A PREDICTIVE MODEL FOR E-COMMERCE CUSTOMER CHURN UNDER AN INTELLIGENT ALGORITHM

Yunting TUO[1]

Every enterprise has its target customers, and all customers are directly related to the profit and business success or failure of the enterprise, and this is especially true for e-commerce enterprises. Retaining customers is the basis of long-term operation of e-commerce enterprises. This paper built a prediction model of e-commerce customer churn situation through the Xgboost algorithm, an intelligent algorithm, to analyze and select the features of e-commerce customer churn to serve as the basis of model prediction of customer churn. The research results showed that the accuracy of the Xgboost prediction model for identifying different churn features was above 90% in most cases; the P for overall customer churn prediction was 92.17%, the R was 91.84%, the F_1 value was 92.00%, and the area under curve (AUC) was 0.90, and its prediction performance was better than Logistic Regression and naive Bayesian models. This paper shows that the model constructed based on the Xgboost algorithm can predict the churn of e-commerce customers.

Keywords: e-commerce; customer churn; Xgboost algorithm; feature selection.

1. Introduction

For e-commerce enterprises, customers are an important resource that each e-commerce enterprise competes for, and getting new customers is usually more expensive than keeping existing customers [1]. Therefore, timely detecting the probable defected customers and implementing retention strategies to improve dependence and customer stickiness has extremely important practical significance for an e-commerce enterprise to occupy the market share and increase the profit of the enterprise. Yang constructed an e-commerce customer churn prediction model based on an inline sequential optimization extreme value learning machine, verified with cases that the proposed model could improve the accuracy of customer churn prediction, greatly reduce the training time of e-commerce customer churn modeling, and improve the speed of e-commerce customer churn prediction [2]. Raja used prediction models based on machine learning techniques such as k-Nearest Neighbor (KNN), Random Forest (RF), and XGBoost to recognize the features that significantly affected customer churn and concluded that the the XGBoost classifier provided the higher accuracy score and F-score than KNN and RF classifiers [3]. Amatare et al. used a convolutional neural network (CNN) model

[1] Zhengzhou Railway Vocational & Technical College, Zhengzhou, Henan 450000, China, e-mail: tuo5293@163.com

to predict customer churn in the telecommunication industry and found through example analysis that The accuracy of CNN models CNN1 and CNN2 were 81% and 89%, respectively [4]. Zhu et al. constructed a least absolute shrinkage and selection operator (LASSO)-RF model to analyze airline membership data and found that the LASSO-RF model had higher prediction accuracy and stronger generalization ability for predicting customer churn probability compared to LASSO model or RF model alone [5]. Ahmed et al. proposed an integration stack and combined it with a boosting-based strategy to construct a customer churn prediction model and found through a case analysis that the proposed model was 50% more cost-efficient than the state-of-the-art integration model by example analysis [6]. In their study, they used three machine learning algorithms to predict churn and measured the prediction score by area under curve (AUC) and found that the proposed model was more accurate than previous studies using the same dataset [7]. This paper used the Xgboost algorithm to construct a prediction model of e-commerce customer churn and collected and processed the information data of e-commerce customers. Different from other literature using machine learning techniques, this paper selected the churn characteristics of e-commerce customers according to Pearson's correlation coefficient and feature weight to train the model and obtain the final e-commerce customer churn prediction results. This work provides a theoretical basis for the subsequent use of intelligent algorithms in modeling for predicting e-commerce customer churn.

2. Predictive model of e-commerce customer churn

2.1. E-commerce customer churn characteristics selection

The online to offline (O2O) e-commerce sector in China has developed rapidly over the years, more and more companies have started to reform and join e-commerce. The competition of e-commerce platforms has become more intense, and customer churn has become a problem that every major platform now encounters [8]. Customer churn in e-commerce is not equivalent to industries such as insurance and telecommunications, where there is a clear contractual binding between the customer and the company, and only when the contractual relationship ends or is not renewed is it considered as customer churn. The e-commerce industry has a non-contractual relationship with its customers, who can choose other stores to buy the products they need without any constraint. Therefore, there are many factors for e-commerce customer churn, such as customer service, logistics and delivery, product quality, store discount, number of store fans, and other aspects. However, the number of customer features is too large. If too few features are selected for customer churn, the prediction performance of the model will be inaccurate, but if too many features are selected, it will cause data redundancy. Therefore, the representative features are selected, and Table 1 shows the names and explanations of the selected features.

