A Study of Mobile User Satisfaction Based on Feature Extraction

Yao Lu *

School of Mathematics, Nanjing Audit University, Nanjing , China

* Corresponding Author Email: 213080520@stu.nau.edu.cn

Abstract. China Mobile is the largest mobile communication operator in China, and with the promotion of its 5G applications, the issue of continuously improving its user satisfaction has become an important goal for sustainable development in the future. In this paper, data pre-processing operations were firstly performed, for data deletion and supplementation, data normalization, null filling and other steps; based on different machine learning algorithms, data feature extraction was performed to construct an effective satisfaction prediction model; entropy value method, XGBoost algorithm and lightgbm algorithm were used to train the model for prediction. The confusion matrix plot of the full variable test set was obtained by the XGBoost method, which shows that the model has some reasonableness and realistic significance.

Keywords: XGBoost; Lightgbm Algorithm; China Mobile; Feature Extraction.

1. Introduction

The telecom industry is changing rapidly, along with the arrival of the 5G era and the acceleration of the opening process, China Mobile will face greater market pressure and competitive pressure [1]. In order to further develop the market in the future, the strategy against competitors requires China Mobile to focus on the four customer brands and increase the loyalty of each customer brand sub-brand, so as to further develop the market, dig deeper into the market and combat competitors. Therefore, in order to enhance its competitiveness, China Mobile should further strengthen the differentiation of its customer brands in terms of business, service, and other aspects, and enhance the attractiveness of China Mobile to its customers with continuous innovation [2].

Therefore, mobile should provide a basis for decision making by analyzing various factors that affect user satisfaction, so as to achieve earlier and more comprehensive improvement of user satisfaction, keep customers in their own domain, and enhance customer loyalty to the brand with excellent customer experience [3]. Therefore, the algorithm model based on mobile user satisfaction is important to help improve the quality of network services, so that the customer's experience of using the network is greatly improved [4].

2. Data set source and pre-processing

2.1. Data set source

The data set used in this paper is from the mathor cup college data modeling challenge. In order to simplify the model and calculation, the following assumptions are made in this paper: (I)some variables are replaced with plastic numbers instead of real-life names to facilitate modeling; (II) all model building is entirely derived from the data given by the tournament questions, and no reference will be made to data from other sites for modeling[5];(III) the less necessary information users with too many missing samples can provide for effective prediction;(IV)the different algorithms based on the geometric mean fusion algorithm give the same weights and do not affect the fusion results[6].

2.2. Data pre-processing

Due to the long time span, complicated basic configuration information and wide range of satisfaction-related features, the original data set may have missing values and outliers, which may

affect the results of data mining and analysis in this paper, so this paper will first perform preliminary data cleaning on the original data set.

3. Data Analysis

3.1. Individual variables and satisfaction analysis

There are multiple discrete variables in the data set and there is no meaningful size comparison between the values of the variables. Therefore, in order to expand the role of features to solve the problem of discrete values of attribute data, this paper performs the coding operation of category features [7].

Since the exact value of the samples in the category features is not the concern of this paper, the discrete category features are transformed into category coding in this paper, which allows the model to ignore the variance of the values taken by the category features and reduces the introduction of unnecessary interference [8]. In this paper, the category features are transformed into one-hot encoding to eliminate the distance between all categories of each feature, which solves the problem that the regression model does not handle attribute data well, as well as plays a role in expanding the features to a certain extent and makes the distance calculation more reasonable [9].

There are 8 category-based features in the data, namely: terminal brand type', 'current month ARPU', 'current month MOU', 'previous March ARPU', 'previous March MOU', 'provincial traffic share', 'total GPRS traffic (KB)', 'GPRS - domestic roaming - traffic (KB)", and the statistical results of each category of features are as follows in Figure 1.

