Air Passenger Image Construction Based on Data Mining Technology

Zeyu Zhou, Yisheng Wang, Yan Yu
(School of Business Administration, Hohai University, Changzhou Jiangsu China)

About the author: Zeyu Zhou, Email: zeyu_zhou@qq.com, undergraduate; Yisheng Wang, undergraduate; Yan Yu, undergraduate;

Abstract This paper aims to reduce the airline's vacant rate and make full use of aviation resources to increase the profit by constructing the air passenger image model and forecasting the passenger loss rate based on mining data technology. Firstly, this paper groups the passengers to establish the LRFMC model and obtain the weights of the five indicators by AHP analytic hierarchy process. After weighting the five indicators, the K-means rapid clustering method is used to cluster the passengers. Secondly, construct a logistic regression model to predict the probability of loss. Lastly, based on the conclusions of the established passenger segmentation and loss model, different services and marketing strategies are proposed for different travel groups to attract passengers to take flights and to improve attendance and benefits.

1. Research background
Nowadays, with the continuous development of big data and data science and technology, how to deal with one of the basic characteristics of big data, Valueless, has become an urgent problem to be solved. Taking airlines as an example, airlines have a large number of passenger user information and member data. How to effectively extract and subdivide these massive data and realize accurate and personalized services for passengers has become a topic that scholars at home and abroad are working on. However, as to the air passenger image construction and prediction method of passenger loss probability, scholars at home and abroad have made some research. Wenxin Zhang(2009) [1] made use of the factual data to classify the frequent flyer sample of airlines by a K-means clustering method based on genetic algorithm, compared and analysed the value of the various customers, determined the type of customers and adopts different management strategies aiming at different types. Tongshui Wu and Liang He(2006) [2] carried out the experiment and analysis of air passenger loss through the ID5 algorithm in the decision tree and obtained some rules of passenger loss. Xiangcui Wang(2008) [3] used the analytic hierarchy process to construct a retail passenger loss analysis model, and a formula for estimating passenger loss was proposed. Yaqi Cui(2018) [4] established the passenger loss warning model by the C5.0 algorithm in the decision tree, and a certain degree of empirical analysis was carried out.

Based on the 62,894 member passenger datasets of domestic airlines, this paper constructs a logistic passenger loss prediction model by constructing the LRFMC model, and using the AHP analytic hierarchy process and K-means rapid clustering method, to predict the loss probability in order to propose different services and marketing strategies for different travel groups.
2. Data preprocessing
Processing missing values: Firstly, the R software is used to process the data set, and 1109 missing values are obtained, accounting for 1.76% of the total sample. These missing samples had little effect on the population, so these missing values are eliminated.

(1) Data reduction: This paper is mainly based on a large number of references, meaningful variables are selected from 58 variables for research, so as to achieve dimensional reduction. Regarding the prediction of passenger loss probability, this paper uses regression analysis in this paper to selects those variables which have significant impact.

(2) Data transformation facilitates data analysis by transforming data into appropriate forms. And certain transformation, smoothing and standardizing processing are used to eliminate dimensional effects.

3. Mining steps and a brief description of methods
Establishing the LRFMC model and calculating the weight of each index using AHP analytic hierarchy process.

This paper use the weight value obtained, weighting each index, and obtained the weighted score. The weighed score results are clustered with K-means fast clustering method, so as to classify the passengers.

4. Air Passenger Image Construction
1. AHP analytic hierarchy process

(1) Indicator System Construction

Based on many database marketing experiences, marketing expert Bob Stone proposed the RFM segmentation, which is detailed by three variables, namely the recent consumption interval (Recency), consumption frequency (Frequency), and consumption amount (Monetary), to identify the most valuable passengers.

Passengers with high FRM indicators have bigger commercial value because they have a higher probability of continuing to choose the company for further cooperation. Passengers with poor RFM indicators represent fewer business opportunities, indicating that the value of the passenger is lower. Therefore, the RFM model can help companies decide who to promote, use optimal marketing tools to capture and retain the most valuable travellers, avoid investing in specific characteristics of airlines, and properly adjust and improve traditional RFM metrics. Based on the LRFMC model, this paper will construct the portraits of passengers and select the five indicators of L, R, F, M and C to quantify the variables of airline passenger segmentation. L represents the length of the passenger relationship (from the date of membership), R represents the length of the passenger's last consumption, F represents the passenger's consumption frequency within a certain period of time, and M represents the passenger's upgrade mileage within a certain period of time ("Upgrade mileage "refers to the basic flight mileage accumulated by airline members on the valid flights of the company and airline partners." C represents the average space discount factor for passengers traveling within a certain period of time (ie, the member corresponds to the seat in a certain period of time). Average discount factor).

(2) Algorithm principle

In this paper, the modeling of AHP is described as follows: 1. Constructing the hierarchical structure of the hierarchy; 2. Constructing the comparison discriminant matrix; 3. Verifying the ordering and consistency under the single criterion; 4. General Sorting selection.