Table 1

Names and explanations of features

Feature Name	Name explanation
Gender	Customer gender
Age	Customer age
Purchase power	Customer buying power
Purchase channel	Customer purchase channels
Membership level	Customer membership level
Purchase frequency	Customer purchase frequency
Purchase satisfaction	Customer purchase satisfaction
Post-purchase evaluation	Customer post-purchase evaluation

2.2. Xgboost algorithm

The e-commerce customer churn prediction model actually calculates the churn probability of a customer based on the various types of customer information data input to the model to predict whether this customer will defect or not. In this paper, the Xgboost algorithm [9] is used to construct the e-commerce customer churn prediction model, and the Xgboost algorithm is an improvement of the boosting algorithm. The prediction function is:

$$\widehat{y_\imath} = \sum_{M=1}^{M} f_m(x_i), f_m \in F, \tag{1}$$

where M represents the number of regression trees, F is the set of all regression trees, f represents a function in F, and $\widehat{y_\imath}$ represents the final prediction result. The objective function is defined as:

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \ \widehat{y_\imath}) + \sum_{M=1}^{M} \Omega(f_m) \tag{2}$$

where l represents the loss function, Ω represents the regularization term, and y_i represents the true outcome of the sample. The following equation is obtained by performing Taylor expansion, making the first t-1 trees be constant, and defining the set of node split candidates of the tree and grouping of leaf nodes [10]:

$$obj^T = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T, \tag{3}$$

where g_i and h_i are the first- and second-order derivatives of loss function $\widehat{y_\imath}$. Let $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$, and the following equation is obtained:

$$obj^T = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T, \tag{4}$$

where w represents the vector composed of the scores on the leaf nodes, i.e., the weight vector. The partial derivative of w_j is calculated. The optimal solution of the objective function is obtained by substituting the w value when the partial derivative is 0 into the above formula to obtain the minimum value. The corresponding formula is:

$$obj = \frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T, \tag{5}$$

where T represents the number of leaf nodes per tree, γ represents the difficulty of node slicing, and λ represents the regularization factor.

3. Case analysis

3.1. Data selection and processing

The data of 250,000 customers of the school-enterprise cooperative e-commerce enterprises within one year were selected as the initial data set. When the customer's last purchase was made within a year was observed, and customers who have not made a purchase for more than six months after the last purchase were regarded as defected customers. It was found that the number of retained customers who had made purchases within six months was 110,267; the number of defected customers who had not made purchases within the period was 139,733. The specific number of customer samples is shown in Table 2.

Table 2

Customer labeling statistics after data pre-processing

Category	Tags	Quantity
Retained	1	110,267
Defected	0	139,733

According to the number of customer samples obtained after the above classification, it was found that there was a gap between the number of retained customers and defected customers. In order to avoid the influence of the gap between the two numbers on the prediction results of the model, the number of defected customers was set to be the same as the number of retained customers, i.e., 110,267, to form a data set consisting of 220,534 customers. The data set was randomly divided according to the proportion; 70% became the training set to train the model; 30% became the test set to test the prediction performance of the model on e-commerce customer churn.

Also, in order to meet the training requirements of the model, the data set was processed. The data operations used are as follows. The first processing was data imbalance processing [11], and this has been performed in the last step, i.e., the number of the two kinds of customers was both 110,267. The second processing was data transformation. Through the observation of the dataset, it was found that gender, post-purchase evaluation, and purchase channel were character variables, which could not be analyzed, so they were transformed. After transformation, gender was transformed to two values, i.e., 0 and 1, representing male and female, respectively; post-purchase evaluation was transformed to three values, i.e., 0, 1, and 2, representing positive, negative, and non-evaluated, respectively; purchase channel was transformed to three values, i.e., 0, 1, and 2, representing APP side, personal computer side, and other software link conversion, respectively. The third processing was data normalization. The data were normalized [12], avoiding some

odd samples to produce adverse effects on the model, and the data were limited to the range between 0 and 1.

3.2. Evaluation indicators

The prediction performance of the model was evaluated by the the most commonly used confusion matrix in the dichotomous classification problem. Precision (P), recall rate (R), and F_1 value in the confusion matrix were selected as the main evaluation indicators [13]. The calculation formulas of the indicators are:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}, \tag{6}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \tag{7}$$

$$F_1 = \frac{precision * recall\ rate}{precision + recall\ rate} * 2. \tag{8}$$

The other evaluation indicator, AUC, was the area enclosed with the coordinate axis under the receiver operator characteristic (ROC) curve, and its value ranged from 0.5 to 1. When the value of AUC was closer to 1, it indicated that the model's prediction was more effective.

