Analysis and Prediction of Telecom Customer Churn based on Machine Learning

Ye Xia[1, *, +], Bohan Cui[2, +] and Yunhuai Duan[3, +]

[1] Beijing Royal school, Beijing, China

[2] Anhui University, Hefei, China

[3] Nanjing University, Nanjing, China

* Corresponding Author Email: xiaye@brs.edu.cn

[+]These authors contributed equally.

Abstract. As the telecommunications market becomes increasingly saturated, major operators are facing an increasingly severe problem of soaring customer churn rates. How to identify high-risk churn customers is the most concerned issue for operators. Thanks to the rapid development of pattern recognition technology, existing machine learning algorithms provide key technical support for telecom customer churn prediction. However, how to choose an appropriate forecasting method combined with the characteristics of the application data is still an open question. To this end, based on the analysis and comparison of the feature correlation between telecom customer data and churn, this paper compares the differences in the prediction results of different machine algorithms, so as to choose the method that best fits the characteristics of the application data to build the final customer churn prediction model. Specifically, the Spearman correlation coefficient is used to calculate the correlation between variables in the dataset, the random forest algorithm is used to score the importance of all variables, and the prediction generated by the gradient boosting tree algorithm is introduced. Finally, the gradient boosting tree algorithm is evaluated by five performance indicators: precision rate, recall rate, precision rate, F1 score and AUC (Area under the ROC curve).

Keywords: Machine learning; customer churn prediction; Spearman single factor analysis; random forest feature importance score.

1. Introduction

Based on the communication services (network infrastructure, data transmission and basic voice communication services, etc.) Provided by telecom operators, communication between people through the intelligent mobile terminals (mobile phones, computers, etc.) has become an integral part of daily social life [1]. However, with the diversification of user demands and the continuous improvement of communication service quality requirements, the pressure of competition among telecom operators has gradually increased, which is directly reflected in the competition for customer resources [1]. In order to win more customers, telecom operators not only develop new customers through price advantage or launch new products but also retain existing customers as much as possible by improving service quality. Compared with developing new customers, retaining and enhancing the value of existing customers has become the preferred solution for operators due to their relatively low cost. The basis and key to retaining existing customers is to achieve customer churn prediction, early warning, cause analysis and retention. To this end, we urgently need to develop an efficient prediction model for telecom customer churn to provide decision-making basis for improving customer experience.

Lost customers usually refer to customers who stop using the company's services or products within a certain period of time. At present, the research objects of telecommunication customer churn mainly focus on the prediction of traditional telecommunication customer churn and the network customer churn. The research method is mainly to introduce a prediction of customer churn using feature vector selection and classifier optimization, where the representative technologies includes artificial neural network algorithm, association rule algorithm, decision tree algorithm, SVM(support vector machine) algorithm and so on. For example, HS Kim (2003) used association rules to analyze

the factors of Korea Telecom users' choice of operators and drew a conclusion that it was related to signal quality and discount of call charges. M.C.Mozer (2000) used logistic regression, decision tree, neural network and other methods to study the personal information, bill, credit, application program, complaint history and other data of 47,000 Telecom users in the United States, and predicted the customers who are imminent leave [2]. Eria Kamya and Marikannan Booma Poolan used logistic regression and random forest models to exclude the feature variables that were least relevant to the model prediction, so as to accomplish the user churn prediction model based on the remaining feature variables. Machine learning algorithm enables the model to get rid of irrelevant variables and provide reliable correlation features relate to user churn. In addition, the study also pointed out the possible prediction errors in the model for predicting user churn, which can be reduced continuously after consideration and finally improve the accuracy of the model for testing user churn. Deri et al. applied graph method to analyze the network of mobile communication users to predict the number of customers which may be lost. Backiel et al. constructed the user's relational network based on the data about Call Detail Records, then extracted the characteristics of the network and compared the methods of predicting user loss by using user attributes. The methods proposed by them scored higher in AUC [3]. Amin et al. used Rough Set based on genetic algorithm to predict churn. NTT DoCoMo Company, a Japanese company, had subdivided customers according to their consumption level, reputation and functional service requirements. Lightbridge collected data from a mobile service provider in New England and then built a model about customer churn based on the CART algorithm.

Although the above work has greatly promoted the research of telecommunication customer churn prediction, due to the complexity and difference of experimental data distribution, the prediction results of different methods often fluctuate greatly, and none of them can well meet the actual application requirements. In this paper, based on the obtained dataset "Telecom Operator Customer Dataset" which contains 21 fields and 7043 records, the reasons for customer churn of telecom operators are firstly analyzed from three dimensions. Then, through a series of sorting, statistics, transformation and processing, we quantitatively analyze the relationship between the characteristics of the first 20 columns and whether customers are lost in the 21st column, and build customer churn prediction models based on different machine learning algorithms. By comparing the models and grid searching for a better model, we end up building a predictive model with higher accuracy.

2. Data pre-processing and Descriptive statistical analysis

2.1. Data description

The original research data we used comes from the “Customer dataset of telecom operators” including 21 fields and 7043 records in total. Each record consists of the information of features of specific customer, whose basic information and specific meaning of variables is shown in Table 1.

Table 1. Basic user information of the telecom operators’ customer dataset.

Variable name	Variable type	Variable value (range)
customerID	irrelevant variable	-
gender	discrete	Male,Female
SeniorCitizen	discrete	0, 1
Partner	discrete	Yes, No
Dependents	discrete	Yes, No
tenure	continuous	0-72
PhoneService	discrete	Yes, No
MultipleLines	discrete	No phone service, No, Yes
Internet Service	discrete	DSL, No, Fiber optic
Online Security	discrete	Yes, No internet service, No

Online Backup	discrete	Yes, No internet service, No
Device Protection	discrete	Yes, No internet service, No
Tech Support	discrete	Yes, No internet service, No
Streaming TV	discrete	Yes, No internet service, No
Streaming Movies	discrete	Yes, No internet service, No
Contract	discrete	One year, Two year, Month-to-month
Billingsgate	discrete	Yes, No
Payment Method	discrete	Credit card (automatic), Mailed check,Electronic check, Bank transfer (automatic)
Monthly Charges	continuous	-50.8-118.75
Total Charges	continuous	1-8684.8
Churn	label	Yes, No

2.2. Data pre-processing

2.2.1 Outlier processing

Outliers, also commonly referred to as anomalies, are those unreasonable points obviously in the data set that are impossible to achieve in the real situations. Outlier processing is generally relative to those continuous variables [3-4]. Here, we first delete the irrelevant variables that does not contain useful information in the customer-ID column and then check the descriptive statistics of the continuous variables, which includes the tenure, Monthly Charges and Total Charges in our data set.

