Construction and Application of Cross-border E-commerce Customer Characteristic Analysis Model Based on Data Mining Technology

Yijia Yu*, Qingxiu Liu

*Corresponding author: workforyu@126.com, 2575413727@qq.com

Chongqing College of Architecture and Technology, Chongqing, 401331, China

Abstract. With the continuous development of digital information technology, the application of data mining technology in precision marketing of cross-border e-commerce platform has become more and more extensive, and it has become a key force to promote the reform of marketing model of cross-border e-commerce platform. However, the current direction of data mining mostly focuses on personalized recommendation of goods, and its application scope is single. In this regard, based on the actual needs of the current cross-border e-commerce platform, combined with data mining technology, this paper will build a set of customer feature analysis model to help cross-border e-commerce platform quickly acquire the target user groups and complete the user purchase forecast, so as to make the platform's operation plan and marketing measures more accurate and effective. Practice has proved that the customer feature analysis model integrates Logistic Regression, XGBoost, CatBoost and other algorithms, which can predict and analyze the user information, user attributes, user behavior and other feature information, and reflect the user's willingness to buy a certain kind of goods. At the same time, the model also supports K-means algorithm to cluster customer characteristics, so as to clarify the target user group of a certain kind of goods. The application of customer characteristic analysis model not only provides a complete data mining analysis process, but also provides necessary technical support for the implementation and management of precision marketing strategy of cross-border e-commerce platform, which has certain promotion significance.

Keywords-data mining technology; Cross-border e-commerce; Precision marketing; Customer characteristic analysis model

1 Introduction

With the rapid development of the global digital economy, as a brand-new innovative economic format, cross-border e-commerce has greatly promoted the digital transformation and upgrading of the international trade industry, and is also the key driving force for high-quality social and economic development in the new era. [1] Compared with traditional export trade, cross-border export e-commerce has outstanding advantages in convenience, diversification and mobility, but it also faces changeable environmental impact and severe market competition. In this context, more and more cross-border e-commerce platforms have begun to focus on the reform of marketing model, in an attempt to improve the accuracy and effectiveness of marketing strategies through digital marketing means.

BDEDM 2024, January 12-14, Ningbo, People's Republic of China

DOI 10.4108/eai.12-1-2024.2347212

As a new marketing model, digital precision marketing can improve the marketing effect and customer satisfaction of cross-border e-commerce platform with the help of the practical advantages of big data and data mining technology. However, the current mining analysis focuses on user traffic, user behavior, user value and personalized recommendation of goods, and lacks the analysis of purchasing tendency of a certain kind or specific goods and the method of dividing the corresponding customer groups. [2] At the same time, combined with the current research at home and abroad, it is found that the user behavior mining analysis proposed in references [3] and [4] is based on historical data, and the prediction results are relatively scattered, which is easily affected by the cold start problem. In the aspect of user characteristics, the research object used in literature [5] is relatively simple, and the correlation analysis with user behavior is completed only from the basic information of users, which has the problem of insufficient feature dimension. In view of this, this paper believes that based on the actual needs of the current cross-border e-commerce platform, a set of customer feature analysis model is built with data mining technology, and the user information, user attributes, user behavior and other feature information are emphatically integrated. It improves the practical performance of the model, expands the application scope of the analysis model, opens up new methods for helping cross-border e-commerce platforms to quickly acquire target user groups and complete user purchase forecasts, and makes the platform's operation scheme and marketing measures more accurate and effective.

2 Model construction

2.1 Prediction algorithm

Under the predictive analysis mining method, there are three main algorithms: decision tree, naive Bayes and artificial neural network. Combined with the actual application requirements and implementation difficulties in this study, three algorithms, namely Logistic Regression, XGBoost and CatBoost, are finally proposed as the core applications of the cross-border e-commerce customer feature analysis model.

