Research on Classification of Railway Freight Customers Based on Customer Value Analysis

Yang Yi[1(✉)] and Huang Lei[2]

[1] School of Economics and Management, Beijing Jiaotong University, Beijing, China
[2] School of Economics and Management, Beijing Jiaotong University, Beijing, China

EMAIL: y18513546911@163.com (Yang Yi);
lhuang@bjtu.edu.cn (Huang Lei)

Abstract. AbstractCustomer segmentation can divide the existing customers of an enterprise into different customer groups according to certain standards, and the research on customer segmentation based on customer value has important theoretical and practical significance. Based on the huge customer data accumulated by freight railways, a customer value segmentation model based on improved K-Means algorithm is proposed. Firstly, the dimensions and indicators of customer segmentation are determined, and then the K-Means algorithm is designed and improved based on the customer segmentation framework. Experiments show that the model has good classification effect, and it has obvious practical significance to apply it to railway freight marketing.

Keywords :ailway freight transportation, customer segmentation, K-Means Algorithm

1 Introduction

There are many types of railway freight customers with complex needs, and there are uneven distribution and waste of freight resources, which is at a disadvantage in the increasingly mature and competitive freight market and the market share. The customer is an important resource of the enterprise. Therefore, if an enterprise wants to obtain sufficient benefits, it must conduct effective difference analysis on customers based on factors such as customer attributes, behaviors, needs, preferences, and values, and provide targeted services and marketing models according to differences, so as to realize the maximization of customer resource value and enterprise profit [1].

The current railway freight system has accumulated a large amount of customer data, but there is no clear data classification method. Establish a three-dimensional segmentation model, and propose a railway customer segmentation method based on the customer's current value, potential value and product matching degree. The improved K-Means algorithm is used to segment customers, and the transportation resources are allocated according to the level of freight customers, which provides a theoretical method for realizing the rational allocation of enterprise resources [2].

Y. Jiao et al. (eds.), *Proceedings of the 3rd International Conference on Internet Finance and Digital Economy (ICIFDE 2023)*, Atlantis Highlights in Economics, Business and Management 1,
https://doi.org/10.2991/978-94-6463-270-5_3

2 Customer value segmentation model based on improved k-means algorithm

2.1 Customer value analysis of railway freight transportation

At present, the widely used customer value classification model has developed from a single-dimensional model based on customer lifetime value to a two-dimensional model based on current value and potential value to calculate customer life cycle value [3]. For the measurement of current value, the net present value of customers' past profit contribution is often used as the evaluation index; The selection of potential value measurement indicators is different, and two measurement variables are often used, on the one hand, the present value of future expected profits, and on the other hand, the ripple effect brought by customer loyalty, such as word-of-mouth publicity, repeated purchases, and expenditure ratio [4]. Considering the characteristics of rapid changes in the transportation market, fierce competition, and large choice of transportation modes for customers, the customer segmentation model should also consider the customer's product matching degree, and establish the railway customer segmentation index system by using the three-dimensional variables of current value, potential value and product matching degree [5]. The railway customer segmentation index system is shown in Figure 1:

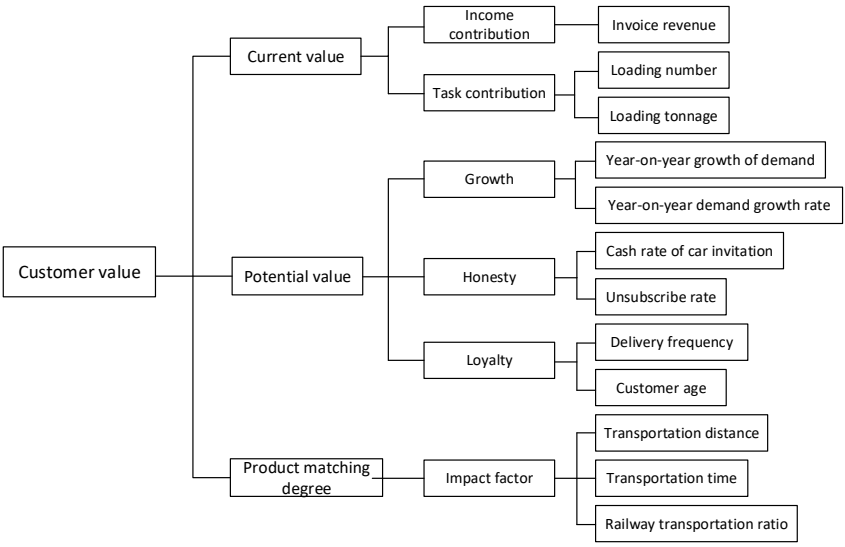

Fig. 1. Customer value evaluation index system

2.1.1 Customer current value analysis.

The current value of railway customers refers to the contribution of customers' historical transportation behavior to railway revenue, which can be measured by two indicators: income contribution and task contribution.

- Income contribution. Combined with the railway operation, it mainly comes from the railway transportation cost in the basic information table and billing information table of the freight ticket system. Transportation expenses here refer to freight in a broad sense, including freight and miscellaneous fees, quoted amount, insurance amount, transportation expenses, etc., which can truly reflect the freight income brought by customers. The Invoive revenue is used as the reference factor of income contribution index.
- Task contribution. It mainly includes the variable cost caused by customers' transportation behavior. As an important means of railway operation management, railway operation index reflects the quality of railway operation. The customer's loading number and tonnage are used as the reference factors of customer's task contribution index.

2.1.2 Customer potential value analysis.

Customer potential value refers to the value and profit that customers may bring to enterprises in the future economic behavior during the railway operation. The potential value of customers is a future-oriented economic behavior. The potential value of freight railway customers can be analyzed from three aspects: growth, honesty and loyalty.

- Growth. Growth represents the improvement ability of customers' freight demand and the embodiment of customers' own development potential. The reference factors of growth indicators include the year-on-year growth of demand and the year-on-year demand growth rate.
- Honesty. Honesty is an important index to measure the potential value of customers, which directly affects the risk of railway freight transactions. By analyzing the customer's economic behavior, the cash rate of car invitation and the unsubscribe rate are used as indicators to measure the customer's integrity.
- Loyalty. Loyalty refers to the dependence and specificity of railway customers on railway freight, and there is a positive correlation between dependence and loyalty. Customer loyalty evaluation indicators mainly include delivery frequency and customer age. Delivery frequency can effectively indicate the frequency of customers using railway transportation, and it is the core index to evaluate customer loyalty. Passenger age is also an important indicator reflecting the degree of customer dependence on railway transportation.

2.1.3 Product matching degree analysis.

The main factors that affect the choice of transportation mode are: transportation time, transportation cost, transportation capacity, transportation reliability, transportation safety and accessibility. Therefore, transportation distance, transportation time and railway transportation ratio are mainly selected as the analysis indexes of product matching degree.

2.2 Improved k-means algorithm.

At present, customer segmentation mainly adopts classification technology and clustering technology. Clustering technology can subdivide railway freight customers into different categories without customer category identification in the railway freight database, and maximize the differences between different categories of customers, so this paper chooses clustering technology to subdivide existing railway freight customers.

K-Means algorithm is a very classic distance-based partition clustering algorithm, which can effectively aggregate objects with similar distance into a group and keep the distance between different groups as large as possible. The core problem of K-Means algorithm is how to measure the distance between samples and how to cluster them.

In order to effectively evaluate the differences between sample data, the K-Means algorithm regards the sample containing n variables as a point in the n-dimensional space, and judges the similarity of the sample by calculating the spatial distance between points. Because the sample data variables analyzed by the K-Means algorithm are all in numerical form, the distance between different points can be measured by Euclidean distance. Assuming that X and Y are sample data containing n variables, the Euclidean distance between X and Y is the square root of the sum of squares of the differences between the values of n variables, and the calculation formula is:

$$d(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}, \tag{1}$$

The disadvantage of distance formula method is that Euclidean distance treats the differences between different attributes in data set equally, and the contributions of each attribute to Euclidean distance are the same, while the contributions of attributes with small differences to Euclidean distance are very small, Euclidean distance often cannot meet the needs of the actual situation. Therefore, the traditional K-Means algorithm is improved based on the identification and processing of outliers [6].

