Analysis of Passengers' Tickets Pre-booking Behavior in High-speed Railway Based on Market Segmentation

Yantao Gong
School of Traffic and Transportation
Beijing Jiaotong University
Beijing, 100044, China
E-mail: 17120801@bjtu.edu.cn

Lei Nie*
School of Traffic and Transportation
Beijing Jiaotong University
Beijing, 100044, China
E-mail: lnie@bjtu.edu.cn
(*Corresponding author)

Huiling Fu, Zhenhuan He
School of Traffic and Transportation
Beijing Jiaotong University
Beijing, 100044, China
E-mail: hlfu@bjtu.edu.cn,
zhhe@bjtu.edu.cn

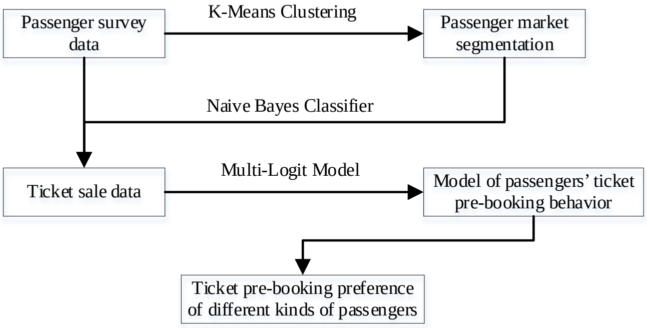
***Abstract*—Chinese high-speed railway has gradually formed into a big network, and passengers pay more attention to the quality of the transport services provided by railway operation departments, so accurately obtaining passengers' tickets pre-booking behavior characteristics becomes the key to improve high-speed transport services. Since there is no research on classifying and describing different types of passengers using both passenger survey data and ticket data in the current study. In this paper, we pay more attention to the use of 2 kinds of data. Firstly, market segmentation of high-speed rail passengers is proposed, which is based on K-Means Cluster analysis and using passenger survey data. Then, this paper draws lessons from Naive Bayes Classifier, the ticket data are classified according to the result of market segmentation. And a passengers' tickets pre-booking behavior model based on Multi-Logit Model is established. Finally, through the analysis of specific parameters, the tickets pre-booking preferences of four different types of passengers (family tourism market, personal visiting market, student market and business market) are classified and described. Using the model, passenger flow forecasting can be realized for different trains. The results can be used for designing high-speed railway products.**

Keywords—High-speed Railway; Market Segmentation; K-Means Cluster; Passenger Choice Behavior; Logit Model

I. INTRODUCTION

With the development of social economy, Chinese high-speed railway has gradually formed into a big network, in the case of travel demand is basically met, passengers pay more attention to the quality of high-speed railway passenger service. Therefore, improving the quality of high-speed rail passenger transport products has become the top priority for the further development of high-speed rail, and the key is to accurately obtain the characteristics of passengers' tickets pre-booking behavior.

For railway enterprises, ticket data is the high-quality basic data for passenger flow analysis and research, which is widely used in various kinds of research. However, the ticket data only contains the relevant attributes of passenger transport products (i.e. trains), and could not reflect the relevant characteristics of the passengers who buy this ticket, such as gender, occupation, travel purpose, etc. If we only rely on the ticket data to study the passenger's ticket purchase choice behavior, we can't classify and describe the different types of passengers' tickets pre-booking behavior, which is not conducive to the railway enterprise to develop the corresponding ticket selling policy to meet the different travel needs of all kinds of passengers. At present, relevant researchers have conducted a lot of research on high-speed railway passenger travel behavior. Based on the passenger survey data, T. Chen et al. have constructed a double-layer nested logic decision-making model for passenger departure time and transportation mode selection, and calculated the model parameters using the maximum likelihood method[1]; M. Su et al. have established a binary logit model for passenger travel selection based on the market survey data Based on the SP survey data[2]; S. Wang et al. established the logit model of high-speed railway passenger choice behavior, and verified the rationality of the model[3]; L. Qiang used the ticket data to carry out regression analysis of passenger travel behavior, and demarcated some typical OD's passenger travel choice parameters[4]; van Ryzin et al. established the discrete choice model and used the maximum likelihood estimation method to describe the customer choice behavior[5].



