A Multi-label Satisfaction Prediction Model Based on Extreme Gradient Boosting Trees and Ensemble Classifier Chains

Xuefen Liu[1, a], Zebin Ma[1, b], Yonglin Zou[1, c] and Ying Zhang[1, d ,*]

[1]School of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang, China

[a]liuxuefen@stu.gdou.edu.cn, [b]mazebin@stu.gdou.edu.cn, [c]zouyonglin@stu.gdou.edu.cn,
[d]zhangying@gdou.edu.cn

Abstract

In order to further improve the quality of network service, all mobile operators need to pay attention to the network experience. Based on the user experience, we sort out the factors that affect the customer's online experience, so as to analyze the main factors that affect the customer's online business experience, and improve the customer's online experience. Based on the XGBoost model node classification gain theory, this paper establishes the XGBoost model and substitutes the preprocessed training data into the model for grid search to obtain the optimal parameter combination, and then brings the best parameter combination into the model, using the best performance The optimized XGBoost algorithm sorts the importance of features, and obtains the top five main factors that affect the satisfaction of voice services and Internet services. In order to solve the problem of multi-objective classification (4 sets of satisfaction labels), this paper constructs a combined multi-objective classification model based on Classifier Chains and ensemble learning, and uses the Accuracy index as the model evaluation. According to the standard, the model with the best performance is selected for integration, and the verification set data is substituted into the model as data to obtain 4 sets of satisfaction label scoring values. Data experiments show that the model has achieved good results in predicting multiple scoring items. This classification method has certain reference and practical significance.

Keywords

XGBoost; Classifier Chains; Ensemble Learning; Multi-Objective Classification.

1. INTRODUCTION

In the context of the current development of science and technology, some advanced technologies, such as mobile communication technology, have been widely used in real life, changing people's lifestyles and ways of thinking. The rapid development of mobile communication technology makes people enjoy the great convenience brought by mobile communication technology. Moreover, with the gradual improvement of network coverage, mobile communication operators have also begun to realize the importance of customers' network experience, and gradually improve the service quality of their own networks.

In the mobile communication service industry, operators can use the index of customer satisfaction to measure the matching degree of customer expectations and customer experience, so as to understand the needs of customers, the problems existing in enterprises and the differences with competitors, so as to improve service work in a targeted manner. In today's era of information explosion, data transparency, and homogeneous operations, the operating status of each mobile operator market is mainly reflected in customer satisfaction, especially in today's gradual development of the digital economy. Combined use of means, focus on establishing a comprehensive system evaluation system with customer experience as the first, to meet the realization of digital transformation of customer satisfaction evaluation, make customer experience and mobile operation business service decision-making, and promote high-quality and sustainable mobile communication technology develop.

One of the traditional ways to improve customer satisfaction is to solve the problems that affect customer experience bit by bit based on the feedback of customer experience. However, the sharp increase in the number of customers and various types of mobile products make it impossible for traditional methods to effectively address customer needs and improve customer satisfaction. Therefore, this research plan conducts a comprehensive analysis through various factors that affect customer satisfaction, and provides methods for business service decision-making to improve customer satisfaction earlier and more comprehensively.

China Mobile Communications Group Beijing Branch, in order to improve customer satisfaction with voice services, allows customers to evaluate the overall experience of voice calls in daily life, network coverage and signal strength, clarity of voice calls, and stability of voice calls. The experience is scored, and the factors that affect the customer's voice service experience are also counted to further analyze the main influencing factors of customer satisfaction with the voice service; Experience, network coverage and signal strength, mobile Internet speed, and mobile Internet stability are scored in four aspects, and the factors that affect customer Internet service experience are also counted to further analyze the main factors affecting customer satisfaction with voice services.

Based on the above research background, this paper needs to solve the following two research task:

Task 1: Mining the main influencing factors of customers' satisfaction with voice service and Internet service.

According to the existing data, analyze the factors of customers' voice service satisfaction and Internet service experience respectively, and dig out the main factors affecting customer voice service satisfaction and Internet service experience. At the same time, it is necessary to give the influence degree of each factor on customer scoring Quantitative analysis and results.

Task 2: Establish a customer scoring prediction model based on influencing factors.

Combined with the analysis of task 1, based on relevant influencing factors, a mathematical model for customer ratings is established for customer satisfaction with voice services and online services, and prediction research on customer ratings is carried out accordingly.