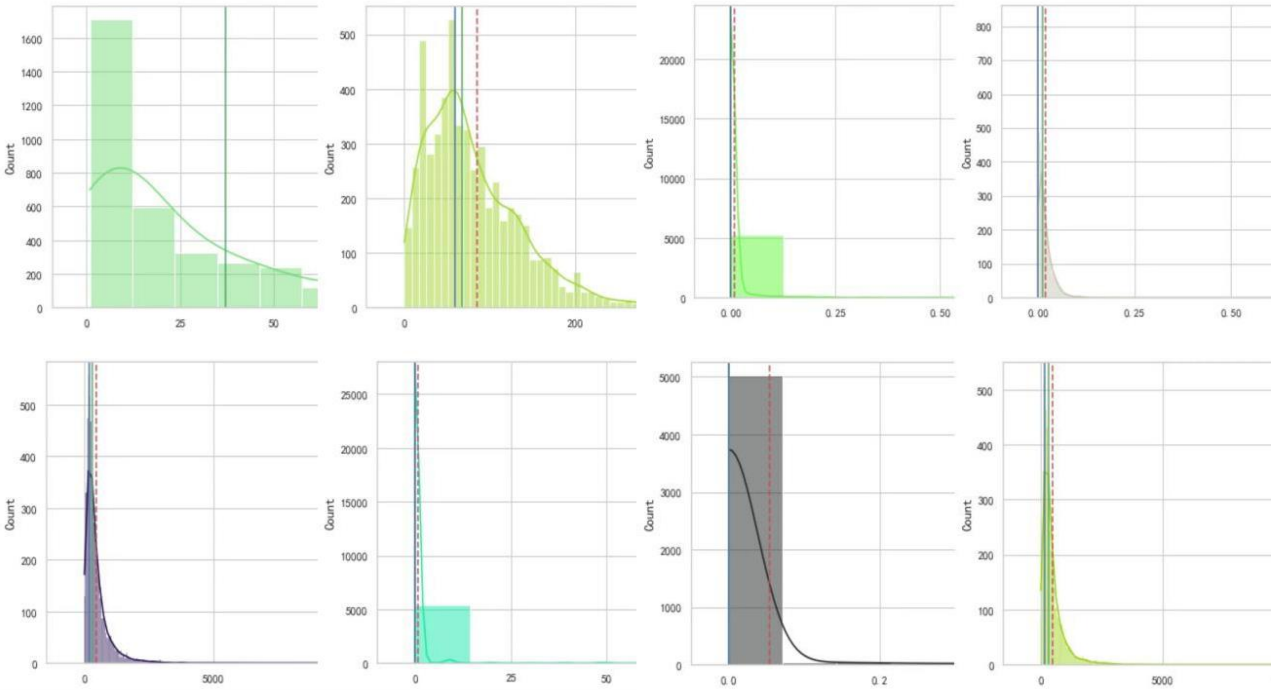

Figure 1. Satisfaction graph for individual variables of voice calls

3.2. Correlation analysis

Among the correlation coefficient methods there are Pearson (pearson) correlation coefficient and Spearman (Spearman) correlation coefficient and chi-square test analysis. Pearson correlation coefficient is a measure of the degree of linear correlation, describing the linear correlation between the variable and the dependent variable [10].

In statistics, the Spearman rank correlation coefficient is named after Charles Spearman and often denotes its value with the Greek letter ρ(rho) [11]. The Spearman rank correlation coefficient is used to estimate the correlation between two variables X, Y, where the correlation between the variables

can be described using a monotonic function [12]. If neither of the two sets in which the two variables take values have the same two elements, then ρ between the two variables can reach +1 or -1 when one of the variables can be expressed as a good monotonic function of the other (i.e., the two variables have the same tendency to change) [13].

Suppose two random variables are X and Y (which can also be considered as two sets), both of them have N number of elements, and the *i*-th (1<=*i*<=N) value taken by the two random variables is denoted by X_i and Y_i, respectively. The calculation is shown as follows:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{1}$$

4. Feature extraction

Before using the algorithm, the data are normalized in this paper, and the features in the dataset are converted to the same magnitude, thus eliminating any negative impact that different magnitudes may have on the algorithm.