3.3. Experimental design

The experimental design process was collecting the information of e-commerce customers, initially understand the characteristics of all attributes, and exploring the relationship between every attribute and customer churn, and processing the data, including data sample balance, feature selection, etc. The processed data sets were randomly divided into training and testing sets. Then, the Xgboost algorithm, Logistics regression algorithm [14], and plain Bayesian algorithm model [15] were trained and used to predict the customer churn. Finally, the experimental results obtained from different models were compared to determine whether the prediction results of the Xgboost algorithm model on e-commerce customer churn were accurate. At the end, the important factors affecting the customer churn prediction model were analyzed, and corresponding suggestions were proposed for the e-commerce customer churn problem.

3.4. Analysis of results

To ensure that the selected e-commerce customer churn characteristics were credible and valid for the prediction model, Pearson correlation coefficients [16] were calculated for the eight customer churn features. Table 3 shows that the Pearson correlation coefficients for customer gender and age ranged from 0.4 to 0.6, i.e., the correlation was moderate; the coefficients for customer purchase channel and purchase power ranged from 0.6 to 0.8, i.e., the correlation was high; the Pearson correlation coefficients for membership level, purchase frequency, purchase satisfaction, and post-purchase evaluation ranged from 0.8 to 1.0, i.e., the correlation was extremely high. The coefficients were all positive, indicating that the above eight churn features were all positively correlated with whether customers

defected or not, so the above churn features could be used in the subsequent prediction model analysis.

Table 3

Pearson correlation coefficients for different churn features

Feature name	Type	Pearson correlation coefficient
Customer gender	Text	0.4773
Customer age	Numerical	0.5925
Customer purchase channel	Text	0.6374
Customer buying power	Numerical	0.7962
Membership level	Numerical	0.8417
Customer purchase frequency	Numerical	0.8634
Customer purchase satisfaction	Numerical	0.8772
Customer post-purchase evaluation	Numerical	0.9362

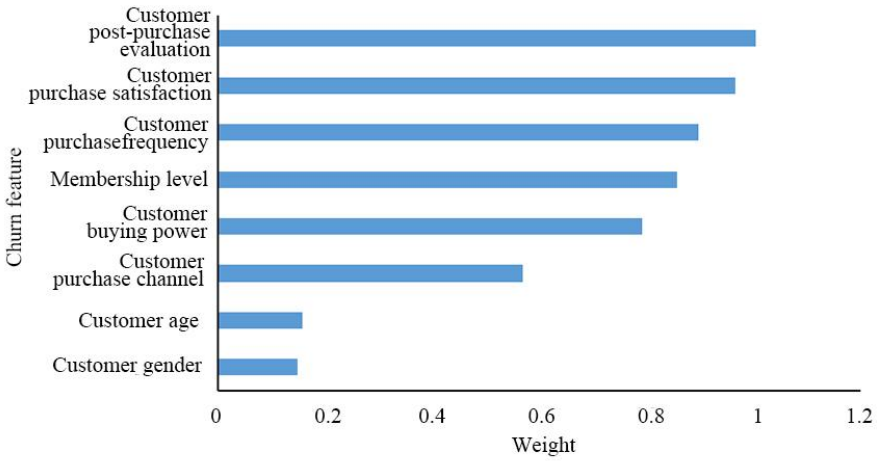

Fig. 1. Weights of different churn features

The importance of the selected features was known from the weight calculation [17]. It was seen from the weight size in Figure 1 that customer post-purchase evaluation was the most important factor affecting e-commerce customer churn, with a weight of 1.0037; second was customer purchase satisfaction, with a weight of 0.9659; third was customer purchase frequency, with a weight of 0.8975. The importance of the remaining features, in descending order, were membership level, customer purchase power, customer age, customer gender, and customer purchase channel.

It can be seen from the data in Table 4 that the accuracy of the Xgboost algorithm prediction model for different churn features was above 85%, and most

of them were above 90%, the highest of which is 97.01% for customer purchase satisfaction and the lowest is 85.56% for customer purchasing power.