2. MODELS ESTABLISHMENT

2.1. Establishment of XGBoost Model

Based on the Boosting framework, the XGBoost algorithm is a synthetic algorithm that combines basis functions and weights to fit data. It has very powerful capabilities in missing value processing and prediction. In order to analyze the data set in an all-round way, we established the XGBoost model to obtain the top 10 influencing factors that affect customer satisfaction with voice services and online services. The process of establishing the model is as follows [1]:

First, for the 5428 data sets of voice service user satisfaction and the 7020 data sets of Internet service user satisfaction in this problem, the XGBoost model can be expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \quad (i = 1, 2, \ldots 5428) \tag{1}$$

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \quad (i = 1, 2, \ldots 7020) \tag{2}$$

where $\hat{y}_i$ represents the predicted value of the ith item, and k represents the decision tree of the k decision tree,

$$F = \left\{ f(x) = w_{q(x_i)} \right\} \left(q : R^K \to \{1, 2, \ldots, T\}, w \in R^T \right) \tag{3}$$

Indicates the set of CART decision tree structures, q is the tree structure where samples are mapped to leaf nodes, T is the number of leaf nodes, and w is the real number score of leaf nodes. When constructing the XGBoost model, it is necessary to find the optimal parameters according to the principle of minimizing the objective function, so as to establish the optimal model. The objective function of voice service user satisfaction Obj_1 can be divided into error function L_1 and model complexity Ω, and the objective function of Internet service user satisfaction Obj_2 can be divided into error function L_2 and model complexity Ω. The formula is as follows:

$$Obj_1 = L_1 + \Omega \tag{4}$$

$$L_1 = \sum_{i=1}^{5428} (y_i - \hat{y}_i)^2 \tag{5}$$

$$Obj_2 = L_2 + \Omega \tag{6}$$

$$L_2 = \sum_{i=1}^{7020} (y_i - \hat{y}_i)^2 \tag{7}$$

$$\Omega = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \tag{8}$$

where γ and λ are two hyperparameters, which are used to control the intensity of punishment.

For the XGBoost model, when using the training set to optimize the model, it needs to keep the original model unchanged, and add a new function f to the model. The specific process is as follows:

$$\begin{gathered} \hat{y}_i^{(0)} = 0 \\ \hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} = \hat{y}_i^{(1)} + f_2(x_i) \\ \text{B B} \\ \hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \end{gathered} \tag{9}$$

where $\hat{y}_i^{(t)}$ represents the predicted value of the t-th model, and $f_t\left(x_i\right)$ represents the new function added by the tth model. At this time, the objective function can be expressed as:

$$Obj_1^{(t)} = \sum_{i=1}^{5428}\left(y_i - \left(\hat{y}_i^{(t-1)} + f_t(x_i)\right)\right)^2 + \Omega \tag{10}$$

$$Obj_2^{(t)} = \sum_{i=1}^{7020}\left(y_i - \left(\hat{y}_i^{(t-1)} + f_t(x_i)\right)\right)^2 + \Omega \tag{11}$$

It is worth noting that after adding the new function *f*, in order to improve the overall expression effect of the XGBoost model, the corresponding objective function Obj should be reduced, that is, the objective function Obj needs to be iteratively optimized next. In order to quickly find the parameters that can make the objective function Obj obtain the minimum value, it is necessary to expand the second-order Taylor of the objective function Obj, which can be obtained:

$$Obj(t)_1 \approx \sum_{i=1}^{5428}\left[\left(y_i - \hat{y}^{(t-1)}\right)^2 + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)\right] + \Omega \tag{12}$$

$$Obj(t)_2 \approx \sum_{i=1}^{7020}\left[\left(y_i - \hat{y}^{(t-1)}\right)^2 + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)\right] + \Omega \tag{13}$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}}\left(y_i - \hat{y}^{(t-1)}\right)^2$, $h_i = \partial^2_{\hat{y}_i^{(t-1)}}\left(y_i - \hat{y}^{(t-1)}\right)^2$.

Considering that only variables need to be considered in the process of optimizing the objective function Obj, removing the constant term can be obtained:

$$\begin{aligned} Obj(t)_1 &= \sum_{i=1}^{5428}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_i^2(x_i)\right] + \Omega \\ &= \sum_{i=1}^{5428}\left[g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i=I_i} g_i\right) w_j + \frac{1}{2}\left(\sum_{i=I_i} h_i + \lambda\right) w_j^2\right] + \gamma T \end{aligned} \tag{14}$$

$$\begin{aligned} Obj(t)_2 &= \sum_{i=1}^{7020}\left[g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega \\ &= \sum_{i=1}^{7020}\left[g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i=I_i} g_i\right) w_j + \frac{1}{2}\left(\sum_{i=I_i} h_i + \lambda\right) w_j^2\right] + \gamma T \end{aligned} \tag{15}$$

where $I_j = \{i| \ q(x_i) = j\}$ means that the jth leaf node stores the collection of samples falling on the leaf node.

