Research on Customer Value of E-commerce Data Based on Machine Learning

Jiazhen Zou

Department of Economics and Management, Northeast Normal University, Changchun, Jilin, China

*Corresponding author: zoujz347@nenu.edu.cn

Abstract. Nowadays, online shopping has popularized all over the world, the e-commerce platforms have to make every effort to compete with each other. In order to survive from the fierce competition, the e-commerce platforms have to make more profits with less cost which means that high-quality customers should be the most important resource for the e-commerce platforms. Therefore, the purpose of this passage is to measure the customer value by analyzing the e-commerce data with the methods of K-Means, RFM model and logistic regression. In the experiment, the RFM model shows that the customers in different clusters have different value for the e-commerce platform, the prediction model formed by logistic regression, visualized by confusion matrix and evaluated by 4 indicators shows that only 50.4% of the customers will repurchase on this e-commerce platform in the future. So, it is necessary for the e-commerce platform to carry out the personalized strategies for the customers in different clusters.

Keywords: K-Means; RFM model; logistic regression; customer value; e-commerce platform.

1. Introduction

With the popularization of the internet, online shopping has become an essential aspect of our daily life, the competition among e-commerce platforms have become more and more fierce because the competition for the customers to ensure the revenue. As is known to all, 20% of the customers create 80% of the profits [1]. The e-commerce platforms have to raise efficiency by screening the high-quality customers and making efforts to compete for them in order to reduce the cost while maintaining their profitability. Therefore, it is necessary for the platforms to put emphasis on how to mine the customer data in order to evaluate and classify the customers, predict the customer behavior, establish effective marketing strategies.

With the development of machine learning, machine learning can analyze massive data and dig out potential information, which is suitable for business analysis, so there are many successful applications in various industries. At present, the existing models mainly include PCA followed by the application of the k-mode clustering algorithm [2], RFM analysis with traditional K-means and Fuzzy C-Means algorithms and a new algorithm RM K-Means [3], artificial neural networks and CART and random forest based on different metrics [4], the shopping behavior analysis and prediction model built by XGBoost hybrid model [5].

Hence, this passage will analyze the data of 5000 customers of an e-commerce platform by using K-Means to classify the customers, RFM model to measure customer value, logistic regression to build a prediction model about whether the customer will repurchase on this e-commerce platform in the future, visualize the prediction model with confusion matrix and evaluate it with 4 indicators, so that the e-commerce platforms can carry out the personalized strategies to maximize the profits.

2. Methodology

2.1 K-Means Clustering

In 1967, James MacQueen first proposed the clustering method of K-Means [6]. K-Means is a well-known clustering algorithm based on unsupervised partition, which divides the input data points into L partitions according to the distance metric [7]. Its main function is to cluster unlabeled samples

into several specified classifications. The ultimate purpose of K-Means is to minimize the squared error E:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|_2^2 \quad (1)$$

The steps of K-Means are shown in Fig. 1:

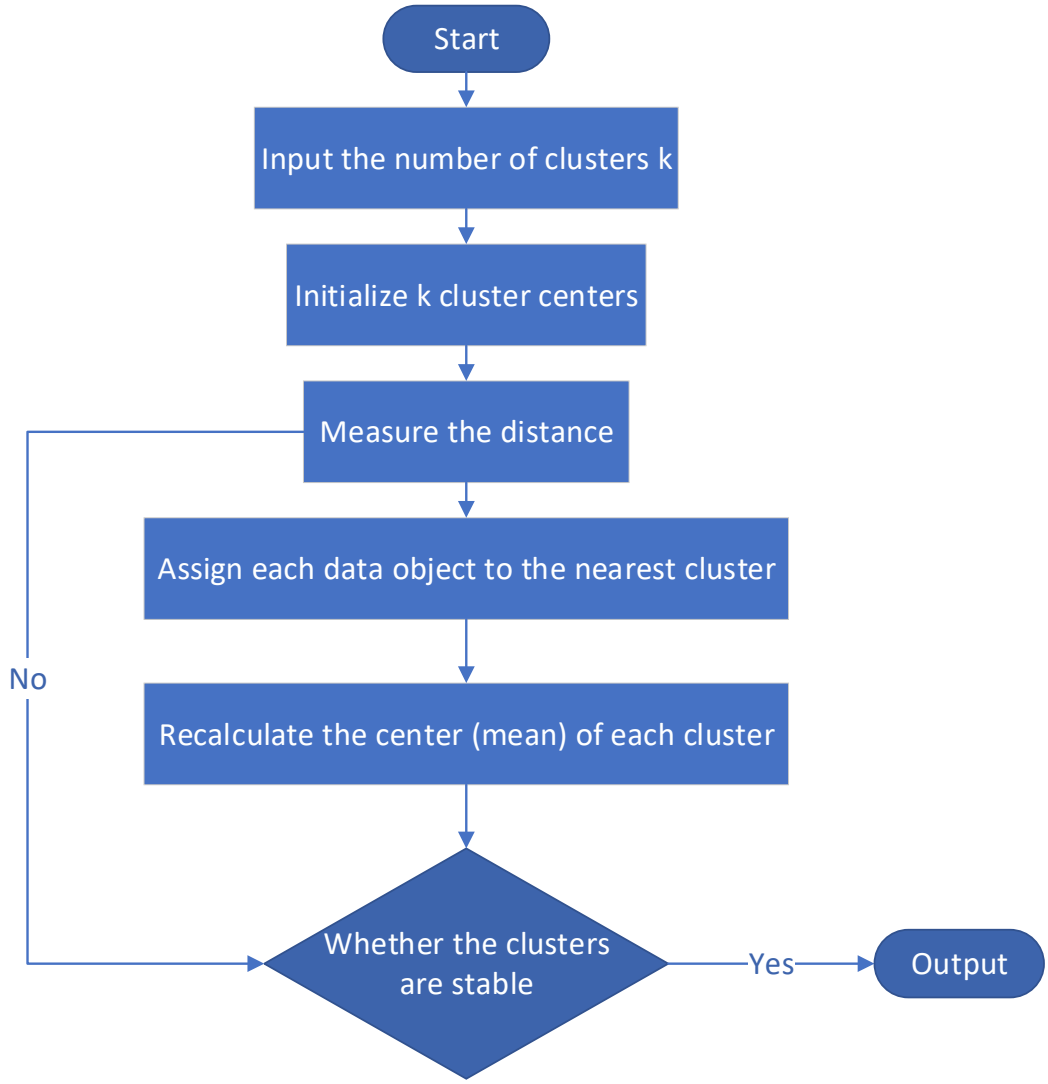

Fig. 1 Steps of K-Means

2.2 RFM Model

RFM model was first proposed by Jim Sellers and Arthur Hughes in 1994 [8]. RFM model is a method to measure the potential value of the customers and the ability to create profits, which contains three dimensions: R (Recency), F (Frequency) and M (Monetary). It is the most frequently adopted segmentation technique based on their prior purchasing history [9]. Recency measures how recently a customer has made a purchase, the smaller the R values, the more likely that the customer will make a new order. Frequency measures how often a customer makes a purchase, the greater the F values, the more likely that the customer will make more new orders in the future. Monetary measures how much money a customer spends on purchases, the greater the M values, the more value the customer creates for the e-commerce platform.

2.3 Logistic Regression

In 1973, McFadden provided a theoretical foundation of the logit model [10]. Logistic regression is a kind of generalized linear regression analysis method, which belongs to supervised learning. The central mathematical concept of logistic regression is the logit—the natural logarithm of odds ratios [11]. It is mainly used to solve binary classification problems by using training set to train the model and classifying test data after training. It can be denoted by:

$$y = \frac{1}{1+e^{-\sum b_i x_i}} \tag{2}$$

2.4 Application in Customer Value of E-commerce Data

The application in customer value of e-commerce data is shown in Fig. 2:

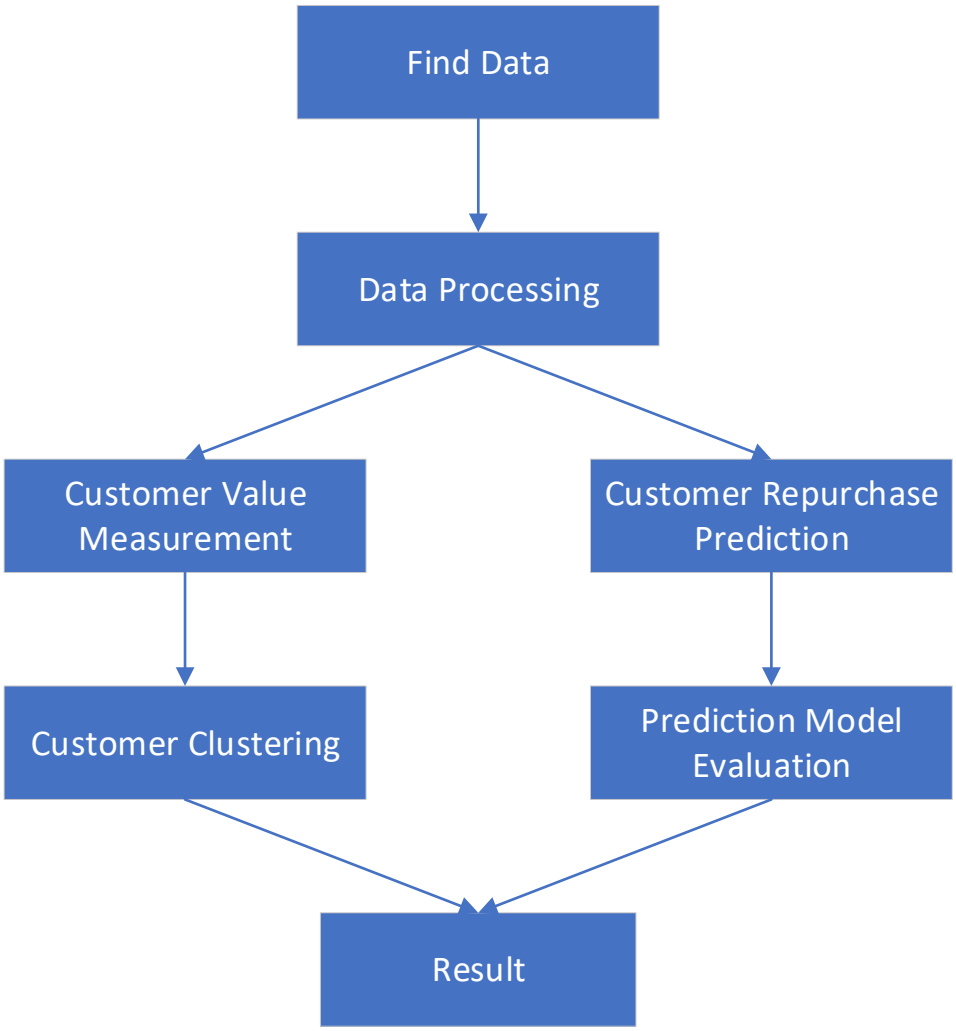

Fig. 2 Research Framework

3. Experiment

3.1 Data Description

The dataset in this research comes from Kaggle. It contains 5000 records and 40 fields, which is about the customer data of an e-commerce platform.

This dataset split the total orders and total consumption of each customer in three dimensions. The first dimension is to divide each month into 4 weeks (the days from 1st to 7th as week1, the days from 8th to 15th as week2, the days from 16th to 23th as week3, the days from 24th to 31th as week4). The second dimension is to divide each week into 7 days (from Monday to Sunday). The third dimension is to divide each day into 4 periods (from 0 to 6, from 6 to 12, from 12 to 18, from 18 to 24).

The variables in the dataset are shown in Table 1:

Table 1. Variable Information of The Dataset

Variables	Description
CustomerID	Each customer has an CustomerID
TOTAL_ORDERS	The total orders
REVENUE	The total consumption
AVERAGE_ORDER_VALUE	The average value of all orders
CARRIAGE_REVENUE	The total carriage consumption
AVERAGESHIPPING	The average carriage consumption
FIRST_ORDER_DATE	The date of first order
LATEST_ORDER_DATE	The date of last order
AVGDAYSBETWEENORDERS	The average days between orders
DAYSSINCELASTORDER	The days since last order
…	…
MONDAY_ORDERS	The total orders on Monday
…	…
MONDAY_REVENUE	The total consumption on Monday
WEEK1_DAY01_DAY07_ORDERS	The total orders in the first week of each month
…	…
WEEK1_DAY01_DAY07_REVENUE	The total revenue in the first week of each month
…	…
TIME_0000_0600_ORDERS	The total orders from 0:00 to 6:00 of each day
…	…
TIME_0000_0600_REVENUE	The total revenue from 0:00 to 6:00 of each day
…	…

3.2 Experiment Design

In order to group the customers with different value, use RFM model to measure the value of each customer and use K-Means to divide customers into different clusters according to the R, F, M value. Therefore, personalized strategies can be made for the customers in different clusters.