Passenger survey data
K-Means Clustering
Passenger market segmentation
Naive Bayes Classifier
Ticket sale data
Multi-Logit Model
Model of passengers' ticket pre-booking behavior
Ticket pre-booking preference of different kinds of passengers

Fig.1. Research Map

Most of the existing studies only use the passenger survey data or ticket data to study the whole passenger choice behavior, but have not yet achieved the classification description of different types of passengers' tickets pre-booking characteristics. To solve this problem, this paper first uses K-means Cluster analysis to segment the high-speed railway passenger market based on the passenger survey data. And the ticket data are classified by using Naive Bayes Classifier according to the result of market segmentation. Then, the passengers' tickets pre-booking behavior model is established, which is based on Multi-Logit

Model, to achieve the classification of different kinds of passengers' tickets pre-booking behavior. Finally, the validity of the model is verified by an example. The specific research ideas are shown in Fig. 1.

II. Market segmentation of the high-speed railway passenger

Market segmentation is a process that enterprises divide the whole market into different consumer groups according to the different needs of consumers. Effective market segmentation of high-speed railway passengers can help railway enterprises to find their own target market. According to the significant characteristics of the target market, further detailed research can be carried out to study the different tickets pre-booking behavior characteristics of different types of passengers, reasonably design passenger transport products, and effectively allocate various resources, to further improve enterprise's income under the premise that market demand is met.

A. Data sources

In July 2016, a questionnaire survey was conducted on passengers for a high-speed railway with a length of about 500km in Guangdong and Guangxi province, mainly including 5 working days and 3 rest days. In the survey, the method of random sampling was used to randomly select passengers in the stations and trains for questionnaire survey, and 4181 questionnaires were finally recovered. In this study, the passenger survey data of the same OD are selected for research, and 10889 ticket data corresponding to the OD and the date are selected as the data basis of Chapter II and III.

B. Variable of market segmentation

Because the ticket data can only reveal all kinds of information contained in each ticket, and cannot show the personal attribute information of the passengers who choose this ticket, so in the market segmentation of passenger groups, we need to start with the passenger survey data, and select the information that can reveal the passenger information for market segmentation.

TABLE I. Value of Explanatory Variable in Market Segmentation

Variable	Variable name	Description
X_1	Gender	1 = "male", 2 = "female"
X_2	Age	1 = "18 below", 2 = "18~25", 3 = "26~35", 4 = "36~49", 5 = "50~59", 6 = "60 above"
X_3	Occupation	1 = "government officials", 2 = "migrant workers",3 = "students", 4 = "freelance",5 = "enterprise manager", 6 = "enterprise staff",7 = "others"
X_4	Annual travel times	1 = "1 time",2 = "2~3 times", 3 = "4~5 times",4 = "6~10 times", 5 = "10 times above"
X_5	Monthly income	1 = "3000 below",2 = "3000~5000", 3 = "5000~7000",4 = "7000~10000", 5 = "10000~15000",6 = "15000 above"
X_6	Travel purposes	1 = "business",2 = "study",3 = "tourism", 4 = "visiting",5 = "migrant", 6 = "commuting",7 = "other"
X_7	Source of the fee	1 = "public",2 = "self-paid"
X_8	Number of peers	1 = "1",2 = "2",3 = "3", 4 = "4",5 = "5",6 = "6 above"

The passenger survey data obtained from the questionnaire mainly includes the passenger's personal attribute information (such as gender, age, occupation, income, annual travel times, etc.), specific travel information (such as travel purpose, source of the fee, number of peers, specific ticket information, etc.), and tickets pre-booking information.

At present, it is necessary to cluster the data collected from the passenger market survey to get the classification of sample groups, that is, different sub markets. Therefore, the choice of segmentation variables should pay more attention to the personal attribute information and specific travel information of passengers. And different sub markets are determined through these two aspects.