2.1.1 Logistic Regression

Logistic Regression is a statistical learning method for solving classification problems. It maps input variables to output variables by fitting a logical function, thus modeling the relationship between input variables and output variables as probability. The final output result of this logic function is between [0,1], and then it is judged by combining with the established threshold: when the output result is greater than the threshold, it is judged as a positive sample, and when it is less than the threshold, it is judged as a negative sample. In the specific algorithm definition, the set of input variables is X={x1, x2, x3, …, xn}, θ={θ1, θ2, θ3, …, θn} represents the corresponding coefficients of the input variables, H(z) is the final output value, and g(x) is the sigmoid function, as shown in Formula 1.

$$H(z) = g(x) = \frac{1}{1+e^{-z}} \quad z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \tag{1}$$

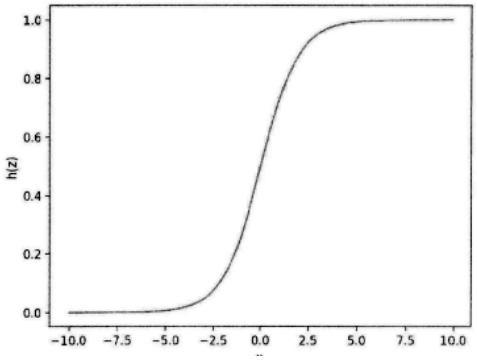

Figure 1. Sigmoid function curve.

As shown in Figure 1, the sigmoid function as a whole is monotonically increasing. When the value is infinitely close to positive infinity, the corresponding probability value is closer to 1. Similarly, when the value of infinity is close to negative infinity, the corresponding probability value is closer to 0. Therefore, the decision threshold of sigmoid function is generally set to 0.5. From the above analysis of Logistic Regression algorithm, we can see that the overall construction of the algorithm is relatively simple and has good explanatory power.

2.1.2 XGBoost

XGboost algorithm is a representative gradient lifting algorithm, and its overall structure is similar to decision tree. XGboost adopts multi-round iterative calculation mode, and each iteration will generate a new weak classifier, which will be trained based on the residual of the previous classifier to continuously reduce the classification deviation and improve the accuracy of the final classifier. [6] In the specific algorithm definition, the preset XGboost objective function formula is shown in Formula 2.

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \Omega(f_t) \tag{2}$$

Where n is the number of samples, y_i is the actual value of sample i, and $\hat{y}_i^{(t)}$ is the predicted value of sample i by t decision trees, and $\Omega(f_t)$ represents the complexity of the t decision tree, which belongs to the regularization term. According to the principle of XGboost algorithm, the number of decision trees increases from 0 until the t decision tree appears. When the t decision tree predicts the sample i, it is necessary to keep the prediction result of the previous t-1 decision tree and add the prediction value of the t decision tree, that is, the generalized derivation formula is shown in Formula 3. When the maximum depth and sample weight of the decision tree are less than the set threshold, stop the cyclic iteration of the tree and get the best tree structure. [7]

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{3}$$

Compared with other prediction algorithms, XGBoost has higher prediction accuracy, and can weaken the influence of each decision tree by multi-round iterative calculation mode, thus reducing the variance of the model and avoiding the risk of over-fitting.

2.1.3 CatBoost

Both CatBoost algorithm and XGboost algorithm belong to gradient lifting algorithm, and the calculation mode is also multi-round iterative calculation. Compared with XGboost algorithm, CatBoost algorithm has an adaptive learning rate, which can better control the contribution of weak classifiers in each iteration, thus improving the prediction accuracy. The calculation formula of adaptive learning rate is shown in Formula 4, where t is the number of iterations, ηt is the learning rate of the t iteration, and αt is the average learning rate of the previous t iterations.

$$\eta_t = \frac{1}{\sqrt{t+1}} \quad \alpha_t = \frac{\sum_{i=1}^{t} \eta_i}{t} \tag{4}$$

In addition, the objective function of CatBoost algorithm is also adjusted, and the square loss function is adopted, as shown in Formula 5.

$$L(\theta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

CatBoost algorithm can adapt to large-scale data set applications, and the training speed is fast. It has certain robustness in the face of irregular and missing data samples, which ensures the accuracy of the prediction results.