2.2.2 Missing value processing

Missing values are mainly processed in two ways, namely deleting features with missing values or filling in missing values. Note that some variables in the customer churn data miss too many their attributes so that they only contains less useful information, the missing ratio can be used to delete those variables with a missing ratio of more than 50% in the data, so as to further improve data quality. For those variables with fewer missing values, the most probable value can be used to fill the missing values, which is more convincing than deleting all samples with missing values, and also helps retain more information [5]. Generally, the mean, mode and median methods can be used to directly fill in the missing values. The column chart of missing features in the data set can be made to better visualize in Figure 2. As the Figure 1 shown, there are only a small number of missing values in the three Total Charges, Payment Method and Monthly Charges, whose missing ratio is less than 1%. The continuous features of Total Charges and Monthly Charges can adopt their respective means to fill in, while the discrete feature of the Payment Method should use the mode to fill in the missing values.

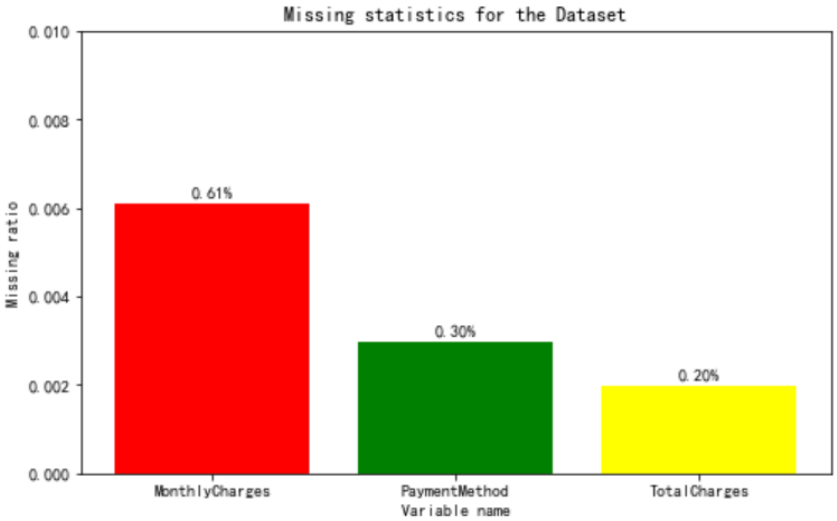


Fig 1. Missing features in the data-set.

2.2.3 Discrete variable coding and sample distribution

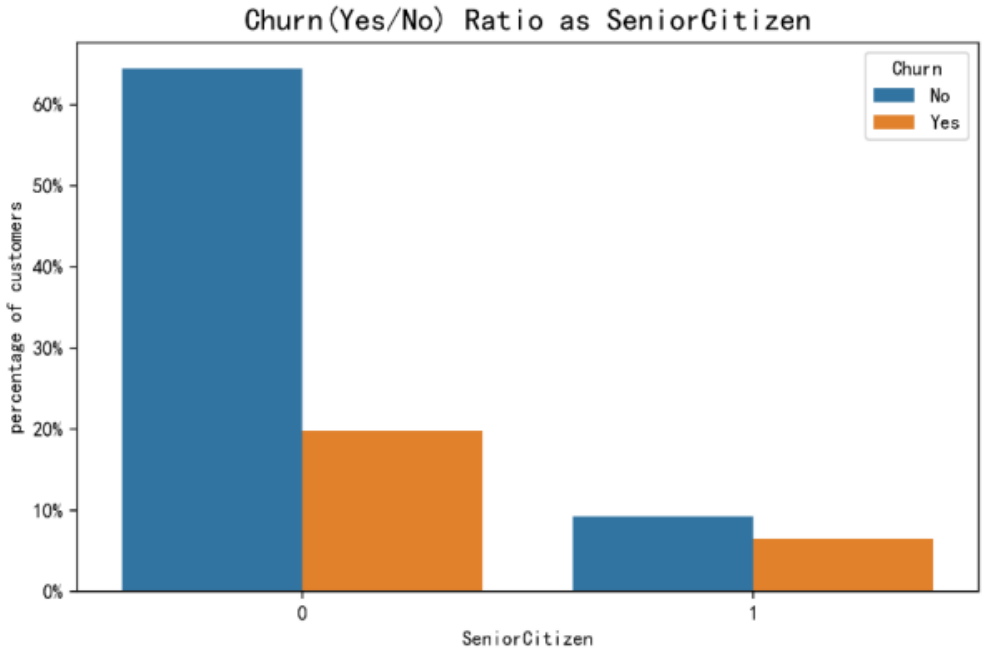
Before constructing a customer churn prediction model for telecom operators, it is essential to first analyze the distribution of sample labels. In detail, the particularly unbalanced distribution can seriously affect the model training, leading to the learning algorithm preferring those types with more samples on the unbalanced data [6]. Generally speaking, when the distribution ratio of positive and negative samples exceeds 4:1, it becomes very necessary to take some undersampling or oversampling processing measures on those unbalanced sample distributions. .According to Dateset, the sample distribution reaches nearly 3:1 so that it is relatively uneven, but the ratio does not exceed 4:1. In addition, to facilitate the following analysis of the input model, we also adopt the one-hot encoding method to process the discrete data.