Therefore, the variables for the market segmentation in this paper are shown in Tab. I.

C. Market segmentation based on K-Means Clustering

K-Means Cluster is a common method of classification and statistics. Firstly, according to a certain method, a group of clustering centers are selected to make the samples agglomerate to the nearest center, so that the initial classification is formed. Then, according to the principle of the nearest distance, unreasonable classification is modified until the clustering is reasonable. Since K-Means Cluster analysis method has the characteristics of high efficiency, accuracy and better controllability, and allows the number of clusters to be specified in advance, when the number of clusters is known, using this clustering method can get the clustering results faster. Therefore, this paper chooses this method to analyze the passenger survey data and realize the segmentation of high-speed railway passenger market.

1) Sample clustering

In SPSS software, according to the basic steps of K-Means Cluster, firstly, the common factors are extracted from the original variables according to the Principal Component Analysis, and the expression of the original variables is realized by using the common factors whose eigenvalues are greater than 1. Then, taking the extracted common factors as input, using K-Means Clustering function in SPSS software, the market segmentation of passenger samples is realized. In the clustering analysis, the sample classification is set to 4, the initial cluster center is automatically selected by the system, the maximum number of iterations is set to 100, and the convergence parameter is set to 0.00. When the number of iterations reaches 100, the distance of each cluster center is far less than the distance of the initial cluster center, the iterative operation ends. After variance verification, we can get the result of sample clustering, that is, the result of passenger market segmentation.

2) The results of market segmentation

Through the sample information statistics of each sub market obtained by clustering samples, the passenger characteristic distribution of the sub market is obtained as shown in Tab. II.

It can be seen from Tab. II that when passengers are divided into four sub markets:

Market (1) is mainly for the relatively old middle-aged and elderly groups, with relatively average income, low annual travel times, mainly for government officials and freelancers, mainly for tourism, with many peers, which is in line with the characteristics of general middle-aged and elderly families, and can be defined as family tourism market.

Market (2) is mainly composed of middle-aged people with relatively high income and high annual travel times. It is mainly composed of enterprise staff, freelancers and some enterprise managers. The travel purpose is mainly for tourism and leisure, visiting relatives and friends. The number of peers is relatively small, which is in line with the characteristics of middle-aged people who love traveling and can be defined as personal visiting market.

Market (3) is mainly for younger people, with relatively low income, few annual travel times, mainly for students. The purpose of travel is mainly for study, visiting relatives and friends, tourism and leisure. The number of peers is small, which is in line with the characteristics of summer travel of general college students. It can be defined as student market.

Market (4) is mainly for the middle-aged people who are rich and powerful. Their income is relatively high, and their annual travel times are also high. They are mainly public servants and enterprise workers. Their travel purpose is focused on official business. The number of peers is small, which is in line with the travel characteristics of ordinary business people. It can be defined as business market.

TABLE II. PASSENGER MARKET CHARACTERISTICS DISTRIBUTION TABLE

Sub market	Gender	Age	Occupation	Annual travel times	Monthly income	Travel purposes	Source of the fee	Number of peers	Proportion
(1)Family tourism market	45% (male)	66% (36-59), 39.5 (average)	37.3% (government officials), 33.6%(freelance)	4.7 (average)	6600 (average)	74.5%(tourism), 14.5%(visiting)	98.2% (self-paid)	4.6 (average)	24.9%
(2)Personal visiting market	63% (male)	67% (18-35), 33.5 (average)	32.6%(freelance), 23.0% (enterprise staff), 21.5%(others)	7.6 (average)	7800 (average)	23.7%(tourism), 27.4%(visiting), 27.4%(others)	97.0% (self-paid)	1.5 (average)	30.5%
(3)Student market	36% (male)	81% (25 below), 21.5 (average)	58.3%(students), 20.5%(freelance)	4.0 (average)	3200 (average)	9.5%(study), 66.1%(tourism), 18.9%(visiting)	98.4% (self-paid)	3.0 (average)	28.7%
(4)Business market	81% (male)	79% (26-49), 31.9 (average)	14.3% (government officials), 20% (enterprise manager), 40% (enterprise staff)	9.1 (average)	8300 (average)	78.6%(business)	88.6% (public)	2.0 (average)	15.8%