Next, to find the optimal score of each leaf node j and its corresponding optimal objective function Obj $\frac{\partial Obj^{(t)}}{\partial w_j} = 0$ can be obtained:

$$w_j^* = \frac{-\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{16}$$

$$-\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{17}$$

Here, Obj is a scoring function that can be used as an evaluation model, and the smaller the value of Obj, the better the effect of the model.

By recursively calling the above tree building method, we can get a large number of regression tree structures, and can search for the optimal tree structure through the objective function, and add it to the existing model, so as to continuously build the optimal XGBoost model.

2.2. Establishment of the Combined Multi-Objective Classification Model of Classifier Chains + Integrated Learning

In general, methods to solve multi-label classification problems include problem transformation methods, adaptive algorithms, ensemble methods, etc. After experimental evaluation, we decided to adopt a combined multi-objective classification model of Classifier Chains + ensemble learning.

Classifier chaining is a machine learning method for problem transformation in multi-label classification. It combines the computational efficiency of binary correlation methods, while allowing for classification by taking label dependencies into account [2].

(1) Problem transformation (single-objective classification to multi-objective classification)

The classifier chain method is based on Binary Relevance and works well even on a large number of labels. In addition, it takes into account dependencies between tags. The principle of this method is: Given a set of labels and a data set L, its form is, *x* is a feature vector, $\mathrm{Y} \subseteq \mathrm{L}$ is a set of labels assigned to the instance. BR converts dataset to $|\mathrm{L}|$ dataset and learns $|\mathrm{L}|$ binary classifier $\mathrm{H{:}X} \to \{\mathrm{l}, \neg \mathrm{l}\}$ for each label $\mathrm{l} \in \mathrm{L}$.

(2) Principle of classifier chain

For a given set of labels L, the classification chain model (CC) learns $|\mathrm{L}|$, the same as the classifier in the binary correlation method. All classifiers are connected in a chain through the feature space.

Suppose a dataset is given where the ith instance is of the form (x_i, Y_i), Y_i is a subset of labels, and x_i is a set of features. Dataset to be transformed into a dataset of instances of $|\mathrm{L}|$.

where the jth data set is of the form $((x_i, l_1, ..., l_{j-1}), l_j)$, $l_j \in \{0, 1\}$. l_j is 1, if the jth label is assigned to the instance, and 0 otherwise. The classifiers thus establish a chain, where each classifier learns the binary classification of a single label. The features given to each classifier have binary values indicating which of the previous labels was assigned to the instance.

By classifying a new instance, by building a chain of classifiers, the label is again predicted. Classification starts from the first classifier C_1 and proceeds to the last $C_{|L|}$, passing label information between classifiers through the feature space. Therefore, inter-label dependencies are preserved. However, for different chain orders, the result will be different. For example, if a label frequently co-occurs with other labels, only instances of the label appearing later in the chain have information about the other labels in their feature vectors. To address this issue and improve accuracy, we use an Ensemble of classifiers [3].

In an ensemble of classifier chains (ECC), several CC classifiers can be trained on random subsets of the dataset in the order of random chains (i.e., random order of labels). The label of a new instance is predicted by each classifier separately. Afterwards, the total number of predictions or "votes" for each label is calculated. A label is accepted if it is predicted by a certain percentage of the classifiers and is greater than a certain threshold.

At the code level, we maintain a separate classifier object for each Target Column. The correlation between tags is ordered by using a classifier chaining strategy and allowing each tag's classifier to be considered in the predictions conditioned on the predicted value of the tag that occurs earlier in the order.

(3) Establish a variety of machine learning integration models

In order to obtain the optimal performance, we have established a variety of machine learning models, and use the Bootstrap aggregating (Bagging) integrated learning method [4] to combine the models. The concept of Bagging is briefly introduced below.

Bootstrap aggregation (Bagging) is an ensemble learning method that combines the prediction results of different regression or classification models (the prediction results have a large variance) to reduce the variance. The results of the model are then averaged. The results predicted by each model are included in the prediction with equal weight. The specific principles are as follows:

Given a training set *D* of size *n*, the Bagging algorithm selects *m* subsets D_i of size n' from it evenly and with replacement (that is, using self-service sampling method) as a new training set. Using algorithms such as classification and regression on the *m* training sets, *m* models can be obtained, and the result of Bagging can be obtained by taking the average and taking the majority vote.

