Analysis and Comparison of Forecasting Algorithms for Telecom Customer Churn

Gui’e Jiao¹* and Hong Xu²

¹Shanghai University/Shanghai Jian Qiao University, No.1111 Hucheng Ring Road, Pudong New District, Shanghai, China
²Shanghai Ocean University, No. 999 Hucheng Ring Road, Pudong New District, Shanghai, China

*Corresponding author e-mail: jiaoguie@gench.edu.cn

Abstract. The integrated algorithm is a highly flexible data analysis and prediction algorithm. In many big data competitions at home and abroad, the winning teams basically adopt the idea of integrated algorithms such as random forest, GBDT, XGBoost and other algorithms. This shows that accuracy of ensemble algorithms is still very advantageous in terms of predictive classification. The main task of this article is to predict the loss of telecom customers. Under the current situation of saturation of the telecom market, how to retain the original customers is the main task of each telecom operator. This article mainly compares the four prediction models on the telecom data set. Predictive performance, the final performance evaluation index also shows that the random forest model and XGBoost model of integrated thought have better predictive models.

Keywords: Data Mining, Customer Churn, Classification Prediction

1. Introduction

For the predictive analysis of the classification status of customers, commonly used machine learning methods include clustering, association rules [1] and machine learning models [2-7]. The machine learning models used this time are mainly used. The data for predicting customer churn this time was borrowed from publicly available data sets on the Internet, and private information such as user names has been deleted.

In the data preprocessing stage, perform a simple statistical analysis of the data, including viewing the characteristic attributes of the data set and its data types, showing the loss of each attribute in the data set according to whether the customer is lost, and dividing it according to the different attribute values. In the two images on the left and right, use the .isna().sum() method to view the missing value of each attribute in the data set, filter out the attributes with missing values, and then deal with the missing values according to the specific data situation. The results show that some attributes in the data have missing values. For the processing of missing values, generally delete and fill. The attribute with missing values here is the total annual consumption. This attribute does not involve time series
and continuity, so it is not necessary to use the previous value or the next value is filled. In order to ensure the least impact on the prediction result, the average value is used here.

Fig 1. Attribute loss ratio

The data presentation form in the original data set is generally not very suitable for mining. Therefore, before the model prediction, the data needs to be normalized. In this data set, many characteristic attribute values exist in the form of strings. In order to better apply to the prediction algorithm of the model, the data type of the string type is converted. For the attribute value of two, the mapping is converted to 0 and 1. For the ternary attribute value, according to it. The actual meaning of is converted into a two-element attribute value or one-hot encoding.

Fig 2. User usage time

2. Experiment and Results Analysis

This article compares four commonly used predictive analysis models, including logistic regression model, random forest model, AdaBoost model and XGBoost model. The last three prediction algorithms are integrated algorithms. In the model test, the distribution ratio of the data set is 7:3.

2.1 Experiments

(1) Logistic regression model is a commonly used machine learning model, usually used for data mining, analysis and prediction [8]. In the prediction of the loss of telecom users, the more traditional Sigmoid hypothesis function is adopted, and the regression prediction model is based on the linear regression of the hypothesis function and independent variables. Pass the processed result data to the logistic regression model to obtain the best weight coefficient. The loss function of this model draws on the linear regression model and improves it. The loss function is

\[ J(\beta) = \frac{1}{m} \sum_{i=1}^{m} \{ -y^{(i)} \log h(x^{(i)}) - (1 - y^{(i)}) \log[1 - h(x^{(i)})] \} \]

The gradient descent method is used to solve the optimal solution of the model. The model is trained to find the best classification regression coefficients, and the user life cycle is predicted based on the characteristic attributes of user information in the data set. Different feature attributes are very important for training the weight coefficients of these data, which will affect the quality of the final prediction results.
Random forest, XGBoost and AdaBoost models are all ensemble learning algorithms. The main idea is to integrate weak learners to form the final strong learner. Integrated learning is divided into two schools, and there is no correlation between each weak learner. Random forest and XGBoost model are both typical algorithms for this ensemble learning. Another ensemble learning algorithm is the mutual influence between weak learners. The typical algorithm is AdaBoost algorithm.

(2) Random forest model: The training process of the random forest model is to build multiple decision trees and merge them to obtain a more accurate and stable model. The final output of the model is determined by many decision trees. In the classification model, the classification results of the final model are produced by voting on these decision trees. Every decision tree that makes up a random forest is very important. The decision tree will select and divide the characteristic attributes of the data. In random forest, the ranking of feature attributes is determined by information entropy. Information entropy is an index to measure the stability of attributes. In this telecommunication customer churn prediction, it is necessary to consider the mutual influence of multiple feature attributes to determine the final.

