International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 4 ǁ 2022 ǁ PP. 05-13

Data Value Mining: A Case Study of Airline Customer Data

Pingshui Wang[1], Tao Chen[2]
[1] School of Management Science and Engineering, Anhui University of Finance & Economics, Bengbu, China
[2] School of Management Science and Engineering, Anhui University of Finance & Economics, Bengbu, China
Corresponding Author: Pingshui Wang

Abstract
With the development of the Internet, cloud computing, big data, artificial intelligence and other new generation of information technology, the potential value of data is easier to be mined and utilized. This paper takes airline customer data as an example, extracts valuable information for enterprise decision-making using effective data processing and cluster analysis technology, and visualizes it, so as to optimize marketing strategy and achieve enterprise profit maximization.

Keywords: *Big data; Python language, Cluster analysis, Customer relationship management, Personalized service.*

Date of Submission: 02-04-2022 Date of acceptance: 16-04-2022

I. INTRODUCTION

With the advent of the information age, the marketing focus of enterprises has shifted from products to customers, and CRM (Customer Relationship Management) has become the core problem of enterprises. The key problem of CRM is customer clustering to distinguish valueless customers from high-value customers by customer grouping. Enterprises develop personalized service plans for customers with different values, adopt different marketing strategies, and focus limited marketing resources on high-value customers to achieve the goal of maximizing corporate profits [1]. Accurate customer clustering result is an important basis for enterprises to optimize marketing resource allocation. Customer clustering is becoming one of the key problems to be solved in customer relationship management.

This paper will combine with the RFM (Recency Frequency Monetary) model and group customers using k-means clustering algorithm and compare the customer value of different types of customers based on use the airline customer data, so as to develop the corresponding marketing strategy and maximize the interests of the company.

II. ANALYSIS OF CURRENT SITUATION AND CUSTOMER VALUE OF AIRLINES

In the face of fierce market competition, airlines have launched more concessions to attract customers. A domestic airline is facing a management crisis, such as frequent passenger loss, declining competitiveness and under-utilization of resources. By establishing a reasonable customer value evaluation model to classify customers, analyze and compare the customer value of different customer groups, and develop the corresponding marketing strategy, to provide personalized service to different customer groups.

2.1 Analysis of the current situation of airlines

2.1.1 Competition within the industry

In addition to the competition among the three major airlines, the competition of civil aviation will also include various newly emerging small airlines, private airlines and even foreign aviation giants. Overproduction of aviation products, product homogeneity is more and more obvious, so airlines from price, service competition gradually turned to customer competition.

2.1.2 Competition outside the industry

With the construction and rapid development of high-speed rail, bullet train and other railway transportation, airlines have been greatly impacted.

2.1.3 Analysis of airline data features

Currently the airline has accumulated a large amount of information about its members' files and their flight records.

We use March 31, 2021 as the end time, select a period with a width of two years as the analysis observation window, and extract the detailed data of all customers with in-flight records in the observation window to form historical data with 44 features, a total of 62988 records, which contains the *membership number, initiation time, gender, age, level of membership card, the city is carried out, the province is carried out, the country is carried out, the end time of the observation window, total accumulative integral, the observation window of total km flight number, flight number within the observation window, average flight time interval and average discount coefficient*. The data table structure features: Shape (62988, 44).

2.1.4 Paper objectives

Based on the current data of airlines, this paper can achieve the following objectives.
(1) Classify customers using airline customer data.
(2) Perform a feature analysis of different customer categories and compare the customer value of different customer categories.
(3) Provide personalized service to different value customer categories and develop corresponding marketing strategy.

2.2 Customer value analysis of airlines

The global economic environment and market environment have changed quietly, and the business of enterprises has gradually shifted from product-oriented to customer-demand-oriented. A new "customer-centric" business model is emerging and being promoted to unprecedented heights. However, maintaining a relationship with customers costs money, and only part of the customers owned by an enterprise can bring profits to the enterprise. The resources of an enterprise are also limited. Ignoring high-potential customers and providing the same service to all customers will make the enterprise's resources unable to exert their maximum utility to create maximum profits. Any enterprise wants to survive and develop, must obtain profit, the pursuit of profit maximization is one of the purposes of enterprise survival and development. Therefore, enterprises cannot and should not maintain the same relationship with all customers.