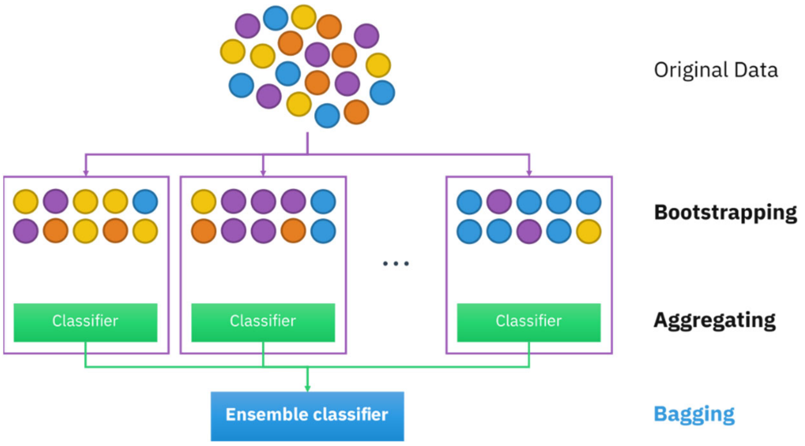

Figure 1. Schematic diagram of Bagging

3. RESULTS AND ANALYSIS

3.1. XGBoost Model Solution Results

(1) Ranking of feature importance—XGBoost algorithm

Tree models are often used for feature selection that is automated by machine learning algorithms. The core of obtaining the importance scores of each feature lies in the calculation of information gain. Information gain is the degree to which a certain condition divides the information uncertainty, and the tree model uses the information gain criterion to select the characteristic variables.

The tree structure is realized through node splitting to form left and right subtrees. Therefore, whether the tree continues to split can be determined by the information gain before and after node splitting, that is, the sum of the structure scores of the left and right subtrees and the structure score of the tree before splitting are calculated. Difference. The calculation formula of XGBoost's node splitting gain is as follows [1]:

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L+\gamma}+\frac{G_R^2}{H_R+\gamma}-\frac{(G_L+G_R)^2}{H_L+H_R+\gamma}\right)-\gamma \tag{18}$$

In the above formula, $\frac{G_L^2}{H_L+\gamma}$ and $\frac{G_R^2}{H_R+\gamma}$ represent the structure scores of the left subtree and right subtree respectively after the node is split, and $\frac{(G_L+G_R)^2}{H_L+H_R+\gamma}$ represents the structure score of the leaf node when it is not split.

By selecting the node whose feature is tree splitting when (Gain > 0) is the maximum value, by calculating the information gain of each feature variable and sorting them, the feature with the best regression ability can be obtained.

We first substitute the preprocessed training data into the above XGBoost model, and first adjust the parameters of the model through Jupyter software. The parameter adjustment range is shown in Table 1.

Table 1. XGBoost model tuning range

Parameter Name	Parameter Explanation	Parameter Range
n_estimators	The Number of Decision Trees	[5，20，30，40，50，100，500]
max_depth	Tree Depth	[3, 4, 5, 6 ,9, 10]
min_child_weight	The Minimum Sample Weight Required On a Leaf Node	[1, 2, 4, 5, 6]
learning_rate	Learning Rate	[0.01, 0.05, 0.1]
gamma	One of The Hyperparameters	[0.001，0.1]

Then, through grid search on the XGBoost model, the best parameter combination was successfully found, as shown in Table 2 below:

Table 2. XGBoost model optimal parameter combination

Parameter name	Optimal parameter value
n_estimators	100
max_depth	6
min_child_weight	1
learning_rate	0.3
reg_alpha	0
reg_lambda	1
gamma	0

Among them, reg_alpha and reg_lambda cannot be adjusted due to insufficient running memory, so the default values are taken here.

Then we brought the best parameter combination into the model. For the 4 sets of satisfaction labels of customer voice services, this paper establishes 4 sets of labels by inputting a training set of 5428 samples, which are the overall satisfaction of voice calls, network coverage and signal strength. , voice call clarity, voice call stability XGBoost model, obtained the top 10 influencing factors for each group of satisfaction labels, as shown in Table 3~6.