When predicting the loss of telecom customers, Random Forest combines 1000 decision trees, the minimum number of split samples is 3, the minimum leaf node is 3, and the largest feature attribute is determined internally by the algorithm. One advantage of the random forest model is that it can output the importance ranking of feature attributes. The ranking of feature attributes in the dataset is shown in Figure 3.

![Feature Importance according to Random Forest](image)

**Fig 3.** Random forest algorithm's ranking of feature attributes

The advantage of the random forest algorithm is that training can be highly parallelized, and it has advantages in training speed for large samples in big data. The algorithm can randomly select decision tree nodes to divide features, so that samples can still be effectively trained when the feature dimension is high. The importance of each function to the output can be given; the random sampling training model has small variance and strong generalization ability. It is relatively simple to implement and not very sensitive to missing values, but it can easily lead to overfitting.

(3) The principle of the XGBoost algorithm is to continuously increase the tree on the basis of the decision tree. Like random forests, weak learners are decision trees. The difference is that the algorithm adds one tree to n-1 trees. When a tree becomes n trees, the accuracy of the algorithm is continuously improved and the effect is improved [9].

The objective function of XGBoost is

\[ L = \sum_{j=1}^{J} L(y, f(x_i) + h_j(x_i)) + \gamma J + \frac{1}{2} \sum_{j=1}^{J} \omega_j^2, \]

which is mainly composed of loss function and regular term. In the function, J is the number of leaf nodes, and \( \omega_j \) is the optimal value of the j-th leaf node. I is the input training sample set, T is the maximum number of iterations, L is the loss function, used to indicate the degree to which the model fits the data, and the regularization coefficient \( \lambda, \gamma \) is used to control the complexity model. The output result of the objective function is the strong learner \( f(x) \). In the XGBoost model, the model uses Taylor’s formula to process the loss function. If the node loss is calculated under the current situation, the relevant calculation formula is

\[ \text{score} = \max \left( \frac{1}{2} \frac{G_i^2}{H_{i}} + \frac{1}{2} \frac{G_j^2}{H_{j}} - \frac{1}{2} \frac{(G_i + G_j)^2}{H_{i} + H_{j} + \lambda} - \gamma \right), \]

where \( \gamma \) is the
threshold for calculating the split loss. Only when it is greater than \( \gamma \) will it choose to split, which plays a pruning role.

The XGBoost algorithm is a very flexible algorithm and has achieved good results in many data mining activities. Like the random forest of integrated learning, the importance of the attributes of the data set will be sorted and output. The following figure shows the importance ranking of each attribute in the XGBoost model.

![Feature importance according to XGBoost](image)

**Fig 4.** Function decision value

(4) AdaBoost algorithm (Adaptive Boosting, adaptive enhancement algorithm), which belongs to the Boost type of algorithm. Its weak learners have a strong dependence. The AdaBoost algorithm is an iterative algorithm idea. The algorithm increases the weight of samples that were classified incorrectly in the previous round, and reduces the weight of samples that are correctly classified; the final ensemble method of strong classifiers is the linear weighting of weak learners Sum up. A weak learner with a low error rate has a larger weight, and a weak learner with a higher error rate has a smaller weight. In the end, the strong learner is composed of the weak learner, and the weight distribution of the weak learner will be determined according to the correct ratio of its classified samples. Here, the correct rate is related to the weight of the data sample. The higher the correct rate, the higher the weight. In order to obtain better weight coefficients, it is best to correctly classify the samples with larger weights in the previous round of classification errors to obtain higher weights during classification training.

2.2 Results and Analysis
In the algorithm debugging stage, grid tuning and cross-validation are mainly used. Grid tuning is mainly used to select the best combination of parameters and cross-validate the fit of the training model. When training a model, debugging is an extremely important step. The best parameters will make the model reach the best state. In the algorithm debugging of this paper, we compare 3-fold cross-validation and 5-fold cross-validation. In general, three-fold cross-validation has better results, and the prediction accuracy of the model has also been improved. The following table shows the final improvement effect of each model, where \( f_1 \) is the evaluation index.

|                         | Training set \( f_1 \) | Test set \( f_1 \) | Cross-validation \( f_1 \) |
|-------------------------|------------------------|-------------------|----------------------------|
| Logistic regression     | 58.45%                 | 56.09%            | 57.37%                     |
| model                   |                        |                   |                            |
| Random forest model     | 67.41%                 | 57.17%            | 57.22                      |
| XGBoost model           | 73.01%                 | 55.78%            | 56.69%                     |
| AdaBoost model          | 99.39%                 | 52.77%            | 55.11%                     |

The most commonly used evaluation indicator for machine learning algorithms is the confusion matrix, which has 4 first-level indicators. TP represents the number of samples whose prediction is true and the actual is true; FP is the number of samples whose prediction is correct but actually is wrong; FN is predicted to be false, but the actual number of samples is true; TN is the number of
samples whose prediction is false but the actual is true. The second level indicator f1 of the confusion matrix used above is the harmonic mean value.