On the other hand, use logistic regression to build a prediction model about whether the customer will repurchase on this e-commerce platform in the future. In this part, the first step is to create a categorical variable called NextPurchaseDayRange. When the unstandardized F value is less than or equal to 90, the NextPurchaseDayRange is assigned to the value of 1, otherwise the NextPurchaseDayRange is assigned to the value of 0. The second step is to split 20% of the dataset as test set randomly, others as training set. Then, use the logistic regression to train the prediction model with the training set and test the prediction model with the test set.

Finally, generate the confusion matrix to visualize the prediction results and evaluate the prediction model by 4 indicators (accuracy, precision, recall, specificity).

In Table 2, it shows the confusion matrix of size 2 × 2:

Table 2. Confusion Matrix (n = 2)

		Predicted Class	
		Negative	Positive
True Class	Negative	TN	FP
	Positive	FN	TP

Then, use 4 indicators to evaluate the prediction model:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

$$Specificity = \frac{TN}{TN+FP} \tag{6}$$

3.3 Experiment Result

3.3.1 RFM Model

In order to have a better grouping result, standardize the RFM value by removing the mean and scaling to unit variance before using K-Means clustering for k from 2 to 10. Then, select the integer k as 4 according to the silhouette score and the balance among clusters.

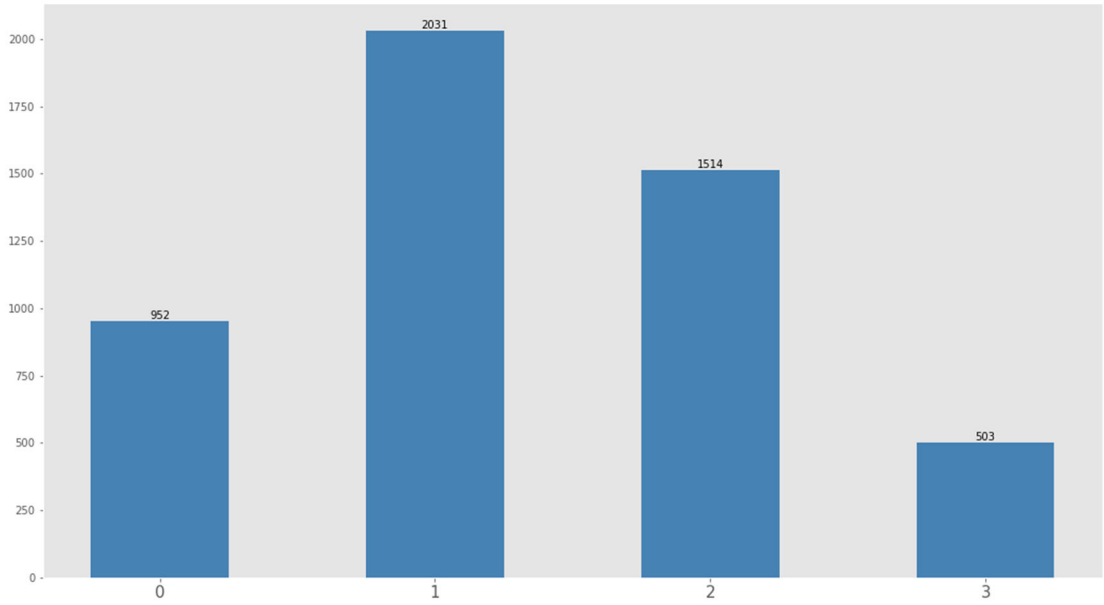

Fig. 3a Customer Clusters Distribution

It can be seen in the Fig. 3a that cluster0 contains 952 customers which is about 19% of the customers; cluster1 contains 2031 customers which is about 41% of the customers; cluster2 contains 1514 customers which is about 30% of the customers; cluster3 contains 503 customers which is about 10% of the customers.