III. MODEL OF PASSENGERS' TICKET PRE-BOOKING BEHAVIOR BASED ON MARKET SEGMENTATION

A. Ticket data classification

To study the tickets pre-booking behavior of different types of passengers, we need ticket data in different markets. Since there is no passenger's personal attribute information in ticket data, it is necessary to introduce a new data classification method to realize the market segmentation of ticket data.

In the field of machine learning, Naive Bayes Classifier is a very common data classification method. Its principle is to use the overlapping information between two events to determine the probability of unknown events from the probability of occurrence of known events. Therefore, according to the principle of Naive Bayes Classifier and the passenger market segmentation determined in Chapter II, the probability of the existence of each ticket information in the diverse sub markets can be obtained by using the overlap between the passenger survey data and the ticket data. Hence, the passenger ticket information in each sub market can be determined accordingly. As shown in Fig. 2, by comparing the passenger survey data with the ticket data, it can be found that there is overlapping information between the two kinds of data in arrival and departure time, seat type and type of ticket. According to this information, we can calculate the different probability that each ticket data belongs to different sub markets, and realize the market segmentation of ticket data by means of random coverage.

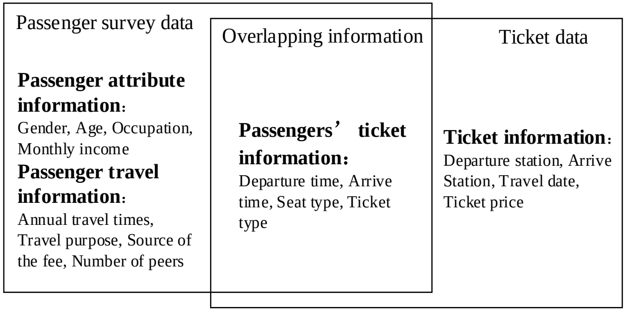

Passenger survey data
Passenger attribute information:
Gender, Age, Occupation, Monthly income
Passenger travel information:
Annual travel times, Travel purpose, Source of the fee, Number of peers
Overlapping information
Passengers' ticket information:
Departure time, Arrive time, Seat type, Ticket type
Ticket data
Ticket information:
Departure station, Arrive Station, Travel date, Ticket price

Fig.2. Data information coupling diagram

B. Mathematical formulations

The disaggregate model has the characteristics of strong logicality, good time and regional transfer, high precision of description and wide range of related factors and variables. It can use the behavioral survey (RP) data and intention survey (SP) data to quickly model. Multi-Logit Model is the most commonly used disaggregate choice model to study passenger travel choice. Since Logit model has strong applicability, simple structure and more mature application, when the hypothesis is subject to the double exponential distribution, using Multi-Logit Model in the study of passengers' pre-booking behavior, the preferences of passengers for observable factors (such as arrival and departure time, ticket price, seat type and other ticket information) can be well fitted. Therefore, this paper chooses Multi-Logit Model as the basis to build a passengers' tickets pre-booking behavior model.

1) The building of the utility function

When studying the problem of travel choice, the disaggregate model uses the concept of utility in economics for reference, regards that the travel choice behavior of travelers is similar to the consumption behavior of consumers, and applies the utility theory to the problem of travel choice. The Stochastic Utility Theory assumes that the decision-maker chooses according to the maximum utility generated by the option, and gives a relationship from the characteristic variable of the option to the decision-maker's choice result. Therefore, according to the Stochastic Utility Theory, the choice of train ticket by passengers can be interpreted as: when passenger n chooses to travel, he will evaluate ticket j from alternative set C , thus generating utility $U_{n,j}$.

