reflects the customer's dependence and trust in the telecom company. The contribution of number_customer_service_calls, total_eve_minutes and total_intl_minutes has a significant positive correlation, *i.e.*, the higher the value, the higher the likelihood of model output churn. The number of calls to customer service reflects the current level of customer dissatisfaction with the telecom network, and the increase in total_intl_minutes increases the related tariffs and leads to the likelihood of customer churn, which is very helpful for the telecom industry to adjust and improve their sales policies and services.

In the SHAP feature analysis effect plot, base_value is the mean value of the telecom customer churn prediction model. The red range indicates that the feature contributes positively to telecom customer churn, and the blue range indicates that the feature contributes negatively to customer churn. **Figure 8** shows
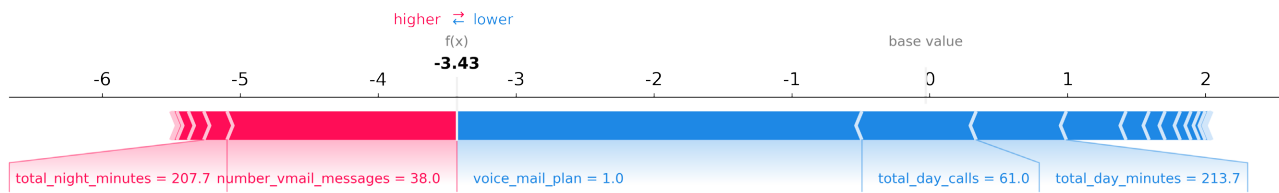
**Figure 8.** Predicted SHAP characteristic analysis effect diagram of non-churn customers.
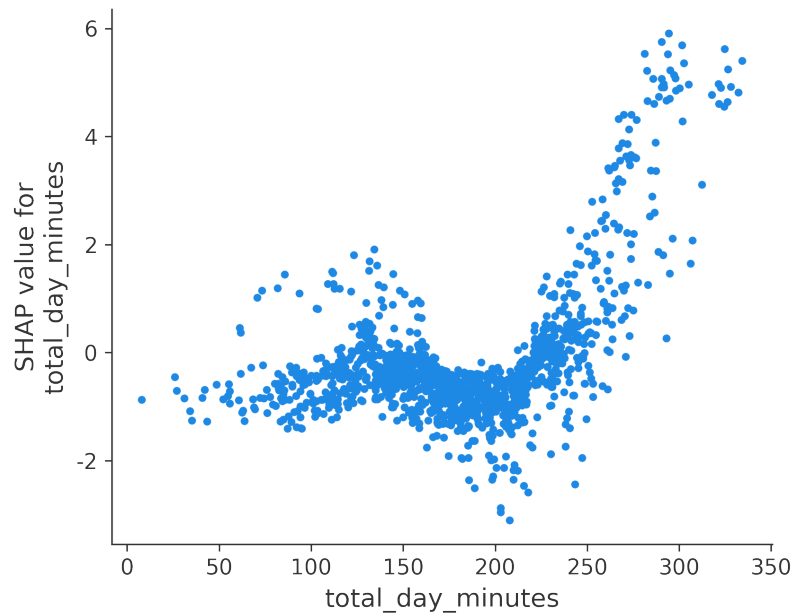


**Figure 9.** Total_day_minutes SHAP scatter plot.

the effect plot of SHAP feature analysis for a customer predicted to be a non-churning customer. The SHAP values for the number of voicemails and the total minutes of calls during the night are 38.0 and 207.7, respectively; the customer has a voicemail plan, the number of calls per day and the total minutes of calls per day are −1.0, −61.0 and −213.7, respectively. Finally, the model prediction for the first data in this test set is −3.43, which means that it is more likely to be non-churning and it is predicted to be non-churning customers due to reasons such as longer total minutes of calls in a day, more calls in a day, still using the voicemail plan, etc. Eventually, through the calculation of the self-attention layer, the short-term behavioral history sequence of the user is converted into a deeper representation that reflects the short-term interest mu t of the current user.

In the SHAP scatter plot in **Figure 9**, total_day_minites reflects the process of customer churn, showing a V-shaped trend of "up-down-up", *i.e.*, the total number of minutes of calls per day is in the range of 150 - 250 minutes, and the model output has a low churn rate. In this paper, we believe that these customers are more willing to continue to use the package of the telecom company because their usage time is relatively stable, while those who have more than 250 minutes a day may choose a more cost-effective telecom package due to their higher daily usage, *i.e.*, they are more likely to churn.