Table 4

Identification accuracy of the Xgboost algorithm prediction model for different e-commerce customer churn features

Analysis dimension	Attrition characteristics	Accuracy
Customers themselves	Customer age	95.37%
	Customer gender	98.27%
	Customer buying power	85.56%
	Customer purchase frequency	90.52%
Store reason	Customer purchase channel	86.45%
	Membership level	90.68%
Shopping experience	Customer purchase satisfaction	97.01%
	Customer post-purchase evaluation	92.64%

It shows that the recognition accuracy of Xgboost algorithm prediction model for different churn features is good, which can effectively support the prediction analysis of customer churn by the model.

Table 5 presents the prediction results of three different models for e-commerce customer churn. Among them, the Xgboost algorithm model had a precision of 92.17%, a recall rate of 91.84%, a F_1 value of 92.00%, and an AUC of 0.90; the logistic regression model had a precision of 86.24%, a recall rate of 87.73%, a F_1 value of 86.59%, and an AUC of 0.85; the naive Bayesian model had a precision of 80.41%, a recall rate of 81.39%, a F_1 value of 81.04%, and an AUC of 0.79. It was seen that the Xgboost algorithm model was better than both logistic regression and naive Bayesian models in predicting e-commerce customer churn.

Table 5

Comparison of the prediction results of different models for e-commerce customer churn

Model category	Precision	Recall rate	F_1	AUC
Xgboost algorithm	92.17%	91.84%	92.00%	0.90
Logistics regression	86.24%	87.73%	86.98%	0.85
Naive Bayesian	80.41%	81.39%	80.90%	0.79

4. Discussion

The number of e-commerce customers is extremely important to the success of e-commerce enterprises, and having a large number of customers is the only way for enterprises to survive successfully in the competition. This paper used the Xgboost algorithm to build a prediction model of e-commerce customer churn. Firstly, e-commerce customer information was collected and processed; then,

Xgboost algorithm, logistic regression algorithm, and naive Bayesian algorithm models were trained by the data and used for prediction; finally, the experimental results obtained by different models were compared. The results showed that the precision, recall rate, and AUC of the Xgboost algorithm model was 92.17%, 91.84%, and 0.90, respectively, which were better than those of the other two models. Based on the above case analysis, the following suggestions are proposed for how to reduce the e-commerce customer churn situation by combining the features of the e-commerce customer churn situation and the model prediction results.

The first suggestion is to improve product links. When customers choose to buy a product, they will generally pay attention to the price, material, physical picture, and other such detailed content. Enterprises can detail the product offers in the main image of the product link and make a significant mark to inform customers of the strength of the product offers to impress them. At the same time, as the e-commerce is online shopping, customers cannot touch and observe the material and details of a product, thus the product details page in its product link is particularly important, which can display the product material, brand, details of the diagram and other information. Also, through the product link, customers can access to the buyer's evaluation page, thus the sellers need to manage the evaluation and top the good feedback containing product details to make customers produce the idea of buying this product and become retained customers.

The second suggestion is good store service and after-sales. Store service and after-sales is an important way to maintain customer relations, which is directly related to whether customers will produce long-term purchase behavior. Store service refers to customer service. When customers are interested in the product but have some confusions, they will seek customer service, so customer service staffs need to have a very good understanding of the product, answer customers according to their needs, and have a good sense of service. The excellent store after-sales will allow customers to have a good shopping experience, which can further improve customer loyalty and retention rate.

The third suggestion is to guarantee logistics services. Logistics services have now become more important. In the process of product transportation, it is likely to happen that the product packaging is broken, or damaged products cannot be used, it is necessary for e-commerce enterprises to connect the logistics enterprise for detailed inquiries and make the corresponding solutions to countermeasures to protect the rights and interests of customers.

In summary, the Xgboost algorithm used in this paper is feasible for predicting e-commerce customer churn, but there are still some shortcomings, for example, the evaluation indicator used was only the evaluation indicators of the algorithm. In future research, the prediction of e-commerce customer churn will take into account the input-return ratio of the enterprise and use profit as one of the

evaluation indicator of the model. Finally, it is hoped that the results of this paper can serve as a theoretical guide and provide some support for researchers in this field.

5. Conclusion

This paper briefly introduced the e-commerce customer churn features and the Xgboost algorithm. A prediction model of e-commerce customer churn was constructed by the Xgboost algorithm. After data collection and pre-processing, the relationship between every attribute and customer churn was explored, and the relatively important attributes were selected for model prediction. Three models, including Xgboost algorithm, logistic regression, and naive Bayesian models, were selected, and their prediction performance was compared to determine whether the Xgboost algorithm model was optimal. The results of this study demonstrated that the accuracy of the Xgboost prediction model for identifying different churn features was above 90% in most cases; the precision, recall rate, F_1, and AUC of the model was 92.17%, 91.84%, 92.00%, and 0.90, respectively, and its prediction performance was better than logistic regression and naive Bayesian models. The results suggest that it is feasible and highly accurate to build a model to predict the churn of e-commerce customers by the Xgboost algorithm.