From the view of the researchers, because of the inexhaustible influence factors of passenger choice, the choice of tickets is also uncertain. Set the observable factors that affect the choice of passengers as the characteristic variable $\boldsymbol{X}_j$, $\boldsymbol{X}_j = (x_1, x_2, x_3 ... x_k)$, and the unobservable factors as the random variable $\varepsilon_{n,j}$.

The utility value of the ticket can be expressed as $U_{n,j}$, $U_{n,j} = U(x,\varepsilon)$. The probability of passengers choosing tickets can be described as $P_n(j \mid x)$. The premise of the application of the discrete choice model is usually to assume that the decision-maker chooses according to the utility maximization theory, that is, the utility value $U_{n,j}$ will be generated when the passenger evaluates the ticket j . The passenger selects the ticket with the most utility from the alternative set with a certain probability, as shown in (1).

$$P_{n,j} = prob(U_{n,j} > U_{n,i}; j \neq i, i, j \in C)$$

(1)

And, the fixed utility function term is composed of the influencing factors that can be observed by researchers, and the random term is composed of the influencing factors that can affect the utility of passengers but cannot be observed. The establishment of discrete selection model focuses on the determination of fixed utility function $V(x)$ and the distribution form of random variable ε .

Therefore, according to the Stochastic Utility Theory, the fixed utility function $V(x)$ is composed of the characteristic variable $\boldsymbol{X}$ and the passenger's choice preference vector $\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3 ... \beta_k)$. If the fixed utility function $V(x)$ is linear with the characteristic variable vector $\boldsymbol{X}$, the fixed utility function of passenger to ticket j is:

$$V_j(x) = \boldsymbol{\beta X}_j^T + \beta_0 \quad (2)$$

To express conveniently, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2 ... \beta_k)$ is used to express passengers' travel preference. $\boldsymbol{X}_j = (x_1, x_2, x_3 ... x_k)$ is the characteristic variable that affects the selection of passengers. In this case, the fixed utility function can be expressed as:

$$V_j(x) = \boldsymbol{\beta X}_j^T$$

(3)

2) Modelling of passengers' tickets pre-booking behavior based on Multi-Logit Model

As mentioned before, this paper chooses Multi-Logit Model as the basis to build the model of passengers' tickets pre-booking behavior. If random variables obey independent and identical double exponential distribution, the probability of being selected for any ticket $j \in C$ in the pre-booking period is:

$$P_j = \frac{e^{V_j(x)}}{\sum_{i=1}^{I} e^{V_j(x)} + 1} = \frac{e^{\boldsymbol{\beta X}_j^T}}{\sum_{i=1}^{I} e^{\boldsymbol{\beta X}_i^T} + 1}$$

(4)

3) Characteristics variable

The passenger's tickets pre-booking preference is based on the difference between different options. This difference is reflected in the difference of utility brought to the passengers first, and then because of the difference of characteristics between different options, which makes the passenger have preference in comparison. Finally, this difference needs to be represented by excessive value. On the other hand, the appropriate value of characteristic variable is the guarantee of accurate parameter estimation.

TABLE III. VALUE OF CHARACTERISTICS VARIABLE

Variable	Variable name	Variable specification	Value
x_1[a]	Departure time	[7:00,10:00), [10:00,13:00) [13:00,16:00), [16:00,19:00]	Binary decision variable
x_2[a]	Arrive time	[11:00,14:00), [14:00,17:00), [17:00,20:00), [20:00,23:00]	Binary decision variable
x_3	Travel time	Actual value of travel time	Normalization: (0,1]
x_4	Price	Actual value of ticket price	Normalization: (0,1]
x_5	Seat type	1 for second-class seat and no seat, 2 for first-class seat, 3 for business seat	Normalization: (0,1]
x_6	Originating train	1 for originating train, 0 for through train	Binary decision variable

[a] *Variables x_1 and x_2 respectively representing the departure time and arrive time are divided into four binary decision variables in the model. $x_{11}, \ldots, x_{14}$ and $x_{21}, \ldots, x_{24}$ respectively indicate which specific period the departure time and arrive time belongs to.*

In addition, the source of data also determines the selection of characteristic variables. Because the ticket data has the advantages of large amount of data and reliable information, the ticket data becomes the best choice for parameter estimation of the model. Therefore, the characteristic variables should also be selected from the information contained in the ticket data. The value selection method of characteristic variable is shown in Tab. III.