3. Descriptive statistical analysis

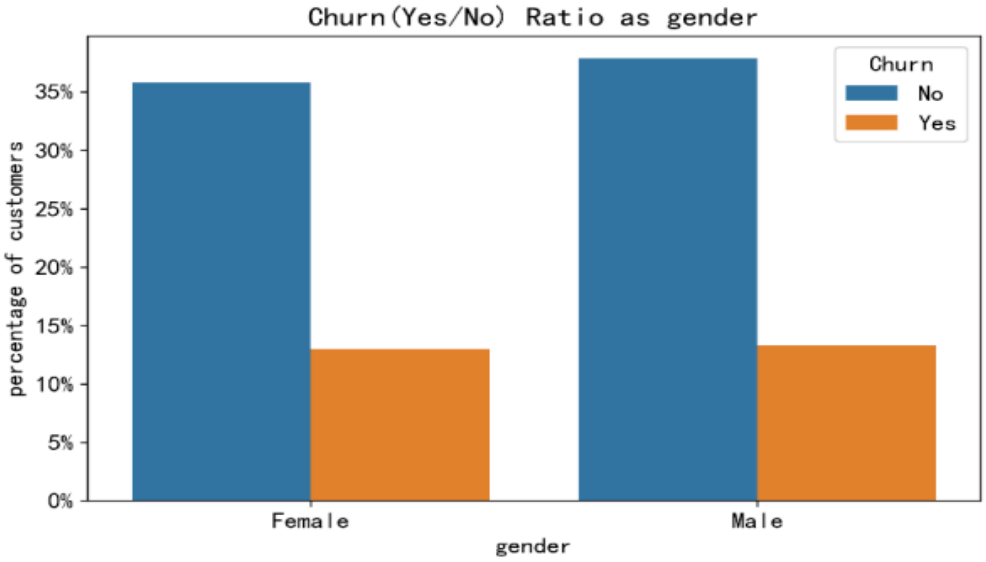
After conducting the data pre-processing, a comprehensive analysis on the customer churn of telecom operators should be made. Only by finding the reasons for the customer churn, we can better formulate some targeted strategies to retain those users who will leave in the future [7]. According to the user information in the table, the user characteristics can be divided into user attribute, service attribute and contract attribute, thus making the corresponding visual analysis from these three dimensions.

3.1. User attribute analysis

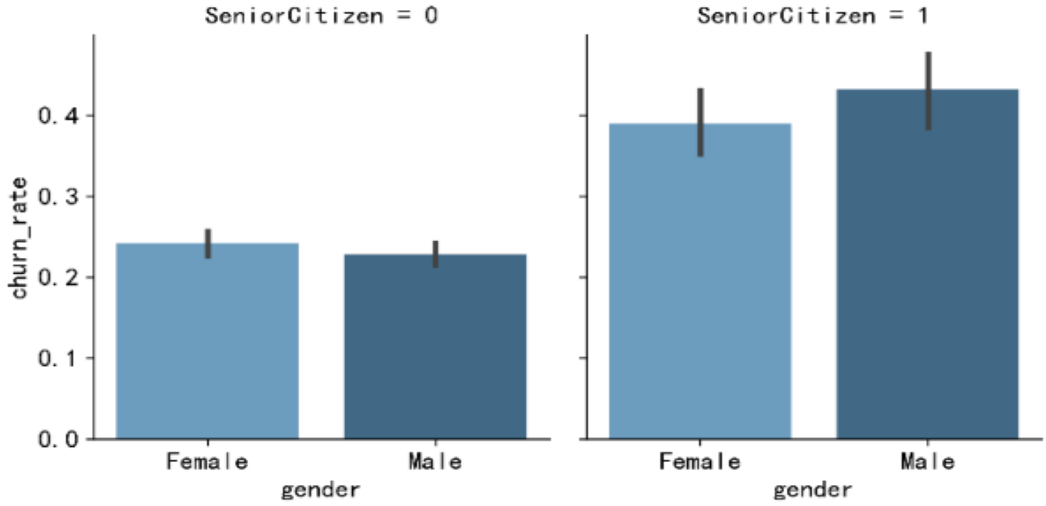
We mainly analyze whether the elderly has an impact on user churn in the user attribute. As shown in Figure 4, we check whether the elderly has an impact on user churn in the user attribute, which mainly analyze the gender distribution of lost customers, the gender distribution of lost customers among senior citizen users and distribution of user churn in Partners and Dependents. According to the Figure 2, the number of senior people utilizing telecom fiber is significantly smaller than that of the youth [8]. However, compared to those younger customers, the churn rate of older customers is significantly higher and at nearly two-thirds. However, the churn rate of young customers is only about one-third. Moreover, it can be proved that user churn is almost irrelevant to gender.



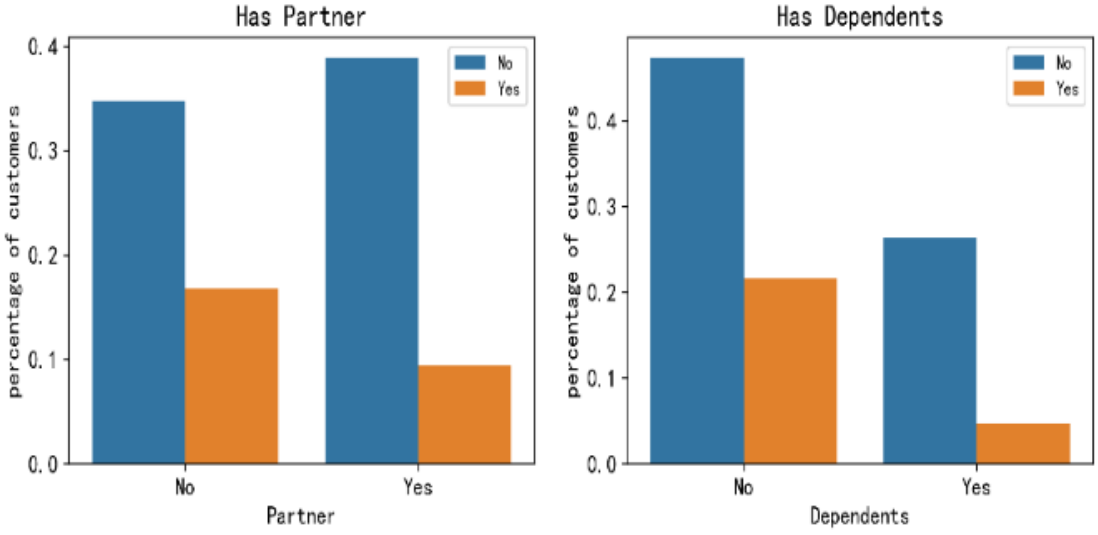
(a) In Senior Citizen.