Jay & Adam Curry, an advocate of customer marketing strategy, extracted the following experience from the experience of hundreds of foreign companies carrying out customer marketing implementation [2].

- derives 80% of revenue from the top 20% of customers.
- 20% of a customer achieves a 100% profit margin.
- derives more than 90% of its revenue from existing customers.
- most of the marketing budget is often spent on non-existing customers.
- 5% to 30% of customers have the potential to upgrade in the customer pyramid.
- customer pyramid 2% customer upgrade translates into a 10% increase in sales revenue and 50% increase in profit.

These experiences may not be completely accurate, but it reveals the trend of customer differentiation in the new era, and also shows the urgency and necessity of customer value analysis. If the profitability of customers is analyzed, it will be found that the structure of customer profitability has changed significantly, and only a certain number of customers bring profits to the enterprise. If enterprises want to obtain long-term development, they must effectively identify and manage such customers. The same approach to all customers who do business with the company will not succeed.

Although many enterprise managers know the importance of customer value analysis, they know little about how to do it. How to consider the factors of customer value in all directions and from many angles and carry out effective customer value analysis is a problem that all enterprises need to consider seriously [3]. Only by selecting valuable customers and focusing on them can we effectively enhance the competitiveness of enterprises and achieve greater development.

In the field of customer value analysis, the most influential and empirically tested theories and models are customer lifetime value theory, customer value pyramid model, strategic evaluation matrix analysis and RFM customer value analysis model [4,5]. This paper will be analyzed using the improved customer value RFM model.

2.3 Steps and processes of airline customer value analysis

The overall process of airline customer value analysis consists of the following four steps.
(1) Extract the data of airlines from April 1, 2019 to March 31, 2021.
(2) Perform data cleaning, feature construction and standardization for extracted data.
(3) Based on RFM model, k-means algorithm is used for customer clustering.
(4) For customers with different values obtained from the model, different marketing methods are adopted to provide customized services.

III. AIRLINE CUSTOMER DATA PREPROCESSING

There are a small number of missing values and outliers in the original data of airline customers, which need to be cleaned before being used for analysis. At the same time, because the original data has too many features, it is inconvenient to be directly used for customer value analysis. Therefore, it is necessary to screen the features and select the key features to measure customer value.

The pre-processing of airline customer data can be divided into the following three steps.

(1) Processing data missing values and outliers.

(2) Feature screening combined with RFM model.

(3) Standardized the screened data.

3.1 Handle missing values and outliers of data

Through data observation, it is found that in the original data, there are records that the ticket price is null, the minimum ticket price is 0, the minimum discount rate is 0, and the total flight mileage is greater than 0. The data with empty ticket price may be caused by the fact that the customer has no flight record.

Other data may be due to customers taking 0 percent off flights or redeeming points. Due to the large amount of original data, such data accounts for a small proportion and has little impact on the problem. Therefore, the data is discarded. The specific treatment method is as follows.

(1) Discard the record that the fare is empty.

(2) Discard a record of a fare of 0, an average discount rate of non-0 and a total flying kilometres of greater than 0.

3.2 Key features of airline customer value analysis

The goal of this paper is customer value analysis, that is, to identify customers of different values through airline customer data. RFM model is the most widely used model to identify customer value.

3.2.1 RFM Model Introduction

- R (Recency) refers to the interval between the most recent consumption time and the cut-off time. In general, the less time elapsed between the last purchase and the deadline, the more likely you are to be interested in goods or services that are immediately available.
- F (Frequency) refers to the number of times a customer consumes in a certain time. It can be said that customers with higher consumption frequency are also customers with higher satisfaction, higher loyalty and higher customer value.
- M (Monetary) refers to the amount consumed by a customer at a certain time. The 80-20 rule that "20 percent of customers account for 80 percent of sales" means that customers who spend more have more spending power.

3.2.2 Interpretation of RFM model results

RFM model includes three features, which are displayed in a Three-Dimensional coordinate system, as shown in Fig. 1. X axis represents Recency, Y axis represents Frequency, and Z axis represents Monetary. Each axis is generally divided into five levels to indicate degree, with 1 being the minimum and 5 being the maximum.