Table 3. TOP10 Feature Importance Scores for Overall Voice Call Satisfaction

Feature Id	Influence Factor	Feature Importance Score
11	Noisy, Hard to Hear, Intermittent During Calls	0.034007
8	Mobile Phone No Signal	0.018049
592	Terminal Brand Type_Nam-Al00	0.013231
10	Sudden Interruption During a Call	0.013216
0	Residential Area	0.012916
562	Terminal Brand Type_M2011K2C	0.012085
2	Colleges and Universities	0.010782
606	Terminal Brand Type_Oce-An10	0.010777
9	There Is a Signal and Cannot Be Dialed	0.010198
571	Terminal Brand Type_M2104K10Ac	0.009916

Table 4. TOP10 Feature Importance Scores of Network Coverage and Signal Strength

Feature Id	Influence Factor	Feature Importance Score
0	Residential Area	0.038865
11	Noisy, Hard to Hear, Intermittent During Calls	0.021222
8	Mobile Phone No Signal	0.014432
13	One Party Cannot Hear During a Call	0.011108
1	office	0.011023
592	Terminal Brand Type_Nam-Al00	0.010997
10	Sudden Interruption During a Call	0.010075
32	4\5G User_2G	0.009432
954	Proportion of Traffic From Other Provinces_100.00%	0.008427
857	Terminal Brand Type_Vog-Al10	0.008294

Table 5. Top 10 Feature Importance Scores for Voice Call Clarity

Feature Id	Influence Factor	Feature Importance Score
11	Noisy, Hard to Hear, Intermittent During Calls	0.037266
0	Residential Area	0.019965
301	Terminal Brand Type_0	0.010690
637	Terminal Brand Type_Pbem00	0.010635
8	Mobile Phone No Signal	0.010024
2	Colleges and Universities	0.009963
510	Terminal Brand Type_Lya-Al10	0.009051
530	Terminal Brand Type_M1901F7Be	0.008974
597	Terminal Brand Type_Noh-Al10	0.008542
1	office	0.008451

Table 6. Top10 Feature Importance Scores for Voice Call Stability

Feature Id	Influence Factor	Feature Importance Score
0	Residential Area	0.038682
11	Noisy, Hard to Hear, Intermittent During Calls	0.023515
10	Sudden Interruption During a Call	0.017172
571	Terminal Brand Type_M2104K10Ac	0.012644
428	Terminal Brand Type_Evr-Al00	0.011771
8	Mobile Phone No Signal	0.011527
904	Terminal Brand Type_Iphone Xr(A1984)	0.011224
125	Proportion of Speech From Other Provinces_3.10%	0.010810
13	One Party Cannot Hear During a Call	0.010162
353	Terminal Brand Type_Alp-Al00	0.008797

From Table 3~6 we can draw:

For the overall satisfaction of voice calls, noise, inaudible, and intermittent calls are the most important feature influencing factors, and its feature importance score is as high as 0.034, followed by no mobile phone signal, with a feature importance score of 0.018.

For the satisfaction of network coverage and signal strength, customers believe that network problems in residential areas are the most important influencing factors, and the feature importance score is 0.038. The second is that there are noises, inaudible, and intermittent calls during the call. The feature importance score is 0.021. No signal on the mobile phone also affects satisfaction. The feature importance score is 0.014. It may be that when the user has no signal on the mobile phone, it will affect the network. Coverage satisfaction drops.

For the satisfaction of voice call clarity, noise, inaudible, and intermittent are the most important influencing factors in the call, and the feature importance score is 0.037. The second is that there is a network problem in the residential area, and the feature importance score is 0.019. It may be related to the laying of communication base stations in residential areas, or it may come from the type of terminal brand in the hands of users.

For the satisfaction of voice call stability, the presence of network problems in residential areas is the most important influencing factor, and the feature importance score is 0.038. Secondly, there are noises, inaudible, and intermittent calls during the call. Sudden interruption during the call will also affect the satisfaction. The feature importance score is 0.023. Maybe the user is satisfied with the stability of the voice call due to interruption problems during the call. Spend.

For the 4 sets of satisfaction labels of customers online business, this paper inputs a training set of 7020 samples to establish an XGBoost model for 4 sets of labels, which are the overall satisfaction of mobile Internet access, network coverage and signal strength, mobile Internet access speed, and mobile Internet stability. , get the top 10 influencing factors for each group of satisfaction labels, as shown in Table 7~10.