(b) In gender.



(c) In gender under similar Senior Citizen.



(d) In Partners and Dependents.

Fig 2. Distribution of user churn rate under different settings.

It is also essential to view the probability distribution density over some continuous variables corresponding to different categories when implementing statistical analysis, so that better useful

properties can be found. Figure 3 shows the relationship between customer churn and online time based on the corresponding kernel density estimation [9]. In detail, its vertical axis can be roughly proportional to the number of data occurrence, and the area of the region under the curve is one. Figure 3-5 shows the relationship between customer churn and online time based on the corresponding kernel density estimation. Note that the churn proportion of users with partners is considerably lower than that of users without partners, and the number of customers with family members is also less [10]. However, compared with those users without family members, the ratio of users with family members is substantially lower and only about one-sixth. Moreover, the longer online time can bring lower churn rate, which is in line with general experience. If the online time reaches 23 months and the churn rate is less than the online rate, it is proved that the user's psychological stability period is generally about 23 months.

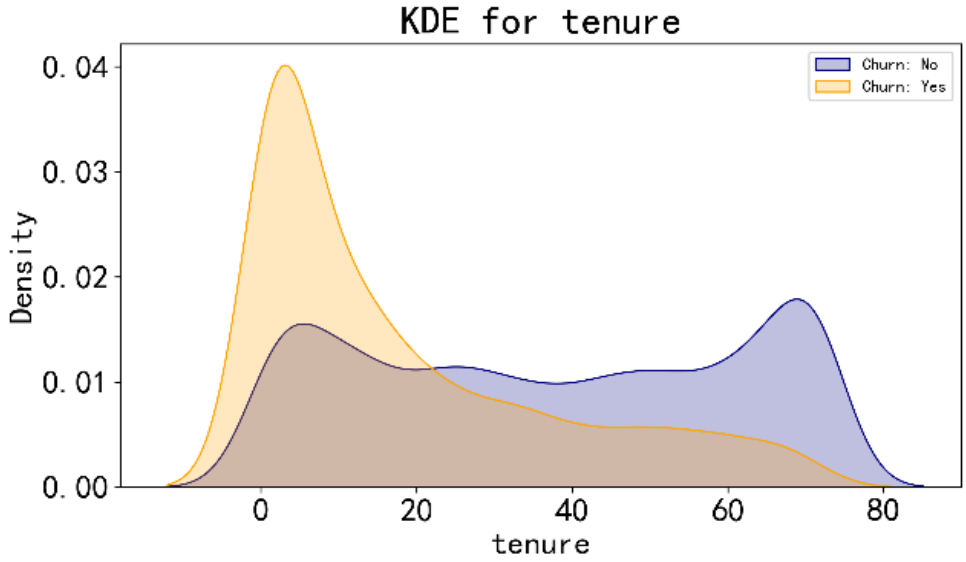


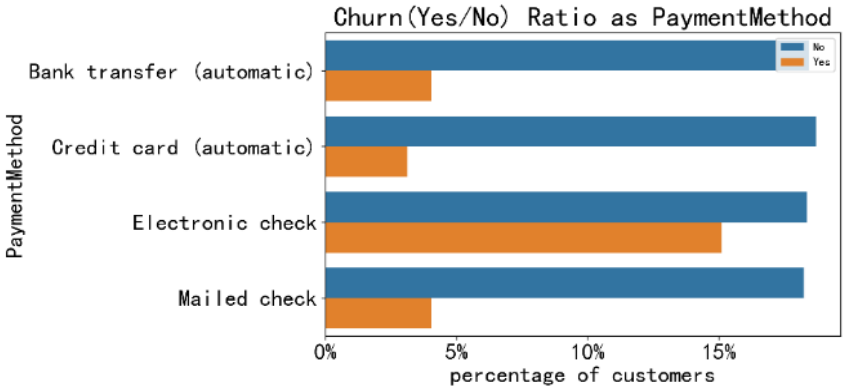
Fig 3. Kernel density estimation for online time.

3.2. Service attribute analysis

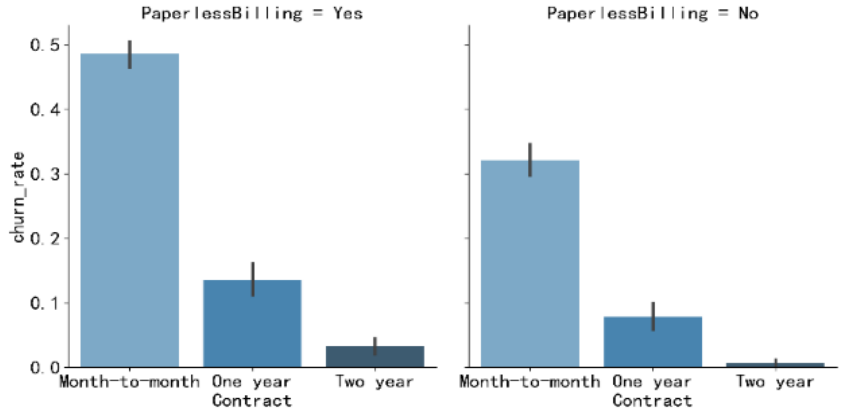
The analysis of service attributes is equally important. We also focus on the distribution of user churn ratio in Multiple Lines and Internet Service below, and also makes the statistics on the number of users using some additional network services and the number of lost customers in various additional services. It can be concluded that the overall impact of telephone service on user churn is relatively small, while the churn ratio of optical fiber network service users is significantly higher than that of digital network and non network service users, especially the churn ratio is nearly three-seventh of that of optical fiber users. Among those users choosing network additional services, the churn ratio of users also selecting security service, backup service, protection service and technical support service is lower. And among those additional services, the churn ratio of users selecting telephone service, multi-line network service, streaming TV and film services is relatively higher.