C. Algorithm and results

1) Algorithm

The natural logarithm of (4) can be obtained as follows:

$$\ln(P_j) = \ln(\frac{e^{\beta X_j^T}}{\sum_{i=1}^{I} e^{\beta X_i^T} + 1}) = \ln(e^{\beta X_j^T}) - \ln(\sum_{i=1}^{I} e^{\beta X_i^T} + 1)$$
$$= e^{\beta X_j^T} - \ln(\sum_{i=1}^{I} e^{\beta X_i^T} + 1)$$

(5)

For any two tickets, j_1 and j_2, there are:

$$\ln(P_{j_1})\text{-}\ln(P_{i_2}) = [\beta X_{j_1}^T - \ln(\sum_{i=1}^{I} e^{\beta X_i^T} + 1)]$$
$$-[\beta X_{j_2}^T - \ln(\sum_{i=1}^{I} e^{\beta X_i^T} + 1)]$$
$$= V_{j_1}(\mathrm{x}) - V_{j_2}(\mathrm{x})$$

(6)

Therefore, if the selected probability of each ticket can be determined, the utility of each ticket can be determined. As the actual utility of each ticket is known, and the relationship between utility and characteristic variables is linear, the solution of the passengers' tickets pre-booking preference vector $\boldsymbol{\beta}$ can be transformed into the solution of the parameter according to the known utility value by using the linear regression method.

In the actual ticket data, the types of tickets are mainly ordinary tickets and seat types are mainly second-class seats. Therefore, this paper assumes that the utility of each train is the utility of all kinds of tickets contained in the train. So, if the transportation capacity of high-speed railway is sufficient, the selection probability of passengers for each train can be determined by the ratio of the tickets selling volume S_i of each train under the same OD to the total ticket selling volume under the corresponding OD, so there are:

$$P_j = \frac{S_j}{\sum_{i=1}^{I} S_i}$$

(7)

Thus, the utility of each ticket can be expressed as:

$$V_{j_1}(\mathrm{x}) - V_{j_2}(\mathrm{x}) = \ln(\frac{S_{j_1}}{\sum_{i=1}^{I} S_i}) - \ln(\frac{S_{j_2}}{\sum_{i=1}^{I} S_i})$$

(8)

Therefore, the utility of each ticket can be determined through ticket data.

2) Results

When the basic variables and specific utility are determined, it is necessary to estimate the parameters of the variables, and determine the degree of their impact on the selection probability. If the degree of impact is small, the variables would be removed, and the variables with a large degree of impact can be retained to build the model. The statistical analysis software SPSS is used to estimate the model parameters. Variable selection is determined according to the significance level α of Chi-square statistics after each variable is added. If α > 0.05, it means that the variable has no impact on the selection results and can be removed; otherwise, it should be retained. The parameter calibration results of submarket I is shown in Tab. IV.

TABLE IV. THE PARAMETER CALIBRATION RESULTS OF SELECT-MODEL IN MARKET(1)

Model	**Non-standardized coefficient**		**Standardization coefficient**	**t**	**Significance level α**
	B	**Standard error**	**Beta**		
constant	3.699	0.243		15.241	0.000
x_{12} [a]	1.015	0.130	0.490	7.812	0.000
x_{13} [a]	0.815	0.117	0.443	6.955	0.000
x_{22} [a]	0.610	0.128	-0.327	-4.759	0.000
x_{23} [a]	0.610	0.115	-0.308	-5.303	0.000
x_{24} [a]	-0.183	0.039	-0.076	-4.634	0.000
x_3	-6.213	0.225	-0.455	-27.624	0.000
x_6	2.154	0.084	0.398	25.687	0.000
x_5	-0.317	0.086	-0.060	-3.680	0.000
x_4	0.274	0.301	0.050	3.049	0.002