REFERENCES

[1]. *P. Berger and M. Kompan*, "User Modeling for Churn Prediction in E-Commerce", IEEE Intell. Syst., **vol. 34**, no. 2, 2019, pp. 44-52.

[2]. *L. Yang*, "Predictions model of customer churn in E-commerce based on online sequential optimization extreme learning machine", J. Nanjing Univ. Sci. Technol., **vol. 43**, no. 1, 2019, pp. 108-114.

[3]. *B. Raja and P. Jeyakumar*, "An Effective Classifier for Predicting Churn in Telecommunication", J. Adv. Res. Dyn. Control Syst., **vol. 11**, no. 1, 2019, pp. 221-229.

[4]. *S. A. Amatare and A. K. Ojo*, "Predicting Customer Churn in Telecommunication Industry Using Convolutional Neural Network Model", IOSR J. Comput. Eng., **vol. 22**, no. 3, 2021, pp. 54-59.

[5]. *Q. Zhu, X. Yu, Y. Zhao and D. Li*, "Customer churn prediction based on LASSO and Random Forest models", IOP Conf. Ser. Mater. Sci. Eng., **vol. 631**, no. 5, 2019, pp. 1-5.

[6]. *A. Ahmed and D. Maheswari*, "An enhanced ensemble classifier for telecom churn prediction using cost based uplift modelling", Int. J. Inform. Technol., **vol. 11**, no. 2, 2019, pp. 381-391.

[7]. *K. Ebrah and S. Elnasir*, "Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms", Comput. Commun., **vol. 7**, no. 11, 2019, pp. 33-53.

[8]. *H. Mei and X. Li*, "Research on E-commerce Coupon User Behavior Prediction Technology Based on Decision Tree Algorithm", Int. Core J. Eng., **vol. 5**, no. 9, 2019, pp. 48-58.

[9]. *G. Jiao and H. Xu*, "Analysis and Comparison of Forecasting Algorithms for Telecom Customer Churn", J. Phys. Conf. Ser., **vol. 1881**, no. 3, 2021, pp. 1-6.
[10]. *A. M. Naser and E. S. Al-Shamery*, "Churners Prediction Based on Mining the Content of Social Network Taxonomy", Int. J. Recent Technol. Eng., **vol. 8**, no. Issue-2S10, 2020, pp. 341-351.
[11]. *A. Deng, H. Zhang, W. Wang, J. Zhang, D. Fan, P. Chen and B. Wang*, "Developing Computational Model to Predict Protein-Protein Interaction Sites Based on the XGBoost Algorithm", Int. J. Mol. Sci., **vol. 21**, no. 7, 2020, pp. 1-13.
[12]. *B. Dmitrii and M. Konstantin*, "Impact of Data Normalization on Classification Model Accuracy", Res. Papers Fac. Mater. Sci. Technol. Slovak Univ. Technol., **vol. 27**, no. 45, 2019, pp. 79-84.
[13]. *A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo and S. Anwar*, "Customer churn prediction in telecommunication industry using data certainty", J. Bus. Res., **vol. 94**, no. JAN., 2019, pp. 290-301.
[14]. *X. Li and Z. Li*, "A Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm", Ing. Syst. Inform., **vol. 24**, no. 5, 2019, pp. 525-530.
[15]. *Y. Yulianti and A. Saifudin*, "Sequential Feature Selection in Customer Churn Prediction Based on Naive Bayes", IOP Conf. Ser. Mater. Sci. Eng., **vol. 879**, 2020, pp. 1-7.
[16]. *P. Chen, F. Li and C. Wu*, "Research on Intrusion Detection Method Based on Pearson Correlation Coefficient Feature Selection Algorithm", J. Phys. Conf. Ser., **vol. 1757**, no. 1, 2021, pp. 1-10.
[17]. *B. E. King and J. Rice*, "Analysis of Churn in Mobile Telecommunications: Predicting the Timing of Customer Churn", AIMS Int. J. Manag., **vol. 13**, no. 2, 2019, pp. 127.