3.3. Contract attribute analysis

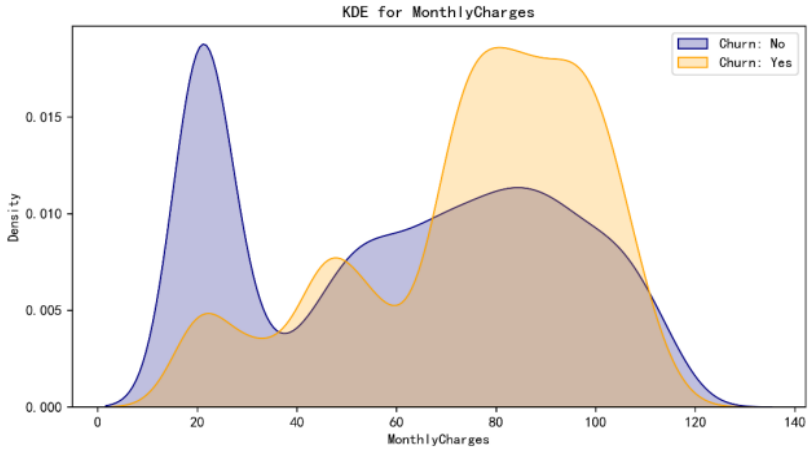
Similarly, the distribution of user churn ratio in the contract attribute can be found in Figure 4. It can be concluded that the user churn rate using electronic payment is the highest, so its corresponding use experience is relatively general. And the impact of the contract signing method on the customer churn rate is shown as monthly signing > signing on a yearly basis > signing on a two-year basis, which further proves that long-term contracts can better retain customers. The churn rate of the user group whose monthly consumption bill is about 75-105 yuan is the highest, and the customer group whose annual consumption amount is about 300 yuan has the highest churn rate. It can be further shown that the increased consumption amount can bring the gradually decreased churn rate of users, which is in line with the actual situation.



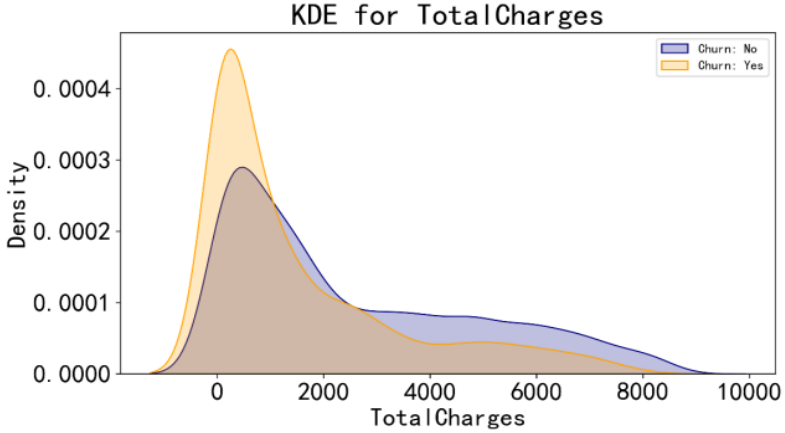
(a) In Payment Method.



(b) In Contract under Paperless Billinge.



(c) Kernel density estimation for Monthly Charges.



(d) Kernel density estimation for Total Charges.

Fig 4. The distribution of user churn ratio in the contract attribute.

4. Methodology

4.1. Dataset construction

After the data pre-processing, there are also 7042 samples left. The data is randomly divided into two parts, namely the training set and test set, according to the proportion of 7:3. And random seed is manually set to 0 to ensure that the result of each division is always the same. After the corresponding division, the training set data has a total of 4929 samples, and the test set data has a total of 2113 samples. Among them, there are 1296 churn customers in the training set and 3633 online customers, while in the test set, there are 555 churn customers and 1558 online customers.

4.2. Spearman correlation analysis

The correlation analysis can help better check the collinearity between features. The correlation threshold is usually set at 0.8, indicating that if the absolute value of the correlation between the other two variables is greater than 0.8, is is necessary to delete one of them. Here, we calculate the Spearman correlation coefficient through the formula (1):

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1}$$

Where the d_i is the difference of ranks, which come from the sorted position information of two variables X and$Y$. n denotes the number of data in the variable [11]. The customer tenure is strongly correlated with the Total Charges, and a positive value also shows a positive correlation. The correlation between these two variables is greater than 0.8, thus deleting the Total Charges variable. In addition, the heat map shows the correlation between any two columns of variables. For example, the customer tenure and the contract signing method are considered too strongly connected since there is a relatively large positive correlation between them. Besides, Total Charges and Monthly Charges are also regarded to be strongly correlated for same reason [12].