[a.] *x_{12} indicates that the departure time of the train belongs to the time period [10:00,13:00), x_{13} indicates that the departure time of the train belongs to the time period [13:00,16:00), x_{22} indicates that the departure time of the train belongs to the time period [14:00,17:00), x_{23} indicates that the departure time of the train belongs to the time period [17:00,20:00), x_{24} indicates that the departure time of the train belongs to the time period [20:00,23:00].*

The model in Tab. IV is the final equation of the Market (1) passenger's pre-booking ticket behavior. In the process of modeling, the regression coefficients of x_{11}, x_{21} are not significant, so they are eliminated. In the final equation, the significance level α =0.05, because the probability P of the significance test of the regression equation is less than the significance level α, the linear relationship between the utility value and the characteristic variable is significant, and the established linear model is appropriate.

It can be determined from Tab. IV that the random utility function of ticket in Market (1) is:

$$V_j^{\mathrm{I}}(x) = \boldsymbol{\beta X}_j^T$$
$$= 3.699 + 1.015x_{12} + 0.815x_{13} - 0.610x_{22} - 0.610x_{23}$$
$$- 0.183x_{24} - 6.213x_3 + 0.274x_4 - 0.317x_5 + 2.154x_6$$

(9)

Therefore, in Market (1), the behavior model of passenger's pre-booking behavior is as follows:

$$P_j^{\mathrm{I}} = \frac{e^{V_j^{\mathrm{I}}(x)}}{\sum_{i=1}^{I} e^{V_i^{\mathrm{I}}(x)} + 1} = \frac{e^{\substack{3.699+1.015x_{12}+0.815x_{13}-0.610x_{22}-0.610x_{23} \\ -0.183x_{24}-6.213x_3+0.274x_4-0.317x_5+2.154x_6}}}{\sum_{i=1}^{I} e^{\substack{3.699+1.015x_{12}+0.815x_{13}-0.610x_{22}-0.610x_{23} \\ -0.183x_{24}-6.213x_3+0.274x_4-0.317x_5+2.154x_6}} + 1} \tag{10}$$

And, we can use the same method to determine the model of Market (2), (3) and (4). Specific parameters are shown in Tab. V.

TABLE V. PASSENGER SELECTION PARAMETERS IN SUB MARKETS

Sub market	β_0	β_{11}	β_{12}	β_{13}	β_{14}	β_{21}	β_{22}	β_{23}	β_{24}	β_3	β_4	β_5	β_6
(1)	3.699	0.000	1.015	0.815	0.000	0.000	0.610	0.610	-0.183	-6.213	0.274	-0.317	2.154
(2)	3.820	-0.269	0.240	0.000	-0.839	0.000	0.000	0.000	0.422	-6.137	0.000	-0.097	2.238
(3)	4.220	-0.839	0.000	-0.241	-1.089	0.000	0.392	0.405	0.000	-5.923	-0.918	0.000	2.176
(4)	3.382	0.793	0.157	0.000	0.892	0.000	-0.474	-0.575	0.000	-5.044	0.565	0.820	2.235

D. Parameter analysis

The following conclusions can be obtained by analyzing the ticket pre-booking behavior parameters of four sub markets, as shown in Tab. VI.