1.Screening outliers.

S_i represents the sum of the distances between point I and other points:

$$S_i = \sum_{j=1}^{n}\sqrt{\sum_{k}^{d}\left(X_{ik} - X_{jk}\right)^2}, \tag{2}$$

Where X_{ik} and X_{jk} respectively represent the coordinate values of two points. V represents the average sum of distances between point I and other points:

$$V = \left(\sum_{i=1}^{n}\frac{D_i}{n}\right) * \alpha, \tag{3}$$

Where n is the dimensionality of the D-bit data of the sample data. For each data point I, firstly, the sum of the distance S_i and sum of average distance V is calculated, and α are the distance adjustment factors, which are determined to be 1.05 by calculation, that is, when the average distance of a point from the population data to be

clustered is 1.05 times, it is considered as an isolated point and screened out from the initial sample.

2.Dealing with isolated points.

Because the contribution of attributes with small dissimilarity to Euclidean distance is very small, it is not possible to judge that isolated points belong to dirty data only by distance and delete them directly. Therefore, we should judge the isolated points according to their similarity with the cluster, and the similarity can be measured by variance. If the variance within the cluster is larger, the similarity within the cluster is smaller. The smaller the variance within the cluster, the greater the similarity within the cluster[7].

The variance D_k of class k is expressed as:

$$D_k^2 = \sum_{x \in A_k, j=1}^{n} \left(X_j - C_j^k\right)^2 / N_{sk}, \tag{4}$$

Where A_k represents the space formed by data set X within class K; X represents the data set of class k; X_j represents the value of the data point on the variable j; N_{sk} represents the number of data sets in class k, and C_j^k represents the value of the center point of k on variable J. Calculate the relative distance between the centers in each cluster of each isolated point, which is expressed as:

$$d^2(i,k) = D_k^2 \sum_{j=1}^{n} (X_{ij} - C_j^k)^2, \tag{5}$$

Calculate the relative distance between the centers in each cluster of each isolated point, and assign the isolated point to the nearest cluster. If the SSE after adding the isolated point is greater than that before adding the isolated point, the isolated point will be abandoned, otherwise, assign the isolated point to the cluster.

2.3 Experimental analysis.

Select some customer data of a railway bureau in November 2022, and the data comes from the railway freight ticket system. Because most railway business systems are electronic, the original customer data is relatively complete, which can basically meet the requirements of customer segmentation [8]. However, due to the different data rules and business parameters of each business system, there are differences and repetitions among various customer data sources, and some noises and missing data need to be preprocessed. The preprocessing methods used include data integration, data cleaning and data conversion.

The preprocessed 3,279 pieces of customer data are classified in three dimensions: current value, potential value and product matching degree, and the results are shown in Table 1.

Table 1. Customer value segmentation results

Field name	Classification 1	Classification 2
Invoice revenue	237850.35	801271.10

Loading number	36.43	188.95
Loading tonnage	2815.65	12662.80
Year-on-year growth of demand	8.24	23.58
Year-on-year demand growth rate	12.63%	7.88%
Cash rate of car invitation	98.26%	94.18%
Unsubscribe rate	1.71%	0.82%
Delivery frequency	77.38	183.74
Customer age	2.52	7.43
Transportation distance	521.60	1832.21
Transportation time	3.51	8.79
Railway transportation ratio	62.87%	84.39%

Based on the customer value classification results of the three variables of current value, potential value and product matching degree, and using the SQL statement to write a customer type discrimination statement to determine the segmentation type of each customer, customers can be divided into eight types: high present value-high potential value-high matching degree (Class I), low present value-high potential value-high matching degree (Class II), low present value-high potential value-low matching degree (Class III), high present value-high potential value-low matching degree (Class IV), high present value-low potential value-high matching degree (Class V), low present value-low potential value-high matching degree (Class VI), low present value-low potential value-low matching degree (Class VII), high present value-low potential value-low matching degree (Class VIII). A total of 278 cases of Class I customers were obtained, accounting for 8.48%. There were 221 customers of Class II, accounting for 6.74%. There were 336 customers in Class III, accounting for 10.28%. There were 367 customers in Class IV, accounting for 11.19%. There were 417 customers in Class V, accounting for 12.72%. There were 438 customers in Class VI, accounting for 13.36%. There were 493 customers in Class VII, accounting for 15.06%. There were 729 customers in Class VIII, accounting for 22.23%. Customer Dimension Classification Results is shown in Figure 2:

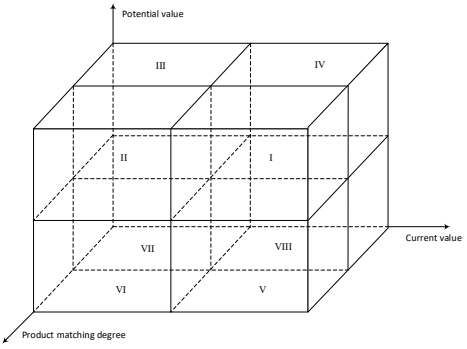

Fig. 2. Customer dimension classification diagram

Based on the segmentation results of railway freight customers, different types of customers have different characteristics. For example, Customers with high present value-high potential value-high matching degree (Class I) have the highest customer value, make great contributions to the railway revenue, and have the potential to continue to bring huge benefits to the railway. Railways can tilt important transportation resources to this kind of customers, give priority to meeting the capacity needs of this customer group [9]. However, customers with low present value-low potential value-low matching degree (Class VII) can bring little income to the railway now and in the future, and occupy the development and maintenance expenses of the enterprise, so they belong to the "eliminated customers" of the railway, and the resource investment for such customers should be reduced [10].

3 Conclusion

Based on the improved K-Means algorithm, the customer segmentation model of freight railway determines the customer classification comprehensively and carefully from a three-dimensional perspective, which provides reference for the railway freight department to distinguish customer categories and provide differentiated and targeted services for customers, so as to improve the competitiveness and market share of railway freight.

References

1. Zhang Wenqing. Research on Railway Freight Customer Evaluation Based on Fuzzy Comprehensive Evaluation Method, J. Railway freight, 2018, 36(08):32-36.
2. Cui Xiaoyun, Wang Huanhuan, Qian Shenyi. Application of Improved K-means Algorithm in Customer Classification, J. Modern Computer (Professional Edition), 2016(24):25-27+31.
3. Peng Jianfeng, Xu Banmin, Zhang Yixiang. Key Technologies and Research on Railway Sensitive Data Security Based on Equal Protection 2.0, J. Network Security Technology and Application, 2021, 241(01):138-142.
4. Zheng Pingbiao, Zhu Kefei, Dai Mingrui. Discussion on Railway Customers Classification Method based on Customer Value Analysis, J. Railway transport and economy, 2014, 36(03):42-46.
5. Chen Ning, Sun Xiaoyang, Gong Depeng. Study on Accurate Identification of Railway Freight Customers Based on Business Intelligence, J. Integrated transportation, 2018, 40(07):103-109.
6. Deng Cheng, Yang Zhuangying, Gu Junjie, Cai Zhi, Li Yue. Improved K-means Algorithm and its application in railway customer segmentation, J. Railway computer application, 2014, 23(06):45-48.
7. Zhang Bin, Peng Qiyuan. Research on Customer Segmentation Method of Chinese Railway Freight Transport Based on KFAV, J. Transportation System Engineering and Information, 2017, 17(03):235-242.
8. Yue Song. Analysis on Data Asset Management of Freight Railway Enterprises, J. Volkswagen standardization, 2021, 346(11):197-199.

9. Jiang Xiaohong, Wang Ye, Zheng Tan. Logistics Customer Classification Based on K-means Clustering Algorithm and R Language Implementation, J. Logistics Engineering and Management, 2018, 40(12):66-68+65.
10. Liu Shuying, Zou Yanfei, Li Hong. Customer Value Analysis of Airlines Based on K-Means Algorithm, J. Digital Technology and Applications, 2021,39(11):10-12.