4.3. Feature importance scoring based on random forest algorithm

We make the importance scoring on all variables based on the random forest algorithm. Through setting a threshold and only retaining the factors whose importance score is greater than the threshold, it can help explore core variables that affect the specific customer churn. The Gini coefficient is used as a measure of node contribution in the process of calculating feature importance. Suppose there are m factors$(X_1, X_2, \ldots, X_m)$, VIM represents the importance score of the factor and GI represents the Gini coefficient, which can be calculated as:

$$GI_m = 1 - \sum_{k=1}^{|K|} p^2{}_{mk} \tag{2}$$

$$VIM_{jm}^{Gini} = GI_m - GI_l - GI_r \tag{3}$$

Here, GI_m is the Gini coefficient of node m; K is the number of categories in the label while P_{mk} denotes the proportion of samples with K in node m. GI_l And GI_r denote the Gini index of the two new nodes of the branch respectively. Suppose that the nodes of factor X_j in the tree are in the set M, let n denote the number of trees in the random forest algorithm then the importance of X_j with an ordinal number is as follows:

$$VIM_{ij}^{Gini} = \sum_{m \in M} VIM_{jm}^{Gini} \tag{4}$$

The Gini coefficients are normalized to obtain the importance scores of the influencing factors. Based on the random forest algorithm, the importance of each variable on the label is analyzed, with all corresponding hyper-parameters set to default values. The set random seed 0 can ensure the reproduced experiment. The two column variables of customer tenure and Monthly Charges have the greatest correlation with the customers churn label of the telecom operators, which follows the results of the descriptive statistical analysis introduced previously. In detail, customers with the longer tenure time and higher monthly charges are less likely to leave. Moreover, the variable Phone Service business has the smallest correlation with the customer churns, and the correlation value is also obviously smaller than other variables. Based on the related experience, the set correlation threshold 0.01 can help delete the column variables of Phone Service whose correlation value is less than 0.01.

4.4. Model construction

4.4.1 Prediction model based on random forest algorithm

The random forest algorithm builds its own weak learner by improving the CART decision tree. Let the input data set$D = \{(x_1, y_1), (x_2, y_2), \ldots (x_m, y_m)\}$, the number of iterations determined by the weak classifier be T and the final output of the strong classifier bef(x). The basic steps of RF are as follows. For$t = 1,2\ldots\ldots, T$, random sampling is performed on the training data set. If a total of m samples are collected, the sampling set containing m samples obtained by the t_{th} is D_t. Select a part of the sample features in D_t to train the nodes of the decision tree, and then select an optimal feature among the selected features to divide the corresponding left and right sub-trees. The set T weak learners will vote on the results to produce the predicted result.

When constructing a customer churn prediction model based on the random forest algorithm, the hyper parameters of the model can be set to: max_depth=6, random seed is 0 and other hyper parameters are default values. After building the model, the 4929 samples of the training set can be used to train the model, and then import the 2113 samples of the test set into the model for the corresponding testing.

4.4.2 Prediction model based on XGBoost algorithm

XGBoost is one of the most efficient implementations of gradient boosting decision tree (GBDT), whose main process is as follows.

If the maximum number of iterations of the classifier is set to T, the loss function is L, the regularization coefficient are λ, γ,and $f(x)$ denotes the output of strong learner. Then for the input data set $I = \{(x_1, y_1), (x_2, y_2), \ldots (x_m, y_m)\}$ and the feature serial number $k = 1,2\ldots K$:

(1)$G_L = 0, H_L = 0$;

(2) Arrange the samples according to the features k, and take the i sample in turn to calculate the sum of the first and second derivatives of the left and right sub trees after placing the current sample in the left sub-tree.

$$G_L = G_L + g_{tt}, G_R = G - G_L \tag{5}$$

$$H_L = H_L + h_{tt}, H_R = H - H_L \tag{6}$$

(3) Try to update the largest score:

$$S = max(\ \text{score}\ , \frac{1}{2}\frac{G_L^2}{H_L + \lambda} + \frac{1}{2}\frac{G_R^2}{H_R + \lambda} - \frac{1}{2}\frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \gamma) \tag{7}$$

(4) Find the corresponding splitting feature and eigenvalue when S becomes the maximum value, so as to continue to divide the sub-tree.

(5) When the maximum value of S is 0, it means that the decision tree has been established to solve w_{tj} of all the leaf regions, so as to obtain the weak learner$h_t(x)$, and then update the strong learner

$f_t(x)$ .Then, the next round of weak learner iteration can be started. If the maximum value of S is not 0, the second step process should be continued.

When building a customer churn prediction model according to the XGBoost algorithm, learning_rate, max_depth, number of trees and n_estimators are set to 0.1, 8, 200, 0 respectively, and other hyperparameters is set to the default value. The training set and test set data are invariant when used for training and testing the model, so the following will not be repeated when building other models.

4.4.3 Prediction model based on logistic regression algorithm

Logistic regression generally refers to such a process: the first is to set the cost function for a regression classification problem, and then use the optimization method to iteratively calculate the optimal model parameters, and to better determine the problem type by the corresponding testing. The following is the calculation process of the logistic regression model with the L_1 regularization term. The first step is to establish a classification prediction function. The classification prediction function is:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \tag{8}$$

In detail, θ^T is the transpose of the vector $\theta = (\theta_0, \theta_1, \ldots, \theta_n)$, $x = (x_0, x_1, \ldots, x_n)$ is an n-dimension vector where $x_0 = 1$, notice that the dimension of the data is $n - 1$. Then, construct a cost function (loss function) with a L_1 regularization term to convert the classification function problem into a vector θ as:

$$J(\theta) = -\frac{1}{m} l(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [(y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))] + \frac{\lambda}{m} \sum_{j=1}^{m} \theta^2 \tag{9}$$

$x^{(i)}$ Is the value of each sample data point on a certain feature, which is actually the value of the ith term of feature vector x. $y^{(i)}$ Denotes the category number, m is the number of sample objects, while $\frac{\lambda}{m} \sum_{j=1}^{m} \theta^2$ is the L_2 regularization term. Finally, to find the solution of the maximum log-likelihood function, the gradient descent method is used, which also obtains the predicted category according to the classification prediction function. When constructing a prediction model based on the logistic regression algorithm, only one random seed is set to 0.