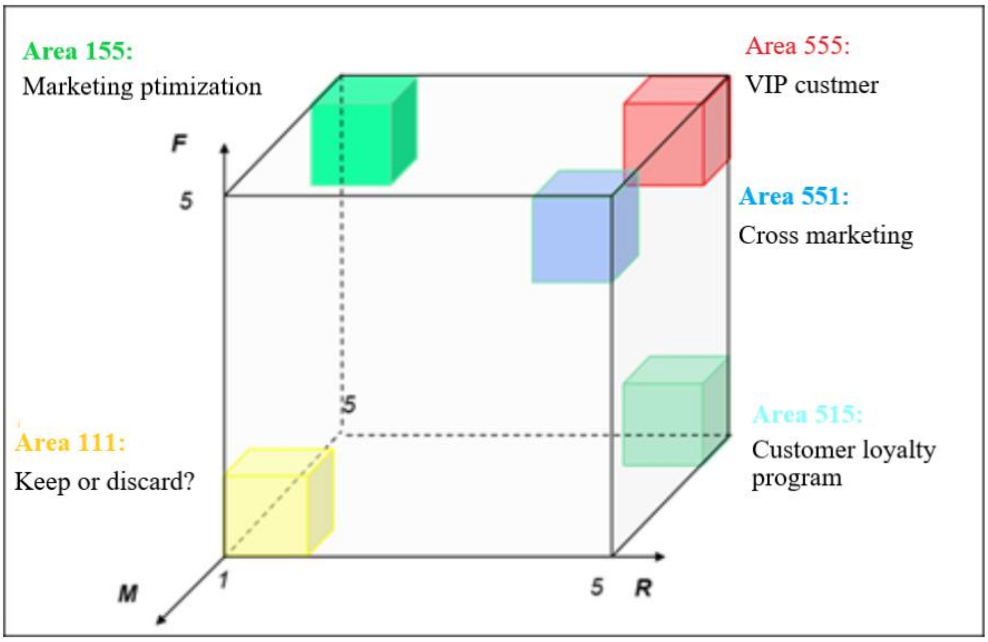

Figure 1: RFM Customer value model

3.2.3 Defects of traditional RFM model in aviation industry

In RFM model, consumption amount over a period of time, the customer purchase amount is the sum of the products of this enterprise, due to airline ticket prices by the transport distance, space level and other factors, also consumption amount of different value of airline passengers is different, so this feature is not suitable for the airlines customer value analysis.

3.2.4 LRFMC model for airline customer value analysis

In this paper, two features are selected to replace the consumption amount: the flight mileage accumulated by the customer within a certain period of time M and the average value C of the discount coefficient corresponding to the passenger seat within a certain period of time. In addition, the membership time of airline members can affect customer value to a certain extent, so customer relationship length L is added to the model as another feature to distinguish customers.

In this paper, customer relationship length L, consumption time interval R, consumption frequency F, air miles M and average value C of discount coefficient are taken as the key features for airlines to identify customer value, and are recorded as LRFMC model.

According to the airline customer value LRFMC model, six LRFMC features FFP_DATE, LOAD_TIME, FLIGHT_COUNT, AVG_DISCOUNT, SEG_KM_SUM and LAST_TO_END are selected. Delete irrelevant, weakly related or redundant features, such as *membership card number*, *gender*, *city of work*, *country of work*, *age*, etc.

3.2.5 Build key features of airline customer value analysis

Since the five features of LRFMC model are not directly presented in the original data, these five features need to be extracted from the original data.

(1) The number of months between the time of membership and the end of observation window L= the end time of observation window - Time of membership (in month), as shown in follows.

L=LOAD_TIME-FFP_DATE

(2) The number of months from the end of the observation window to the last flight time of the customer R= the time from the last flight time to the end of the observation window (unit: month), as shown in the following.