Then, visualize the RFM value of each cluster with barplots:

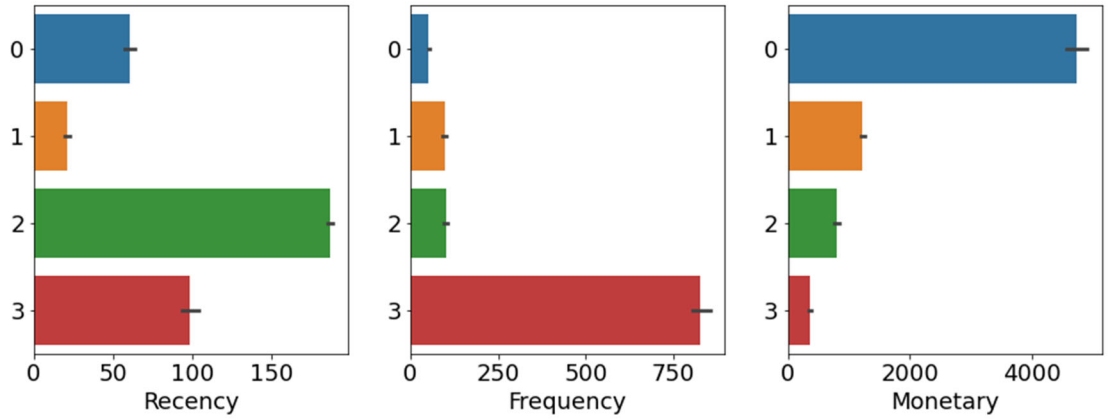

Fig. 3b RFM in Each Cluster

It can be seen in the Fig. 3b that cluster0 has the moderate R value, the lowest F value and highest M value, which means that the customers in this cluster seldom make orders on this e-commerce platform, but these customers tend to buy expensive products and occupy the most revenue of the e-commerce platform; cluster1 has the lowest R value, the F value only higher than cluster0 and the M value only lower than cluster0, which means that the customers in this cluster make the order recently; cluster2 has the highest R value, the F value a bit higher than cluster1 and the moderate M value, which means that the customers in this cluster have not made orders for a long time; cluster3 has the R value only lower than cluster2, the highest F value and the lowest M value, which means that orders made by the customers in this cluster occupy the majority of e-commerce business, but have the lowest value to the e-commerce platform.

3.3.2 Prediction Model

The prediction results are visualized in Fig. 4:

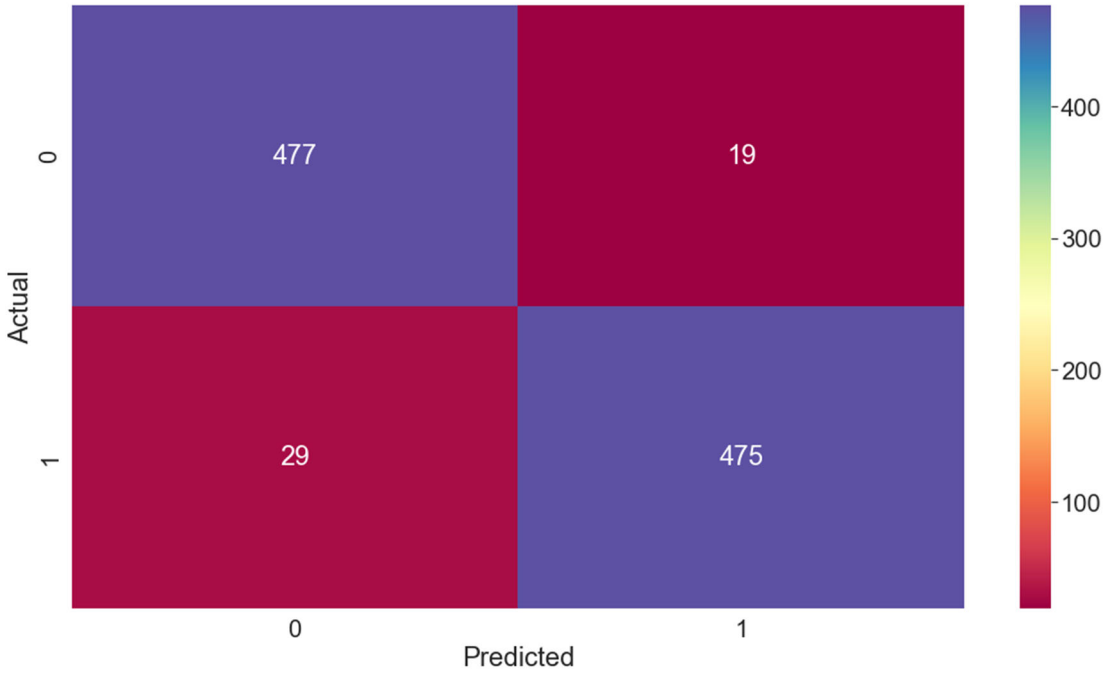

Fig. 4 Confusion Matrix

In this prediction model, the accuracy is 95.2%, the precision is 96.15%, the recall is 94.25%, the specificity is 96.17%. So, this prediction model runs well.