4.4.4 Prediction model based on Naive Bayes Algorithm

The Naive Bayes algorithm is a classification method based on Bayes' theorem and the conditional assumption of independent factors. Suppose that the training data is $D = \{d_1, d_2, \cdots d_n\}$, the sample features in the training data are $X = \{x_1, x_2, \cdots x_d\}$, the class variable in a column of features is $Y = \{y_1, y_2, \cdots y_m\}$, meaning to divide D into y_m types. Among them, $x_1, x_2, \cdots x_d$ are independent of each other and random, the priori probability Y is $P_{prior} = P(Y)$, the posterior probability Y is $P_{post} = P(Y \mid X)$. Based on priori probability $P_{prior} = P(Y)$, evidence $P(X)$ and class conditional probability $P(X \mid Y)$, it can help calculate the posterior probability:

$$P(Y \mid X) = \frac{P(Y)P(X \mid Y)}{P(X)} \tag{10}$$

Since each feature is independent, the above formula can be interpreted as follow:

$$P(X \mid Y=y)=\prod_{i=1}^{d} P(x_i \mid Y=y) \tag{11}$$

The posterior probability can be obtained through the above two equations:

$$P_{post}=P(Y \mid X)=\frac{P(Y)\prod_{i=1}^{d} P(x_i \mid y)}{P(X)} \tag{12}$$

The value $P(X)$ is fixed, so the posterior probability can be known by comparing the size of the molecule, thus determining whether a certain sample data belongs to the category y_i:

$$P(y_i \mid x_1, x_2, \cdots, x_d)=\frac{P(y_i)\prod_{i=1}^{d} P(x_i \mid y_i)}{\prod_{i=1}^{d} P(x_j)} \tag{13}$$

4.4.5 Prediction Model Based on gradient boosting decision tree algorithm

GBDT generally uses the CART regression tree for both classification and regression process. For the binary classification, the complete algorithm flow of GBDT is as follows:

(1) Initialize a weak learner $F_0(x)$:

$$F_0(x)=\log\frac{P(Y=1 \mid x)}{1-P(Y=1 \mid x)} \tag{14}$$

$P(Y=1 \mid x)$ Is the ratio of the training samples $y=1$, and the prior information can be used to initialize the learner.

(2) For building a classification and regression tree of M $m=1,2,...,M$:

a. For $i=1,2,...,N$, calculate the corresponding response value of the m tree:

$$r_{m,i}=-\left[\frac{\partial L(y_i,F(x_i))}{\partial F(x)}\right]_{F(x)=F_{m-1(x)}}=y_i-\frac{1}{1+e^{-F(x_i)}} \tag{15}$$

b. For $i=1,2,...,N$, fitting data ($x_i, r_{m,i}$) by CART regression tree can obtain the m regression tree. In detail, the leaf node of these trees can be $R_{m,j}$, $j=1,2,...,J_m$, of which J_m is the number of leaf nodes of the m regression tree.

c. The optimal fitting value can be obtained by fitting the leaf node area $j=1,2,...,J_m$ of the J_m :

$$c_{m,j}=\frac{\sum_{x_{i\in R_{m,j}}} r_{m,i}}{\sum_{x_{i\in R_{m,j}}} (y_i-r_{m,i})(1-y_i+r_{m,i})} \tag{16}$$

d. Update the strong learner $F_m(x)$:

$$F_m(x)=F_{m\text{-}1}(x)+\sum_{j=1}^{J_m} c_{m,j} I(x\in R_{m,j}) \tag{17}$$

(3) Then the formula of the final strong learner $F_M(x)$ is obtained:

$$F_M(x)=F_0(x)+\sum_{m=1}^{M}\sum_{j=1}^{J_m}c_{m,j}I(x\in R_{m,j}) \tag{18}$$

The binary GBDT classification algorithm is the same as the logistic regression, meaning to fit the log probability through a series of gradient boosting trees. The specific classification model can be expressed as:

$$P(Y=1|x)=\frac{1}{1+e^{-F_M(x)}} \tag{19}$$

Where $P(Y=1|x)$ denotes the probability of the output being positive when the input is x in the customer churn prediction model, the corresponding hyperparameters are set to: the max_depth=8, the number of trees n_estimators=300, the random seed is 0, and other hyperparameters are default values.

5. Experiment

5.1. Calculate the evaluation index

The prediction effect of the corresponding model constructed based on five machine learning algorithms on the test set is mainly evaluated through the accuracy , recall, precision, F1 score and AUC.

Among them, the accuracy represents the percentage of correct prediction cases in the total. Precision represents the correct rate of prediction among the predicted lost cases. Recall represents the correct ratio of prediction among the actual lost cases.F1-score is calculated by precision and recall, which is equivalent to a comprehensive evaluation of the two results. Their calculation formulas are as follows:

$$Accuracy=\frac{TP+TN}{TP+FP+FN+TN} \tag{20}$$

$$\mathrm{Re}call=\frac{TP}{TP+FN} \tag{21}$$

$$Precision=\frac{TP}{TP+FP} \tag{22}$$

$$F1\_score=\frac{2}{\frac{1}{Prcision}+\frac{1}{Recall}}=2\cdot\frac{Prcision\cdot Recall}{Prcision+Recall} \tag{23}$$

Where TP, FP, FN and TN denotes true positive, false positive, false negative, true negative respectively.

5.2. Performance comparison

We first compare the accuracy of different models, whose results is shown in Table 2. According to the performance indicators of the five models introduced above on the test set, the gradient boosting tree model is significantly better than other 4 models in terms of all these indicators. The value of each indicator exceeds 0.78, especially the accuracy rate reaches 91% and shows better results.

Table 2. Performance indicators of each model on the test set.

Model	AUC	Accuracy	Precision	Recall	F1 score
random forest	0.8332	0.8173	0.6907	0.5514	0.6132
logistic regression	0.8513	0.8002	0.6406	0.5459	0.5895
XGBOOST	0.9281	0.8978	0.8253	0.7748	0.7993
Naive Bayes	0.8669	0.7582	0.5273	0.7568	0.6218
Gradient boosted tree	0.9504	0.9110	0.8605	0.7892	0.8233

5.3. Grid search parameters adjustment

The following is to timely select the optimal model for grid search tuning. Grid search is actually a traversal combination search. Among all the tuned parameters, different hyper-parameters can realize the traversed combination to predict each possibility, while the hyper-parameters combination with the best test results is used as the parameter tuning result [13]. During the parameters tuning process, it mainly focuses on the two hyper parameters that is n_estimators and the max_depth due to their greatest impact on the model effect. The range of n_estimators and max_depth are set from 10 to 400 with a size of 10 and from 1 to 11 with a step of 1 respectively. Taking the comprehensive evaluation index of AUC value as the parameter tuning goal, the AUC value of the model on the test set when grid searching is mainly shown in Figure 5.