Table 7. Top 10 Feature Importance Scores of Mobile Internet Overall Satisfaction

Feature Id	Influence Factor	Feature Importance Score
8	Poor Network Signal/No Signal	0.035388
11	Mobile Internet Speed Is Slow	0.015339
10	The Network Is Intermittent or Fast and Slow During The Process of Surfing The Internet	0.015024
14	Game Lag Is High	0.013738
29	All Stuck	0.007313
0	Residential Area	0.007290
9	It Shows That There Is a Signal and Cannot Connect to The Internet	0.006816
50	Other Webpage or App Problems	0.006595
51	All Web Pages or Apps Are Slow	0.006447
1	office	0.006004

Table 8. Top 10 Feature Importance Scores of Network Coverage and Signal Strength Satisfaction

Feature Id	Influence Factor	Feature Importance Score
8	Poor Network Signal/No Signal	0.041117
11	Mobile Internet Speed Is Slow	0.014305
10	The Network Is Intermittent or Fast and Slow During The Process of Surfing The Internet	0.012098
14	Game Lag Is High	0.011691
15	Slow To Open Web Pages or App Pictures	0.008238
9	It Shows That There is a Signal and Cannot Connect to The Internet	0.007362
494	Terminal Brand Type_Pdvm00	0.007194
83	Terminal Brand_Apple	0.006569
0	Residential Area	0.006264
51	All Web Pages or Apps Are Slow	0.005978

Table 9. Top 10 Feature Importance Scores of Mobile Internet Speed Satisfaction

Feature Id	Influence Factor	Feature Importance Score
8	Poor Network Signal/No Signal	0.035719
11	Mobile Internet Speed is Slow	0.024523
10	The Network Is Intermittent or Fast and Slow During The Process of Surfing The Internet	0.016858
14	Game Lag Is High	0.008793
15	Slow to Open Web Pages or App Pictures	0.008649
0	Residential Area	0.006995
184	Terminal Brand Type_Brq-An00	0.006784
268	Terminal Brand Type_Jad-Al50	0.006738
51	All Web Pages or Apps Are Slow	0.006602
9	It Shows That There Is a Signal and Cannot Connect to The Internet	0.006042

Table 10. Top 10 Feature Importance Scores of Mobile Internet Stability Satisfaction

Feature Id	Influence Factor	Feature Importance Score
8	Poor Network Signal/No Signal	0.034710
10	The Network Is Intermittent or Fast and Slow During The Process of Surfing The Internet	0.021057
11	Mobile Internet Speed Is Slow	0.013911
14	Game Lag Is High	0.011871
15	Slow to Open Web Pages or App Pictures	0.008248
9	It Shows That There Is a Signal and Cannot Connect to The Internet	0.007865
0	Residential Area	0.007055
87	Terminal Brand_Hammer	0.006956
29	All Stuck	0.006284
258	Terminal Brand Type_Hwi-Al00	0.006138

From Table 7~10 we can draw:

For the overall satisfaction of mobile Internet access, poor network signal/no signal is the most important feature influencing factor, and its feature importance score is as high as 0.035, followed by slow mobile Internet access, with a feature importance score of 0.015. Then, in the process of surfing the Internet, the network is intermittent or fast and slow, and the delay in playing games is large. The feature importance scores are 0.015 and 0.013 respectively. It can be seen that the overall satisfaction of users' mobile Internet access is mainly affected by network signals and Internet access. Speed and stability issues.

Regarding satisfaction with network coverage and signal strength, customers believe that poor network signal/no signal is the most important influencing factor, with a feature importance score of 0.041, followed by slow mobile Internet access, with a feature importance score of 0.014.

For mobile Internet speed satisfaction, customers believe that poor network signal/no signal is the most important influencing factor, and the feature importance score is 0.035. The second is that the slow speed of Internet access and intermittent or fast or slow Internet access affect customer satisfaction, and the feature importance scores are 0.024 and 0.016, respectively. It can be seen that the main factors affecting satisfaction are the slow Internet speed of the mobile phone and the lack of mobile phone signal.

Regarding the satisfaction with the stability of mobile Internet access, customers believe that poor network signal/no signal is the most important influencing factor, and the feature importance score is 0.034. Secondly, the slow speed of Internet access and intermittent or fast and slow Internet during the process of Internet access affect customer satisfaction, and the feature importance scores are 0.021 and 0.013, respectively. It may be because of the fast and slow time, which makes customers think that the stability of the Internet is not enough, which affects their satisfaction.