4. Conclusions

In this passage, RFM model is used to visualize the customer value and standardized the RFM value in order to have a better clustering result by using K-Means clustering, the clusters show huge differences in the dimensions of recency, frequency and monetary. In the other hand, logistic regression is used to build a prediction model about whether the customer will repurchase on this e-commerce platform in the future by splitting 80% of the dataset for training and other 20% for testing. The prediction results are visualized by confusion matrix and the 4 indicators which is used to evaluate the prediction model show the prediction model has high accuracy.

In terms of the prediction model, it shows that only 50.4% of the customers will repurchase on this e-commerce platform in the future in the test set, which is too low for the e-commerce platform if the platform wants to stand out in the fierce competition. Therefore, the platform should carry out the personalized strategy to retain customers and cultivate the customer loyalty.

For the customers in the cluster0, these customers are the highest-quality customers. According to the Pareto principle, these customers are the 20% of customers who can create 80% of the profits [1]. However, the frequency of these customers making orders on this e-commerce platform is too low. The platform should ensure the interest of the customers including the exclusive customer service, paying more attention to the customer feedback, and establishing a credit system aimed at motivating customers to participate. Additionally, the platform should also offer large coupons and mine the product preference of the customers.

For the customers in the cluster1 and cluster2, both of them have the similar F value and M value, the only difference is the customers in the cluster2 have a higher R value. In view of the fact that these 2 clusters own 71% of the customers, it is necessary for the platform to maintain this source of income. The platform should ensure the quality and service of the corresponding products. Besides, moderate coupons can stimulate the customers to repurchase on this e-commerce platform and cultivate the customer loyalty.

For the customers in the cluster3, although these customers have the highest F value, it is not worthy for the e-commerce platform to make too much efforts because of the M value. What the platform should do is to ensure the customer evaluation.

All in all, the methods and the conclusions in this passage can be applied to the competition for high-quality customers in not only e-commerce platforms, but also every industry.

References

[1] Pareto V. Cours d'économie politique: professé à l'Universjté de Lausanne[M]. F. Rouge, 1896.

[2] Kamthania D, Pawa A, Madhavan S S. Market segmentation analysis and visualization using K-mode clustering algorithm for E-commerce business[J]. Journal of computing and information technology, 2018, 26(1): 57-68.

[3] Christy A J, Umamakeswari A, Priyatharsini L, et al. RFM ranking–An effective approach to customer segmentation[J]. Journal of King Saud University-Computer and Information Sciences, 2021, 33(10): 1251-1257.

[4] Ekelik H, Şenol E. A Comparison of Machine Learning Classifiers for Evaluation of Remarketing Audiences in E-Commerce[J]. Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi, 2021, 16(2): 341-359.

[5] Liu C J, Huang T S, Ho P T, et al. Machine learning-based e-commerce platform repurchase customer prediction model[J]. Plos one, 2020, 15(12): e0243105.

[6] MacQueen J. Classification and analysis of multivariate observations[C]//5th Berkeley Symp. Math. Statist. Probability. 1967: 281-297.

[7] Goicovich I, Olivares P, Román C, et al. Fiber Clustering Acceleration With a Modified Kmeans++ Algorithm Using Data Parallelism[J]. Frontiers in Neuroinformatics, 2021: 46.

[8] Sellers J, Hughes A. RFM Analysis: A New Approach to a Proven Technique[J]. www. relation-shipmktg. com/Free Articles/rmr017. pdf, 1994.

[9] Jo-Ting W, Shih-Yen L, Hsin-Hung W. A review of the application of RFM model[J]. African Journal of Business Management, 2010, 4(19): 4199-4206.

[10] McFadden D. Conditional logit analysis of qualitative choice behavior[J]. 1973.

[11] Peng C Y J, Lee K L, Ingersoll G M. An introduction to logistic regression analysis and reporting[J]. The journal of educational research, 2002, 96(1): 3-14.