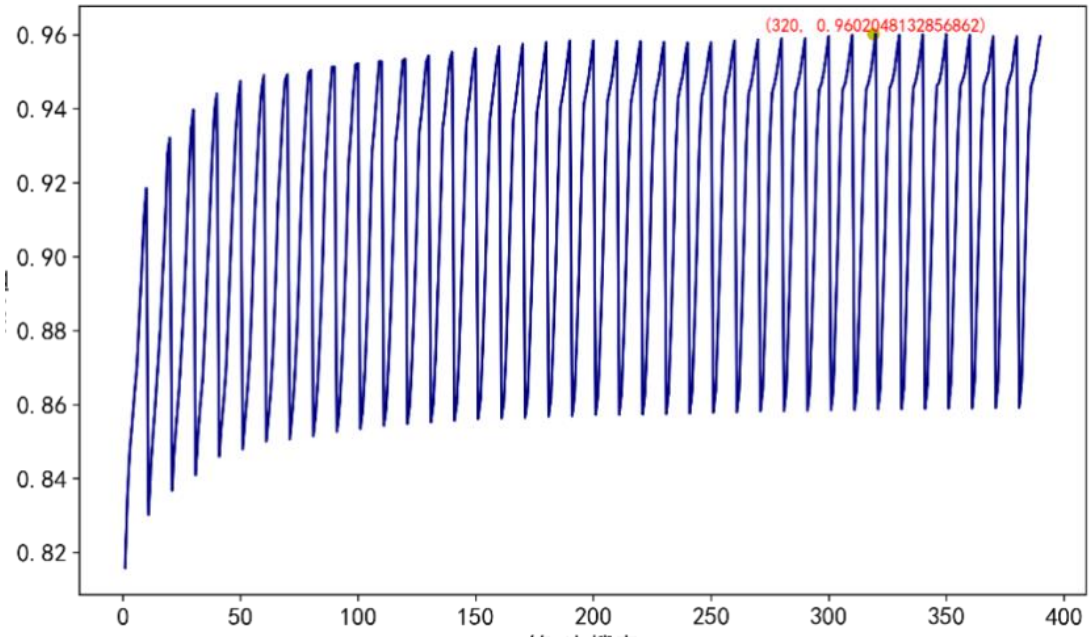


Fig 5. Grid search parameter tuning results.

As the line graph illustrates, the largest AUC value generated by grid search is around 0.9602. In this context, the corresponding n_estimators is 310 and max_depth is 10. The indicator comparison of the prediction results before and after model parameter tuning is shown in Table 3. After the parameters tuning, all the indicators of the model on the test set get fully improved to a certain extent. Therefore, it can be considered that when n_estimators is 320, max_depth is 10, and other parameters are default values, the gradient boosting tree model is the best for predicting the customer churn of telecom operators in this work.

Table 3. Comparison of results before and after model parameter tuning.

	AUC	Accuracy	Precision	Recall	F1 score
Before tuning	0.9504	0.9110	0.8605	0.7892	0.8233
After tuning	0.9596	0.9190	0.8595	0.8270	0.8429

6. Conclusion

According to the above analysis, elderly users, unmarried users and users without relatives are more likely to leave and have the higher churn rate characteristics. And those users with about 4 months online time, telephone services, fiber users/fiber users with additional streaming TV and film services, and also customers with no Internet value-added service, shorter signed contract period, electronic payment, electronic bill and a monthly rental fee of about 70-110 yuan have higher churn rate. And other attributes have less impact on user churn, so the above characteristics remain independent.

It is important to construct a list of users with a higher churn rate based on a predictive model. The related user research can help form a minimum viable product feature and timely invite seed users to have a try. User aspect: it is essential to provide customized services for elderly users, and users without relatives and partners based on their characteristics, such as relative and warm services. On the one hand, it can strengthen the relationship with other users, and on the other hand, it can help provide personalized services for specific users. Service aspect: to overcome the peak period of user churn which generally result from those newly registered users, it is feasible to launch half-year discounts, such as gift coupons. For those optical fiber users and users of additional streaming TV and movie services, we should focus on improving network and value-added service experience. On the one hand, the technical department should be required to improve network indicators, and on the other hand, we can promise users to timely provide free network upgrading and monthly subscription services such as TV and movies, so as to better increase their stickiness. As for those value-added services, such as technical support, online security and equipment protection and technical support, we should focus on the promotion and introduction for users, such as free experience for the first month/half a year. Contract aspect: Those single-month contract users should be provided annual contract payment discounts, which can effectively transform monthly contract users into annual contract users, further increase users' online time and achieve higher user tenure. It is strongly recommended to launch targeted coupons for those who pay by electronic check and then change their pay method to encourage them to do so.

Conclusion of building the user churn prediction model: A satisfied result about the precision of the user churn prediction model has been achieved by the gradient boosting tree algorithm. After the grid search tuning, the five evaluation index values of the model achieve above 0.8, but only 188 wrong predictions when testing on 2,113 users. The result can not only help telecom operation managers or decision makers to accurately predict the churn of some customers, but convenient for telecom staff to formulate some targeted strategies to retain those potential customers.

Reference

[1] Cheng Qiyun, Sun Caixin, Zhang Xiaoxing, et al. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic. Transactions of China Electrotechnical Society, 2004, 19(10): 53-58.

[2] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[3] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[4] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[5] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[6] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[7] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[8] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[9] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[10] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[11] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[12] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[13] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.