4.1. Entropy value method

Entropy refers to a degree of chaos, defining that the more chaotic something is, the greater its entropy value, and the neater something is, the greater its entropy value. The entropy method uses the concept of entropy to assign weights to indicators. Based on the above idea, the entropy method is established. It is based on the fluctuation of the data itself and is not subjective in nature.

Assume that the original data matrix of m samples of n indicators is

$$X_{ij} = /left(X_{ij}/right)_{m/timesn} \tag{2}$$

Calculate the composite score level for each sample:

$$F_i = \sum_{j=1}^{n} W_j X'_{ij} \tag{3}$$

Based on the entropy method, the Internet service user satisfaction data were further screened, and the data weight matrix calculated for each feature and the "overall satisfaction of voice calls" was used to calculate and statistically obtain the six features with the highest correlation, which are: 'whether or not to visit the business hall, whether or not to care for the user, whether or not to be a 4G network customer (local exclude IoT), user description.1, user description, redirection times, redirection residence time, terminal brand', Whether 4G network customers (local exclude IoT), user description.1, user description, redirection number, redirection residence time, terminal brand'. These features are extracted, so that the Internet service user satisfaction data are 39 main features.

Similarly, the indicators with importance greater than 0 are extracted as the main factors affecting the voice service as follows in Table 1:

Table 1. The main factors affecting voice services

Characteristics	Importance
Whether encountered network problems	0.073728
Commercial street	0.028324
Subway	0.101304
High-speed rail	0.053029
Other network problem areas	0.013298
Can't dial with signal	0.140409
Other network problems	0.005016
ARPU (home broadband)	0.034149
Out-of-suite traffic (MB)	0.091799
Out-of-province voice share	0.076236
Inter-provincial roaming - hours (minutes)	0.032996
Foreign traffic share	0.230994
Amount in arrears for the current month	0.058884
Amount owed for the previous 3rd month	0.080433
Whether care user	0.003317
Terminal brand	0.341147
Customer star identification	0.261705

4.2. LightGBM model

LightGBM is an efficient framework for implementing the GBDT algorithm. It is similar to XGBoost in principle, but it is 10 times faster than the XGB model in processing data, and its memory usage is only 1/6 of the XGB model, and its accuracy is also improved. Leaf-wise leaf growth algorithm with depth limitation is used to ensure high efficiency and prevent overfitting.

4.3. XGBoost model

The XGBoost algorithm is an integrated learning method based on the CART decision tree model, which provides the accuracy of the prediction model by constructing multiple CART decision trees, and finally sums the prediction results of each training round to obtain the final prediction value. The objective function of its t_{th} CART decision tree is defined as follows:

$$L^{(t)} = \sum_{i=1}^{n}(y_i, y_i'^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{4}$$

Among them.

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \tag{5}$$

Using Taylor's formula for the expansion of the objective function:

$$f(x + \triangle x) = f(x) + f'(x) \triangle x + \frac{1}{2}f''(x) \triangle x^2 \tag{6}$$

The second-order Taylor expansion of the loss function results in

$$\sum_i L(y_i, \hat{y}_i^{t-1} + f_t(x_i)) = \sum_i [L(y_i, \hat{y}_i^{t-1}) + L'(y_i, \hat{y}_i^{t-1})t(x_i) + \frac{1}{2}L''(y_i, \hat{y}_i^{t-1})f_t^2(x_i)] \tag{7}$$

In the feature node selection, the values of all feature variables in the training set are iterated, and the gain value is calculated by subtracting the sum of the objective function values of the 2 leaf nodes after splitting from the objective function values of the nodes before splitting to obtain the optimal cut point of the tree model, where the gain value is calculated as Equation (8), Xgboost all-variable test set obfuscation matrix diagram can be shown in Figure 2.

$$L_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i+\lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i+\lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i+\lambda}\right] - \gamma \quad (8)$$