R = LAST_TO_END

(3) Number of customers taking the company's aircraft in the observation window F=(number of consumption in a certain period of time)

F=FLIGHT_COUNT

(4) Customer flight mileage in the observation window M= total flight kilometers in the observation window (unit: kilometers)

M=SEG_KM_SUM

(5) The average discount coefficient of the passenger seat in the observation window C= average discount rate (Unit: none)
C=AVG_DISCOUNT

3.2.6 Standardize the five features of LRFMC

After the construction of five features is completed, the distribution of data of each feature is analyzed, and the value range of data is shown in Table 1. It can be seen from the data in the table that the value ranges of the five features differ greatly. In order to eliminate the influence brought by the order of magnitude data, standardized data processing is required.

Table 1 Five features of airline data

Feature names	L	R	F	M	C
Minimum	12.17	0.03	2	368	0.14
Maximum	114.57	24.37	213	580717	1.5

Data samples of L, R, F, M and C are shown in Table 2 for original data and Table 3 for data after standard deviation standardization. The k-means algorithm is used for customer clustering

Table 2 The sample of original airline data with five features

LOAD_TIME	FFP_DATE	LAST_ TO_END	FLIGHT_ COUNT	SEG_K M_SUM	AVG_DIS COUNT
2021/3/31	2020/3/16	23	14	126850	1.02
2021/3/31	2021/6/26	6	65	184730	0.76
2021/3/31	2018/12/8	2	33	60387	1.27
2021/3/31	2018/12/10	123	6	62259	1.02
2021/3/31	2020/8/25	14	22	54730	1.36

Table 3 The sample of standardized airline data

L	R	F	M	C
1.44	-0.95	14.03	26.76	1.30
1.31	-0.91	9.07	13.13	2.87
1.33	-0.89	8.72	12.65	2.88
0.66	-0.42	0.78	12.54	1.99
0.39	-0.92	9.92	13.90	1.34

IV. USE K-MEANS ALGORITHM TO CLASSIFY CUSTOMERS

4.1 Understand k-means clustering algorithm

4.1.1 Basic Concepts

K-means clustering algorithm is a classification method based on centroid (clustering center). Input the number of clustering K and database containing N data objects, and output K clusters that meet the minimum standard of error sum of squares. The algorithm steps are as follows.
(1) Randomly select K objects from N sample data as the initial clustering center.
(2) Calculate the distance between each sample and the center of mass of each cluster, and assign the sample to the cluster center category with the closest distance.

(3) After all samples were allocated, the centers of k clusters were recalculated.
(4) Compared with the k cluster centers obtained in the previous calculation, if the cluster centers change, go to (2); otherwise, go to (5).
(5) Stop and output the clustering results when the clustering center does not change.

4.1.2 Data Types

K-means clustering algorithm is studied on the basis of numerical type data. However, the samples of data analysis are complex and diverse, so it is required not only to analyze the data with numerical type characteristics, but also to adapt to the changes of data types and make different transformations for different features to meet the requirements of the algorithm.

4.1.3 Kmeans function and its parameters

The core function of the K-means algorithm in Python is kmeans, which comes from the Cluster software package of The Python machine learning library SciKit-learn. Its basic syntax is as follows.
kmeans(n_clusters=4, init=’k-means++’, n_init=10, max_iter =300, tol=0.0001, precompute_distances=’auto’, verbose=0, n_jobs=1, random state=None)

After the k-means model is constructed, different information can be viewed through attributes, as shown in Table 4.

Table 4 Properties and Description

Attributes	Meaning
cluster_centers	Return *ndarray*: Represents the mean vector of the classification cluster
labels_	Return *ndarray*: A tag that represents the cluster to which each sample belongs
Inertia_	Return *ndarray*: Represents the sum of the nearest cluster centers of each sample

4.2 Analysis of clustering results

The clustering results of aviation customer data are shown in Table 5.

Table 5 The result of clustering

Clustering category	Clustering number	Clustering center				
		L	R	F	M	C
Customer group 1	5336	0.483	-0.799	2.483	2.425	0.309
Customer group 2	4171	0.056	-0.003	-0.226	-0.229	2.200
Customer group 3	15742	1.160	-0.377	-0.087	-0.095	-0.156
Customer group 4	24663	-0.700	-0.415	-0.161	-0.161	-0.254
Customer group 5	12132	-0.313	1.686	-0.574	-0.537	-0.173

Feature analysis was conducted based on the clustering results, as shown in Fig. 2.