2.2 Clustering algorithm

K-means algorithm is a clustering algorithm based on partition. It takes k as the parameter and divides n data objects into k clusters, so that the similarity within clusters is high, while the similarity between clusters is low. K-means algorithm uses distance as the similarity evaluation index between data objects, that is, the closer the distance between two target objects, the higher the similarity, and vice versa. In the specific algorithm definition, the unlabeled input sample set $S=\{x_1, x_2, x_3, \ldots, x_m\}$ and the initial center set of K categories $u=\{u_1, u_2, u_3, \ldots, u_k\}\}$, and some samples are marked as the classes closest to the class center, as shown in Formula 6.

$$label_i = \arg\min_{1<j<k} \left\| x_i - u_j \right\| \tag{6}$$

Then, calculate the sample average of the category to update the center positions of K categories, as shown in Formula 7.

$$u_j = \frac{1}{\left| C_j \right|} \sum_{i \in C_j} x_i \tag{7}$$

Finally, when the class center no longer changes, stop and output the clustering results, and then sort out the information we need, such as the class to which each sample belongs, for subsequent statistics and analysis. [8] From the definition process of K-means algorithm, it is clear that the algorithm itself also adopts iterative calculation mode, and each iteration will change the clustering situation of samples. Therefore, when K-means algorithm is used for clustering, it is necessary to preprocess the sample data in advance to reduce the influence of bad data on the final result.

3 Practical application

3.1 Data set processing

The data used in this study is derived from some data of an enterprise on overseas eBay platform. The main category is 3C electronic products. The data was collected in September 2022, and the data has been desensitized, which is limited to research use. There are about 1.3 million original data, including basic user information, user attributes, user's recent behavior and user's operation data on 3C electronic products. Table 1 shows the field attributes and description information in the original data. After preprocessing operations such as deleting missing values, duplicate records and abnormal values, the original data finally obtained 93,426 valid data.

Table 1. Attribute and description information of original data fields.

Category	Field name	Field property description	Data type
User basic information	Gender	Male / female / unknown	Category
	Age	18-25, 26-35, 36-45, 46-55, more than 55	Category
	Location	Five-star/four-star/three-star/two-star/one-star city	Category
User attributes	Member	Yes / no	Category
	Membership grade	Level I / Level 2 / Level 3 / Level 4 / Level 5	Category
	Promotion sensitivity	High sensitivity/medium sensitivity/average sensitivity/low sensitivity/insensitivity	Category
User's recent behavior	Number of orders	The total number of orders generated in the past 2 months	Numerical value
	Page views	The total number of commodities visited in the past 2 months	Numerical value
	Amount of consumption	Total consumption in the past 2 months	Numerical value
	Login frequency	Login times in the past 2 months	Numerical value
Page operation	Operational behavior	Purchase/ join the shopping cart/ collect /consult/ enter the store/ browse the goods	Category

3.2 Customer purchase forecast

After one-hot coding and normalization, the original data will be used to predict the customer purchase of 3C electronic products. According to the user's page operation behavior, users are divided into customers who want to buy and customers who don't want to buy, which are used as the output variables of the prediction model, while the user's basic information, user attributes and user's recent behavior are used as the input variables to input the Logistic Regression, XGBoost, CatBoost and other algorithms respectively, so as to verify the final prediction results and clarify the importance of input features. [9] In this study, the precision (P), recall (R) and F1 value are selected to measure the performance of the algorithm. Each index is calculated based on the confusion matrix (Table 2), as shown in Formula 8.

Table 2. Confusion matrix.

	Forecast is positive	Forecast is negative
Truth is positive	TP	FN
Truth is negative	FP	TN

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad F_1 = \frac{2 \times P \times R}{P + R} \tag{8}$$

The final predicted performance pairs are shown in Table 3. The results show that all three algorithms in the customer feature analysis model can realize customer purchase prediction; Based on the value of F_1, CatBoost algorithm performs best.

Table 3. Comparison of the prediction results of each model.

Single algorithm	F_1 value	Precision（P）	Recall（R）
Logistic Regression	0.4366	0.6445	0.3302
XGBoost	0.4798	0.5231	0.4431
CatBoost	0.4815	0.5868	0.4082

In order to further clarify the influencing factors of customers purchasing 3C electronic products, we will continue to output the importance ranking of each input feature based on CatBoost algorithm, and the output results are shown in Table 4. The results show that among the important characteristics, membership grade, gender, age and location have obvious influence on users' purchasing desire, while users' recent behavior has a low influence.

Table 4. Importance of characteristics output by CatBoost algorithm.

Field attribute	Category	Ranking
Is a member	User attributes	1
Not a member		2
Gender male	User basic information	3
Gender female		4
26-35 years old	User basic information	5
36-45 years old		6
Five-star city	User basic information	7
Three-star city		8
...	...	...