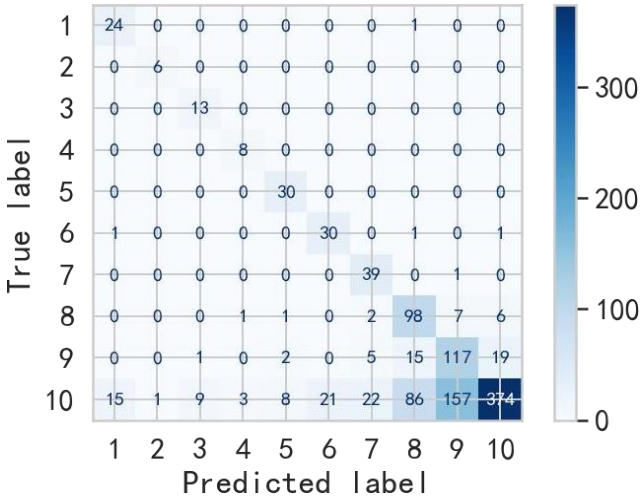

Figure 2. Xgboost all-variable test set obfuscation matrix diagram

5. Conclusions

In this paper, based on the significance and correlation analysis of the impact of each feature on the output, mathematical models for predicting customer scoring results are established separately for voice and Internet services, and the entropy value method, xgboost algorithm, and lightgbm algorithm are used for model training and prediction. The confusion matrix plot of the full variable test set is obtained by the Xgboost method, which shows that the model has certain reasonableness and realistic significance.

References

[1] Li D.X., Wu C.L., Zou L.F. Study on the factors influencing customer satisfaction of Meituan take-out platform under O2O model Advances in Applied Mathematics 11, 5536, 2022

[2] Xu JW, Yang Y. Integrated learning methods: A review of research. Journal of Yunnan University (Natural Science Edition) 40 (6), 1082-1092, 2018

[3] Dong Yingying, Ge Yang, Li Kunshu, Shen Bin, Huang Shuangshuang. Research on the application of fusion model in mobile network user satisfaction prediction Post & Telecom Design Technology, 08, 2022

[4] Zhang Yuchun. Research on happiness prediction and enhancement path based on stacked fusion method. Pure Mathematics 12, 1679, 202

[5] Bebe Wang. Research on customer churn hierarchical prediction based on stacked integration learning in Zhejiang Mobile Company. Zhejiang University of Industry and Commerce, 2018

[6] Ying Weiyun. Research on random forest method and its application in customer churn prediction. Management Review 24(2), 140-145.2012

[7] Adeola Ogunleye, Qing-Guo Wang.XGBoost model for chronic kidney disease diagnosis. IEEE/ACM transactions on computational biology and bioinformatics 17(6), 2131-2140, 2019

[8] Tetiana Borysova, Grygorii Monastyrskyi, Anetta Zielinska, Mariusz Barczak, Development of innovative activities of urban public transport service providers: a multifactorial economic and mathematical model, Sumy State University, Annual 2019

[9] Jeff W Johnson, A heuristic method for estimating the relative weight of predictor variables in multiple regression, Multivariate behavioral research 35 (1), 1-19, 2000

[10] Tang Q, Xia GN, Zhang Xianquan, Long F. Customer churn prediction model based on XGBoost and MLP.2020 International Conference on Computer Engineering and Applications (ICCEA), 608-612, 2020

[11] Mohammad Javad Shabankareh, Mohammad Ali Shabankareh, Alireza Nazarian, Alireza Ranjbaran, Nader SeyyedamiriJournal of Relationship. A Stacking-Based Data Mining Solution to Customer Churn Prediction. Journal of Relationship Marketing 21 (2), 124-147, 2022

[12] Sivasankar Karuppaiah, NP Gopalan.Enhanced Churn Prediction Using Stacked Heuristic Incorporated Ensemble Model.Journal of Information Technology Research (JITR) 14(2), 174-186, 2021

[13] Anton Borg, Martin Boldt.Using VADER sentiment and SVM for predicting customer response sentiment. Expert Systems with Applications 162, 113746. 2020