In the analysis of the results in this experiment, the ROC curve and the AUC area are mainly used. In the coordinate system, the abscissa and ordinate are related to the confusion matrix, the abscissa is $TP = TP + FN$, the ordinate is $FPR = FP + TN$, and AUC represents the area under the ROC curve. Is mainly used to measure the generalization ability of the model. AUC is a specific value, which is more suitable for comparative analysis than the curve. The criterion of AUC value for model performance is: when AUC=1, the performance of the classifier is very good, the perfect prediction effect can be achieved when using this model. This model is ideal and generally does not exist; if AUC<0.5, then the prediction effect of the model is very poor. If it is such a model, then it needs to be performed Reverse prediction; the general performance model AUC value is between 0.5 and 1, the larger the value, the better the performance [10].

The results of the four algorithms are compared as follows:

![Comparison of the results of the four models](image)

3. Conclusion
A comprehensive comparison of four machine learning prediction algorithms, logistic regression algorithms and three ensemble learning algorithms found that the random forest algorithm, XGBoost algorithm and AdaBoost algorithm are more effective, but for many modeling scholars, it is similar to the black box principle. The internal operation of random forest is opaque. The model used this time is a relatively basic model, and then simple parameter adjustments have been made. Future research can improve the loss function and objective function. In addition, the current computer development has moved towards artificial intelligence, and the development of data mining is gradually moving towards the field of deep learning. Neural network is a very successful technology for computer deep learning, so the next step is to consider using deep learning, neural network processing data set [11-12].

References
[1] Wang, Xinyan and Jiao, Guic. ‘Research on Association Rules of Course Grades Based on Parallel FP-Growth Algorithm’. 1 Jan. 2020 : 759 – 769.
[2] Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining August 2016 Pages 785–794.
[3] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal and A. Rehman, "Telecommunication subscribers' churn prediction model using machine learning," Eighth International Conference on Digital Information Management (ICDIM 2013), Islamabad, 2013, pp. 131-136, doi: 10.1109/ICDIM.2013.6693977.
[4] TELE-INFO'06: Proceedings of the 5th WSEAS international conference on Telecommunications and informaticsMay 2006 Pages 281–286
[5] So Young Sohn, Jae Kang Lee. Competing risk model for mobile phone service[J]. Technological Forecasting &amp; Social Change, 2008, 75(9).
[6] Proceedings of the 2017 12th International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2017, 2017, 2018-January pp. 1 - 6.
[7] Scott A. Neslin, Sunil Gupta, Wagner Kamakura, Junxiang Lu, Charlotte H. Mason. Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models[J]. Journal of Marketing Research, 2006, 43(2).
[8] Gopal R.K., Meher S.K. (2008) Customer Churn Time Prediction in Mobile Telecommunication Industry Using Ordinal Regression. In: Washio T., Suzuki E., Ting K.M., Inokuchi A. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2008. Lecture Notes in Computer Science, vol 5012. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-68125-0_88
[9] Abdelrahim Kasem Ahmad, Assef Jafar, Kadan Aljoumaa. Customer churn prediction in telecom using machine learning in big data platform[J]. Journal of Big Data, 2019, 6(1).
[10] Ascarza, Eva, Retention Futility: Targeting High-Risk Customers Might Be Ineffective (January 6, 2018). Columbia Business School Research Paper No. 16-28, Available at SSRN: https://ssrn.com/abstract=2759170 or http://dx.doi.org/10.2139/ssrn.2759170.
[11] Guolin Ke, Zhenhui Xu, Jia Zhang, Jiang Bian, and Tie-Yan Liu. 2019. DeepGBM: A Deep Learning Framework Distilled by GBDT for Online Prediction Tasks. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining. KDD ’19. Association for Computing Machinery, New York, NY, USA, 384–394. DOI:https://doi.org/10.1145/3292500.3330858 GBDT
[12] J. Hu et al., "pRNN: A Recurrent Neural Network based Approach for Customer Churn Prediction in Telecommunication Sector," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4081-4085, doi: 10.1109/BigData.2018.8622094.