(3) Result analysis

In order to more objectively reflect the influence of different indicators in the LRFMC model on passenger behavior, this paper uses the analytic hierarchy process to calculate the weights of five indicators in the LRFMC model and construct the judgment matrix and then the consistency of the judgment matrix is tested. Finally, the calculation results of the weights of the indicators are as follows:
### Table 1 LRFMC Model

| Indicators System | Meaning                                                                 | Symbol | Units | Value change | Definition                                                                 | Weight |
|-------------------|-------------------------------------------------------------------------|--------|-------|--------------|-----------------------------------------------------------------------------|--------|
| LOAD_TIME - FFP_DATE | Number of months from the end of the observation window for membership | L      | Month | ↑            | Length of passenger relationship                                             | 0.039  |
| DAYS_FROM_LAST_TO_END | The last flight time to the end of the observation window | R      | Day   | ↓            | the length of the passenger's last consumption                             | 0.088  |
| FLIGHT_COUNT      | Number of flights                                                       | F      | Time  | ↑            | consumption frequency within a certain period of time,                       | 0.239  |
| SEG_KM_SUM        | Total flight kilometers in observation window                           | M      | Kilometre | ↑          | Upgrade mileage within a certain period of time                               | 0.123  |
| avg_discount      | Average discount rate                                                   | C      | Percentage % | ↑          | The average space discount coefficient during a certain period of time       | 0.511  |

Note: The symbol "↑" indicates that the larger, the better, "↓" indicates that the smaller, the better. According to the weights obtained, this paper weights the five major indicators: L, R, F, M, and C to obtain weighted scores. The weighted scores not only reflect the difference in the importance of the five major indicators, but also lay the foundation for the rapid clustering.

### 2. K-means rapid clustering method

(1) Basic principle of rapid clustering

The advantage of K-means fast clustering over other clustering methods is that it is suitable for larger data sets, but the disadvantage is that it may be affected by outliers.

(2) Determine the number of clusters

In order to determine the number of clusters, this paper first makes a preliminary exploration. Before the four graphs are shown in the following figure, there is a significant downward trend in the square synthesis in the group. After being grouped into three categories, the rate of decline has been significantly reduced, indicating that the choice which is clustered into four to five categories is a suitable fit cluster for this data set.
In order to further determine the number of clusters, this paper divides the passengers into five categories: important development passengers, important retention passengers, important maintenance passengers, general passengers, and low-value passengers based on a large number of references. Therefore, the number of clusters is determined to be five which is more appropriate.

(3) Passenger value analysis

The histogram of the passenger group is as follows:

Based on the results of the clustering, this paper extracts the characteristics of different types of passengers. It is found that the average discount rate of passenger group 1 is high (C), which is medium passenger group; the average discount rate of passenger group 2 is low (C) and the time for becoming a member of the company is short (L), which is a large passenger group; passenger group 3 has a long time for becoming a member of the company (L) which is the largest passenger group; the passenger group 4 has a higher frequency (F) or mileage (M), and has recently taken the company's flight (R) for the small passenger group; Passenger group 5 has recently taken less of the company's flight (R), and the frequency (F) or mileage (M) is lower, which is a medium passenger group. By comparing the size of each index of different passenger groups among groups, the evaluation and analysis of each passenger group are carried out. Among the indicators, the R indicator is good when smaller, and the L, F, M, and C indicators are excellent when bigger. The radar chart visually shows the scores of different categories of passengers on various indicators. Based on the above analysis, we define the value of passengers:

Passenger value: important maintain passengers (4)> important development passengers (1)> important retention passengers (3)> general passengers (2)> low value passengers (5)

| Result | Number | Average score | Category             | Level    |
|--------|--------|---------------|----------------------|----------|
| 4      | 2571   | 1.918         | important maintain passengers | Smaller  |
Passengers that should be maintained: C (the average discount rate) is higher, R is lower, the F (frequency) and M (mileage) are also higher. They are the value travelers of airlines, which is a most ideal type of passengers. They contribute the most to airlines but their proportion is relatively small. Airlines should give priority to their resources on them, increase the loyalty and satisfaction of such passengers, and maximize the high level of consumption of such passengers.

Passengers that should be cultivated: C (the average discount rate) is higher, and the cabin level is high with a lower R, but the F (frequency) and M (mileage) are lower. Some of these passengers have just become members of the company (L is low), and some of them are old members but do not often travel on the company. This paper believes that such passengers are the potential value passengers of airlines, occupying a high proportion of passengers, and have a good development space in the future. Airlines must strive to strengthen the satisfaction of such passengers, making them gradually become loyal passengers of the company.

Passengers that should be detained: C (the average discount rate), the F (frequency) and M (mileage) is higher, but R is lower. The value of their passengers varies greatly and the reasons for the changes vary. Airlines should take certain marketing measures based on the relevant consumption indicators of these passengers to prevent the loss of passengers.

General passengers and low-value passengers: the current value and growth potential are low. Compared with other tourists, they cannot create more value for the company. The company should not waste too much resources on these passengers. More manpower and resources should be invested in more valuable travelers.