## 5. Conclusions

To address the problem of telecom customer churn prediction, a model combining genetic algorithm and XGBoost algorithm is proposed in this paper, and the SHAP framework is used to supplement the model interpretability. The factors affecting telecom customer churn are analyzed using the proposed method, and the main factors affecting telecom customer churn are identified as the length of calls per day, the number of calls per day, and whether to subscribe to a voicemail plan, etc. Through the analysis, it is found that customers with longer and more frequent daytime calls are easier to retain, and this group may be dependent on the telecom industry due to their work demands, and telecom companies can offer daytime call packages with appropriate discounts so as to retain more and more valuable users.

Although the GA-XGBoost algorithm proposed in this paper achieves better results in terms of evaluation indexes compared with traditional machine learning algorithms, there are still shortcomings, as the model uses genetic algorithm for parameter optimization, which leads to long running time in the process of finding optimal parameters. The next step will be to combine the feature selection method to filter the features of the model and remove the redundant features in order to reduce the prediction time of the telecom customer churn model, which will be the research direction after this topic.

## Acknowledgements

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Yu, S.B., Cao, J. and Kung, Z.W. (2012) A Review of Customer Churn Problem Research. *Computer Integrated Manufacturing Systems*, **18**, 2253-2263.

[2] Qian, S.L., He, J.M. and Wang, C.-L. (2007) An Improved Support Vector Machine Based Telecom Customer Churn Prediction Model. *Management Science*, No. 1, 54-58.

[3] Xia, G.Y. (2008) Customer Churn Prediction Based on Kernel Principal Component Analysis Feature Extraction. *Computer Applications*, No. 1, 149-151.

[4] Qiu, Y.H. (2014) Customer Churn Prediction Based on Pruned Random Forest in Telecommunication Industry. *Journal of Xiamen University* (*Natural Science Edition*), **53**, 817-823.

[5] Wang, Z.Y. (2019) Design and Implementation of Telecommunication Customer Churn Prediction Model Based on Particle Swarm Optimization Algorithm. University of Chinese Academy of Sciences (Institute of Artificial Intelligence, Chinese Academy of Sciences), Beijing.

[6] Effendy, V., *et al*. (2014) Handling Imbalanced Data in Customer Churn Prediction Using Combined Sampling and Weighted Random Forest. 2014 2*nd International Conference on Information and Communication Technology* (*ICoICT*), Bandung, 28-30 May 2014, 325-330. https://doi.org/10.1109/ICoICT.2014.6914086

[7] Ahmed, A.A.Q. and Maheswari, D. (2017) Churn Prediction on Huge Telecom Data Using Hybrid Firefly Based Classification. *Egyptian Informatics Journal*, **18**, 215-220. https://doi.org/10.1016/j.eij.2017.02.002

[8] Awang, M.K., Rahman, M.N.A. and Ismail, M.R. (2012) Data Mining for Churn Prediction: Multiple Regressions Approach. Springer, Berlin. https://doi.org/10.1007/978-3-642-35603-2_47

[9] Tavassoli, S. and Koosha, H. (2022) Hybrid Ensemble Learning Approaches to Customer Churn Prediction. *Kybernetes*, **51**, 1062-1088. https://doi.org/10.1108/K-04-2020-0214

[10] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[11] Xu, S.Q. (2019) Application Research on Telecom Customer Churn Prediction Based on XGBoost's Bagging Method. South China University of Technology, Guangzhou.

[12] Holland, J.H. (1992) Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. The MIT Press, Cambridge. https://doi.org/10.7551/mitpress/1090.001.0001

[13] Zhang, C.H.F., Wang, S., Wu, Y.D., Wang, Y. and Zhang, H.Y. (2020) Diabetes Risk Prediction Based on GA-XGBoost Model. *Computer Engineering*, **46**, 315-320.

[14] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 4768-4777.

[15] Chen, S.K., Zuo, H.X., Liao, B. and Sun, R.N. (2022) Fusing XGBoost and SHAP for Coronary Heart Disease Prediction and Its Feature Analysis Model. *Computer Application Research*, **39**, 1796-1804.

[16] Yi, M., Yao, Y.J. and Hu, M. (2022) An Interpretable Model of Public Value Consensus in New Government Media Integrating XGBoost and SHAP: An Example of the Top Ten Municipal Government Numbers of "Today's Headline". *Library Intelligence Work*, 1-12.

[17] Liao, B., Wang, C.-N., Li, M. and Sun, R.-N. (2022) A Method for Predicting and Characterizing the Price of Soccer Players by Fusing XGBoost and SHAP Models. *Computer Science*, 1-13. http://kns.cnki.net/kcms/detail/50.1075.tp.20220411.2010.002.html