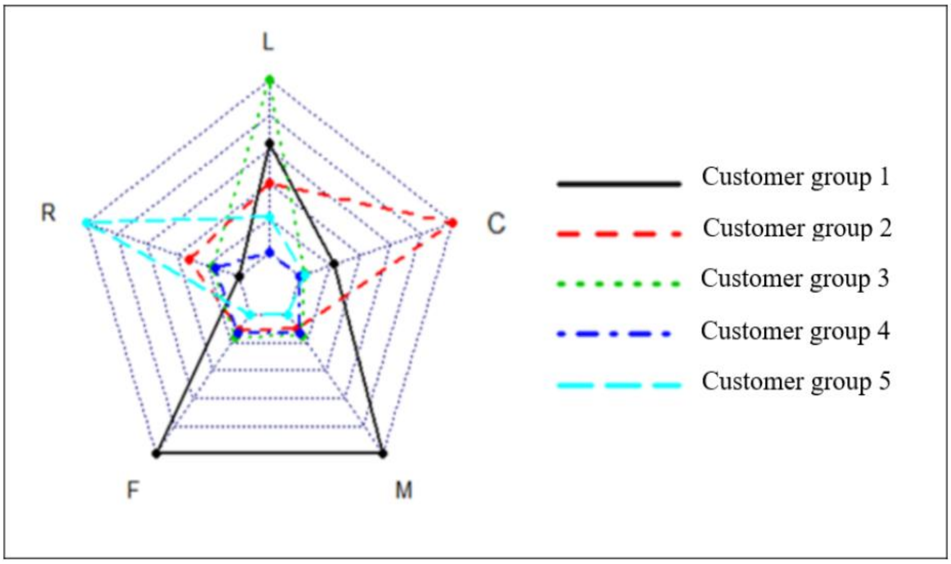

Figure 2 Characteristic analysis of clustering results

Combined with the business analysis, the characteristics of a certain group were evaluated and analyzed by comparing the size of each feature between groups, so as to summarize the advantages and disadvantages of each group. The specific results are shown in Table 6.

Table 6 Evaluation and analysis of cluster features

Group category	Advantage feature	Weakness feature
Customer group 1	F M *R*	
Customer group 2	*C*	R F M
Customer group 3	F M L	
Customer group 4		*L* *C*
Customer group 5		*F* *M* R

Note: Normal font indicates the maximum value, bold font indicates the sub-maximum value, italic font indicates the minimum value, and underlined font indicates the sub-minimum value.

Based on the characteristic description, this paper defines five customer categories: important maintain customers, important development customers, important retention customers, general customers and low value customers.

(1) Important maintain customer. The average discount coefficient (C) of such customers is high (the class of the flight is generally high), the length of time since the most recent flight (R) is low, and the number of flights (F) or total mileage (M) is high. They are the high-value customers of airlines, the most ideal type of customers, the largest contribution to the airline, but a small proportion. Airlines should give priority to their resources, carry out differentiated management and one-to-one marketing to improve the loyalty and satisfaction of these customers, and extend the high level of consumption of these customers as much as possible.

(2) Important development customers. The average discount coefficient (C) of such customers is high, the length of time since the most recent flight (R) is low, but the number of flights (F) or total miles flown (M) is low. Such customers have a short membership time (L), and they are the potential value customers of airlines. Although the current value of this kind of customer is not very high, but it has great potential for development, airlines should strive to promote this kind of customer to increase their in-flight consumption and the consumption of partners, that is, to increase the share of customers' wallets. Through the improvement of customer value, the satisfaction of such customers will be strengthened, and their transfer costs to competitors will be increased, so that they will gradually become loyal customers of the company.

(3) Important retention of customers. The average discount factor (C), number of flights (F) or total mileage (M) of flights taken by such customers in the past are high, but they have not taken flights of the company for a long time (R high) or have taken less frequent flights. This type of customer value change is highly uncertain. As the reasons for the decline of such customers vary, it is important to keep abreast of the latest information and maintain interactions with customers. According to the recent consumption time and consumption times of such customers, airlines should predict the change of customer consumption, and make a list of customers, focus on their contact, take certain marketing measures to extend the life cycle of customers.