3.3 Target groups cluster analysis

According to the ranking of the importance of user features output by CatBoost algorithm, the customer feature analysis model will continue to cluster customer features with K-means algorithm, so as to clarify the target user groups of a certain kind of goods and facilitate the subsequent cross-border e-commerce platform to specify more accurate and effective operation schemes and marketing measures.

The three core elements of K-means algorithm are cluster number K, center point and distance, among which the determination of K value will affect the performance of K-means algorithm, and the contour coefficient method is usually used to select the best K value. The formula for calculating the profile coefficient is shown in Formula 9, where s(i) is the profile coefficient of a single sample, s is the average value of the profile coefficients of all samples, a(i) is the average distance between a single sample and other samples, and b(i) is the minimum average distance between a single sample and other cluster samples. [10] After calculation, when K=4, the SSE decline amplitude of K-means algorithm decreases rapidly and tends to be flat, so K=4 is the best value.

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad S=\frac{1}{n}\sum_{i=1}^{n}s(i) \tag{9}$$

After that, the membership grade, gender, age, location and other characteristics are input into the K-means algorithm for cluster analysis, and the obtained results are shown in Table 5. The results show that users who want to buy 3C electronic products are divided into four categories: male members aged 26-35 who live in five-star cities; male non-members aged 36-45 who live in four-star cities; female members aged 18-25 who live in three-star cities; female members aged 26-35 who live in five-star cities. Based on this, it is convenient for the cross-border e-commerce platform to quickly identify the target user groups, and take corresponding operation plans and marketing measures to increase the sales of the platform.

Table 5. User characteristic cluster analysis results.

Category	Member	Gender	Age	Position	Quantity proportion
1	Yes	Male	26-35	Five-star	55.13%
2	No	Male	36-45	Four-star	22.63%
3	Yes	Female	18-25	Three-star	11.57%
4	Yes	Female	26-35	Five-star	10.67%

4 Conclusions

In order to improve the application efficiency of data mining technology in cross-border e-commerce platform, this paper proposes a set of customer feature analysis model based on the problems of single data mining direction and insufficient feature dimension, which helps cross-border e-commerce platform to quickly acquire target user groups and complete user purchase prediction, so as to make the platform's operation scheme and marketing measures more accurate and effective. Practice has proved that the customer feature analysis model can predict the purchase intention of users by using CatBoost algorithm, and the overall performance is the best. By using K-means algorithm, users with purchasing intention can be divided into four categories, and the target group can be effectively defined. In the follow-up research, it is necessary to further optimize the construction scheme of each algorithm, and gradually promote the application of deep neural network in this field, so as to try to implement the precise marketing strategy and intelligent management of cross-border e-commerce platforms.

References

[1] Geng Jingjing, Zhu Yanfang. Research on the Development of Cross-border E-commerce under the Background of Big Data[J]. Economic & Trade Update.2023.04.79-81

[2] Zhang Hao. Research on the Optimization of Precision Marketing Strategy of H Company under the Background of Big Data[D]. Xi'an University of Technology.2023.06

[3] M. Nasir C. I. Ezeife. A Survey and Taxonomy of Sequential Recommender Systems for E-commerce Product Recommendation[J]SN Computer Science.2023.09.6

[4] Tingzhong Wang, Nanjie Li et al. Visual Analysis of E-Commerce User Behavior Based on Log Mining[J].Advances in Multimedia.2022.05.1-22

[5] Muniappan Ramaraj Jothish Chembath. A new fuzzy rule-based optimization approach for predicting the user behaviour classification in M-commerce[J].IJRES.2023.11.320

[6] Li Zhanshan, Liu Zhaogeng. Feature Selection Algorithm Based on XGBoost[J]. Journal on Communications.2019.09.101-108

[7] Lu Wanwan et al. Research on Early Warning Model Based on XGBoost Algorithm[J]. Electronic Design Engineering.2020.10.49-54

[8] Guo Yongkun et al. K-means Clustering Algorithm for Optimizing the Initial Clustering Center[J]. Computer Engineering and Applications.2022.04.172-178

[9] Jing Xiuli, Shi Mingxi. Prediction of Repeat Purchase Behavior of E-commerce Users Based on XGBoost Algorithm[J]. Journal of Liaoning University,2023.05.134-145

[10] Li Qiuyun, Liu Yanwu. An Initial Center Selection Method for K-means[J]. Application of Electronic Technique.2023.03.134-138