In this paper, by jointly comparing variables with feature importance scores greater than 0.3% in the XGBoost algorithm in different labels, and taking their intersection to obtain the feature variables jointly selected by the four, the main factors of the entire service satisfaction can be obtained, and finally the voice service satisfaction is screened out. The top 5 main factors are as follows:

Table 11. Top 5 Factors Affecting Voice Service Satisfaction

Serial Number	Feature Variable Name
1	Noisy, Hard to Hear, Intermittent During Calls
2	Residential Area
3	Sudden Interruption During a Call
4	Mobile Phone No Signal
5	Terminal Brand Type

The top 5 main factors affecting online business satisfaction are as follows:

Table 12. Top 5 Factors Affecting Internet Service Satisfaction

Serial Number	Feature Variable Name
1	Poor Network Signal/No Signal
2	The Network Is Intermittent or Fast and Slow During The Process of Surfing The Internet
3	Mobile Internet Speed Is Slow
4	Game Lag Is High
5	Slow to Open Web Pages or App Pictures

3.2. The Combined Multi-Objective Classification Model of Classifier Chains + Integrated Learning Solution Results

We used lightGBM[5], two K-nearest neighbors with different weights[6] (1. 'uniform': uniform weight, all points in each neighborhood are weighted equally; 2. 'distance': the reciprocal of the passing distance Points are weighted), the NeuralNet classification model [7] based on the deep learning attention network, and the weighted model of the above model. The training is based on Google Colab and Amazon machine learning platform (AWS machine learning platform). The training results are shown in Table 13~20.

Table 13. Prediction of Mobile Internet Speed

Model Name	Model Score (Accuracy)	Training Time	Integration Layers	Training order
NeuralNet_BAG_L2	0.622222	130.4425	2	5
WeightedEnsemble_L3	0.622222	130.4483	3	6
WeightedEnsemble_L2	0.619658	88.14607	2	4
NeuralNet_BAG_L1	0.614815	86.53028	1	3
KNeighborsDist_BAG_L1	0.478632	0.030883	1	2
KNeighborsUnif_BAG_L1	0.469801	0.017912	1	1

Table 14. Prediction of Mobile Internet Stability

Model Name	Model Score (Accuracy)	Training Time	Integration Layers	Training order
NeuralNet_BAG_L2	0.597721	132.1587	2	5
WeightedEnsemble_L3	0.597721	132.165	3	6
WeightedEnsemble_L2	0.590028	78.30121	2	4
NeuralNet_BAG_L1	0.589316	76.43293	1	3
KNeighborsDist_BAG_L1	0.467664	0.018891	1	2
KNeighborsUnif_BAG_L1	0.457977	0.02026	1	1

Table 15. Overall Satisfaction with Mobile Internet Access

Model Name	Model Score (Accuracy)	Training Time	Integration Layers	Training order
WeightedEnsemble_L2	0.431054	82.34116	2	4
NeuralNet_BAG_L2	0.427778	125.669	2	5
WeightedEnsemble_L3	0.427778	125.6748	3	6
NeuralNet_BAG_L1	0.426211	80.92295	1	3
KNeighborsUnif_BAG_L1	0.288034	0.017864	1	1
KNeighborsDist_BAG_L1	0.276923	0.012308	1	2

Table 16. Network coverage and signal strength (call)

Model Name	Model Score (Accuracy)	Training Time	Integration Layers	Training order
NeuralNet_BAG_L2	0.66028	129.3741	2	6
WeightedEnsemble_L3	0.66028	129.3796	3	7
WeightedEnsemble_L2	0.645357	99.04512	2	5
NeuralNet_BAG_L1	0.643884	67.45878	1	3
LightGBM_BAG_L1	0.590641	30.00896	1	4
KNeighborsUnif_BAG_L1	0.369934	0.034633	1	1
KNeighborsDist_BAG_L1	0.351879	0.049752	1	2

Table 17. Network coverage and signal strength (Internet access)

Model Name	Model Score (Accuracy)	Training Time	Integration Layers	Training order
NeuralNet_BAG_L2	0.584758	127.4823	2	5
WeightedEnsemble_L3	0.584758	127.4886	3	6
WeightedEnsemble_L2	0.582336	82.97496	2	4
NeuralNet_BAG_L1	0.576781	81.46588	1	3
KNeighborsDist_BAG_L1	0.433761	0.01545	1	2
KNeighborsUnif_BAG_L1	0.423932	0.015674	1	1

Table 18. Voice call clarity

Model Name	Model Score (Accuracy)	Training Time	Integration Layers	Training order
NeuralNet_BAG_L2	0.711127	131.5876	2	5
WeightedEnsemble_L3	0.711127	131.5954	3	6
WeightedEnsemble_L2	0.708917	74.95218	2	4
NeuralNet_BAG_L1	0.708732	73.60765	1	3
KNeighborsUnif_BAG_L1	0.422255	0.008423	1	1
KNeighborsDist_BAG_L1	0.402174	0.012502	1	2