5. Passenger loss probability prediction

(1) Establishing index system

| index | Sign |
|-------|------|
| age | X1 |
| Average seasonal flight counts from observation window | X2 |
| Sum of basic accumulated points from observation window | X3 |
| Total weighed flight mileage | X4 |
| Other accumulated points from observation window | X5 |
| Average intervals between flights | X6 |
| Average-discount-rate | X7 |
| Counting of non-plane-taking accumulated points changes | X8 |
| Counting of accumulated points being exchanged | X9 |

Logistic regression model
The problem of air passenger loss is a classification and prediction problem in data mining. First, the classification model is established by training samples, and then the classification of test set samples is predicted according to the model. Because the loss marker is a binary variable, and logistic regression analysis is a generalized linear regression model with dependent variable as classified variable and independent variable as continuous or classified variable, and it can predict the probability of occurrence or non-occurrence of events, if the probability of prediction is greater than 0.5, the prediction will occur and vice versa. Therefore, this paper uses logistic regression model to study the probability prediction of air passenger loss, and randomly selects 80% of the samples as the training set, 20% of the samples as the test set.

Logical regression assumes that the data follows Bernoulli distribution. By using maximum likelihood function and gradient descent method to solve the parameters, so as to classified the data into two categories. The graph of the relation between probability and independent variable in binary classification problem is usually a S-type curve, which is realized by using Sigmoid function in logistic regression analysis. In practical applications, probability (P) and dependent variable are usually nonlinear. In order to solve this type of problem, we introduce logit transforming, which makes the relationship between logit (p) and independent variable linear.

Establish the following two-element Logistic regression model:

\[
\log it(p) = \ln \frac{P}{1-P} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5
\]  

To avoid the occurrence of multiple collinearity, this paper first calculate the correlation of the selected indexes. In the correlation diagram below, a darker color indicates a stronger correlation between the indicators. Overall, the correlation between the indicators is within an acceptable range.

Next, logistic regression analysis is conducted. After step-by-step regressing, the variables that had significant impact on the model were screened out, namely: X1-age, X2-quarter average flight count from observation window, X3-total basic score from observation window, X4-total weighted average flight mileage from observation window, X6-average time interval between flights and X9-times accumulated points been exchange.

| Significance index                          | Sign | B    | SE   | Sig  |
|--------------------------------------------|------|------|------|------|
| age                                        | X1   | -0.008 | 0.003 | 0.02 |
| Average seasonal flight counts from observation window | X2   | -2.92 | 0.03  | 0.01 |
| Sum of basic accumulated points from observation window | X3   | 1.23  | 0.045 | 0.0069 |
Based on the result of Logistic regression calculation, we establish the following prediction model of air passenger loss probability.

\[
p(y) = \frac{\exp(-0.008X_1 - 2.92X_2 + 1.23X_3 - 1.696X_4 + 0.215X_5 + 0.461X_6 + 1.62)}{1 + \exp(-0.008X_1 - 2.92X_2 + 1.23X_3 - 1.696X_4 + 0.215X_5 + 0.461X_6 + 1.62)}
\]

The ROC curve is used to evaluate the prediction results of logistic regression model. Through the following chart, we can find that the ROC curve is convex to the upper left, and the AUC value is 0.874, indicating that the prediction of this model is good.

### 6. Strategy Formulation

1. Customer retention: For the important retention travelers, airlines need to provide priority resources and personalized services for such high-value passengers according to the value of different passenger groups.

2. Customer Churn Alarm: For those passengers who need to be retained or maintained, the airlines should observe the changes of L, R, F, M, and C of such passengers from time to time to calculate the transaction status of passengers and make key contacts based on the possibility of passenger loss. Visit and take the most effective way to prevent the loss of important passengers.

3. Customer development: For travelers who need important development, airlines should strengthen their communication with such passengers so that they can better understand the airline's membership services and various member information, promotional information, etc. Due to the immeasurable potential losses caused by customer churn, if airlines can better use the passenger information they have, build a passenger database and strengthen the analysis of passenger loss. This will have far-reaching implications for the airline's profit growth and the company's personalized development.

### Acknowledgements

This paper was financially supported by National College Students Innovation and Entrepreneurship Training Program of China—project number: 201810294084).
References
[1] Wenxin Zhang, Study on Airline Frequent Passenger Segmentation [D]. Nanjing University of Aeronautics and Astronautics, 2009
[2] Tongshui Wu, Liang He, Analysis of Airline Passenger Loss Based on Decision Tree [J]. Market Modernization, 2006
[3] Xiangcui Wang, Analysis of Passenger Loss in Commercial Retail Industry Based on Analytic Hierarchy Process [J]. Market Forum, 2008
[4] Yaqi Cui, Analysis of Air Passenger Loss Based on C5.0 Algorithm [J]. Journal of XIAN Aeronautical University, 2018
[5] Liangsheng Luo, Wenxin Zhang, Research on airline passenger segmentation method based on frequent flyer database [J]. Modern Business, 2008(23)