(4) General customers and low value customers. The average discount coefficient (C) of the flights taken by such customers is very low, the longer companies have not taken the flights of the company (R is high), the number of flights (F) or total mileage (M) is low, and the membership time (L) is short. They are regular users and low value customers of airlines and may fly with airlines when their tickets are on sale.

Among them, important development customers, important retention customers and important retention customers can be classified into three stages of customer life cycle management: development period, stable period and declining period.

According to the characteristics of each customer type, we rank the customer value of all kinds of customers, provide different products and services for different types of customers, improve the value of important customers, stabilize and prolong the high level of consumption of important customers, prevent the loss of important customers and actively restore the relationship.

This model uses historical data for modeling, and the observation window for analyzing data changes over time. Therefore, for the detailed information of newly added customers, considering the actual situation of the business, the model suggests to run once a month, judge the information of newly added customers through the cluster center, and analyze the characteristics of newly added customers. If the actual situation of incremental data differs greatly from the judgment result, the business department needs to pay attention to it, check the cause of the big change and confirm the stability of the model. If the model stability changes greatly, the model needs to be retrained for adjustment. At present, there is no unified standard for the time of model retraining, and most cases are determined by experience. According to experience, it is appropriate to train the model every six months.

4.3 Model Application

According to the analysis of the characteristics of each customer group, the following marketing means and strategies are adopted to provide reference for the management of value customer groups of airlines.

(1) Member upgrade and retention: An airline may, prior to the time of evaluation of a member upgrade or retention, provide appropriate reminders and even some promotional activities to customers who approach but do not meet the requirements to stimulate them to achieve the corresponding standards through consumption. In this way, we can not only gain revenue, but also improve customer satisfaction and increase the elite members of the company.

(2) **First conversion:** An approach to extract members from a database who are close to, but have not yet achieved, the first conversion standard and to remind or promote them to achieve the standard through consumption. Once the first exchange is made, it is much easier for the customer to re-exchange at the company than at other companies, which to some extent increases the cost of the transfer.

(3) **Cross-selling:** through collaboration with non-aviation enterprises, such as the issue of co-branded card, the customer gains points of the company in the consumption process of other enterprises, enhancing the relationship with the company and enhancing their loyalty.

V. CONCLUSION

This paper takes airline customer data as an example, and focuses on the application of k-means clustering algorithm in data analysis algorithm in airline customer value analysis. Aiming at the deficiency of RFM customer value analysis model, k-means algorithm is used to construct aviation customer value analysis LRFMC model, which describes the whole process of data analysis in detail, and puts forward relevant countermeasures and suggestions according to the data analysis results.

ACKNOWLEDGMENT

We thank the anonymous reviewers and editors for their very constructive comments. This work was supported by the virtual teaching and research section of computer foundation of Anhui University of Finance & Economics of China under Grant No. acxnjys2021003, the teaching research project of Anhui University of Finance & Economics of China under Grant No. acjyzd2020025, the key teaching research project of graduate school of Anhui University of Finance & Economics under grant No. cxjhjyzda1803 and the teaching research project of Anhui Provincial Education Department under grant No. 2020jyxm0037 and No. 2020zyrc020.

REFERENCES

[1] Li Y. F., Luo T., Wang J. L. (2020) “Combination and Coexistence of Dasi Teting Business Model: Complementary or Mutually exclusive?” Case Library of China Management Case Sharing Center.

[2] Zhang L. G. (2016) “Data Mining in Python” Beijing: China Machine Press.

[3] Huang H. M., Zhang L. G. (2017) “Python Data Analysis and Application” Beijing: Posts and Telecommunications Press.

[4] Quan Y. B. (2020) “Customer Profitability Analysis as a Breakthrough Point to Explore Customer Value” China Rural Finance. No.4, pp.77-78.

[5] Liu H., Li J. S. (2018) “Reader potential Value Prediction Based on Customer Relationship Management” Publishing science. Vol. 20, No.6, pp. 69-73.