Table 19. Voice call stability

Model Name	Model Score (Accuracy)	Training Time	Integration Layers	Training order
NeuralNet_BAG_L2	0.703758	131.1745	2	6
WeightedEnsemble_L3	0.703758	131.1801	3	7
WeightedEnsemble_L2	0.689388	91.16178	2	5
NeuralNet_BAG_L1	0.686993	66.16654	1	3
LightGBM_BAG_L1	0.662675	23.84896	1	4
KNeighborsUnif_BAG_L1	0.373434	0.007265	1	1
KNeighborsDist_BAG_L1	0.364407	0.012536	1	2

Table 20. Overall Satisfaction of Voice Calls

Model Name	Model Score (Accuracy)	Training Time	Integration Layers	Training order
WeightedEnsemble_L2	0.581061	92.81635	2	4
NeuralNet_BAG_L1	0.572402	91.45008	1	3
KNeighborsUnif_BAG_L1	0.459654	0.045139	1	1
KNeighborsDist_BAG_L1	0.432388	0.033556	1	2

This section builds a combined multi-objective classification model based on Classifier Chains and ensemble learning, using the Accuracy indicator as the model evaluation standard. As shown in the above table, we can find that the accuracy rate of the NeuralNet classification model based on the deep learning attention network is almost higher than that of other models. Therefore, we choose the NeuralNet classification model with the best performance as the prediction model, the verification set data is substituted into the model as data, and 4 sets of satisfaction label scoring values are obtained, that is, the prediction results.

4. CONCLUSION

In order to further improve the quality of network service, all mobile operators need to pay attention to the network experience. We need to sort out the factors that affect the customer's online experience based on the user's experience. We hope that we can analyze the main factors that affect the customer's online business experience and improve the customer's experience. Internet experience.

For task 1, the first step (data cleaning): This paper uses statistical indicators, data density maps and other methods to find out the missing values in attachments 1 and 2, and removes, fills, and quantifies the data in attachments 1 and 2 one by one; The second step (feature engineering): This article adopts one-hot encoding for the category data in Annex 1 and 2, and then uses feature construction to combine home broadband complaints and tariff complaints into one field whether there is a complaint, and then perform feature description. There are 38 voice services in total. field, a total of 62 fields for online business; the third step (visualization): use the pandas-profiling tool to generate a research report, and statistically describe the characteristic variables by drawing a correlation coefficient heat map; the fourth step (model preparation): this article is based on XGBoost model node classification gain theory, by establishing the XGBoost model and substituting the preprocessed training data into the model for grid search, the optimal parameter combination is obtained, and then the best parameter combination is brought into the model, and the optimal performance is adopted The XGBoost algorithm sorts the importance of features; the fifth step (model solution): In this paper, the training sets of 5428 and 7020 samples are respectively input to establish the XGBoost model of 4 groups of voice service and Internet service label satisfaction, and the four groups are obtained through joint intersection. Based on the characteristic variables selected by the researchers, it can be concluded that the top five factors that affect the satisfaction of voice services are noise during the call, inaudible, intermittent, residential area, sudden interruption during the call, no signal on the mobile phone, and terminal brand. Type, the top five main factors that affect the satisfaction of online business are poor network signal/no signal, intermittent or fast and slow network during the process of surfing the Internet, slow mobile Internet access, long delay in playing games, opening web pages or APPs Pictures are slow.

For task two, in order to solve the problem of multi-objective classification (4 sets of satisfaction labels), this paper constructs a combined multi-objective classification model based on Classifier Chains and ensemble learning, using the Accuracy index as a model evaluation criterion. First, feature engineering is performed on the data set, and then models such as LightGBM, attentional neural network, and K-nearest neighbors are trained for each column, and according to the evaluation results of the above models, the model with the best performance is selected for integration as a prediction model. Substituting the verification set data into the model as data, four groups of satisfaction label scoring values are obtained. Data experiments show that the model has achieved good results in predicting multiple scoring items. This classification method has certain reference and practical significance.

REFERENCES

[1] Wu Guangyu. Research on Prediction of Patient Satisfaction in Online Medical Community Based on XGBoost Algorithm [D]. Kunming University of Science and Technology, 2021.

[2] Read J, Pfahringer B, Holmes G, et al. Classifier chains for multi-label classification[J]. Machine learning, 2011, 85(3): 333-359.

[3] Rokach L. Ensemble-based classifiers[J]. Artificial intelligence review, 2010, 33(1): 1-39.

[4] Breiman L. Bagging predictors[J]. Machine learning, 1996, 24(2): 123-140.

[5] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30.

[6] Altman N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175-185.

[7] Arik S Ö, Pfister T. Tabnet: Attentive interpretable tabular learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(8): 6679-6687.