TABLE VI. DESCRIPTION OF PASSENGERS' CHOICE BEHAVIOR IN SUB MARKETS

Sub market	**Time of departure and arrive**		**Others**
	Time of departure	**Time of arrive**	
(1) Family tourism market	Passengers are more inclined to choose trains departing at [10:00~13:00] and [13:00~16:00].	Passengers are more inclined to choose trains arriving at [10:00~13:00] and [13:00~16:00].	Passengers tend to have shorter travel time and more originating trains. However, the two factors of ticket price and seat type have no significant effect on the choice behavior of passengers.
(2) Personal visiting market	Passengers are less inclined to choose trains departing from [16:00-19:00].	Passengers are less inclined to choose trains arriving at [14:00-17:00] and [17:00-20:00].	Passengers tend to have shorter travel time and more originating trains, don't care much about the ticket price, but prefer the more comfortable seats.
(3) Student market	Passengers are less inclined to choose trains departing at [7:00-10:00] and [16:00-19:00].	Passengers are more inclined to choose trains arriving at [14:00-17:00] and [17:00-20:00].	Passengers tend to have shorter travel time, more originating trains and lower fares, even don't care much about the seats.
(4) Business market	Passengers are more inclined to choose trains departing at [7:00-10:00] and [16:00-19:00].	Passengers are less inclined to choose trains arriving at [14:00-17:00] and [17:00-20:00].	Passengers tend to have shorter travel time and more originating trains, don't care much about ticket prices, but prefer more comfortable seats.

Comparing the different preferences of passengers in the four sub markets horizontally, the comparison shows that although there are obvious different needs of passengers in each sub market, the commonalities among the sub markets are also obvious, for example, overall, passengers pay more attention to travel time and preference for originating trains, etc.

IV. MODEL VALIDATION

Based on the ticket data used for model calibration in the previous chapter, according to the ticket data's submarket information determined in Section A, Chapter III, after deleting the train number and extracting the relevant variables, we could import the ticket data to the corresponding sub market model to calculate the train selection of passengers. Taking the consistency of passengers' selection to evaluate the accuracy of model, the results are shown in Tab. VII (the actual value is the actual number of tickets for four sub markets classified by Naive Bayes Classifier based on the result of market segmentation in Chapter II, and the predicted value represents the number of passengers in the corresponding sub market whose train selection predicted by the model is consistent with the actual train selection). In the results, except that the prediction accuracy of the Market (1) due to the small sample size is slightly lower, the prediction accuracy of the model is more than 76%, which verifies the accuracy of the model, and indicates that the model built in Chapter III can better describe the pre-booking behavior preference of passengers.

TABLE VII. COMPARISONS BETWEEN MODEL PREDICTIONS AND ACTUAL RESULTS

Sub market	**The number of passengers who selected original trains**		**Accuracy rate**
	Actual value	**Predicted value**	
(1)	2419	1727	71.39%
(2)	2625	2028	77.26%
(3)	2970	2281	76.80%
(4)	2875	2263	78.71%
Total	10889	8352	76.70%

When the ticket data is imported into the model for calculation, it can not only verify the accuracy of the model, but also predict the selection of passenger for the trains. By comparing the predicted value with the actual value, part of the train information with large prediction difference can be obtained, as shown in Tab. VIII (the actual value represents the actual number of tickets for a train in the ticket data, and the predicted value represents the predicted number of tickets for the corresponding train by the model). Under the condition of sufficient tickets (all trains are available with sufficient tickets), the selection of some trains by passengers has been obviously shifted, in the prediction, the passenger flow of some trains favored by passengers has increased greatly, while the passenger flow of some unpopular trains have lost a lot.

TABLE VIII. TRAIN INFORMATION WITH GREAT DIFFERENCE IN PREDICTION

Train number	Actual value	Predicted value	Difference value
D21	705	852	147
D23	726	1089	363
D25	694	1030	336
D27	340	504	164
D31	611	787	176
D41	236	93	-143
D37	547	350	-197
D35	691	483	-208
D33	210	85	-125
D39	517	310	-207

Through the prediction of passengers' train selection, we can get the specific passenger flow prediction, which is very important for the line planning and the ticket selling strategy.

Such as train D25, the predicted passenger flow of this kind of train is far greater than the actual passenger flow. Some improvements can be adjusted in the stage of the line planning. The smaller capacity train is upgraded to the bigger capacity train, or the corresponding train is added in the same period, which can be taken to meet the travel demand of passengers. Therefore, based on the prediction of the passengers' train selection, the optimization of line planning can better meet the travel demand of passengers.

Such as train D35, the predicted passenger flow is less than the actual. The change of predicted passenger flow can be analyzed according to different submarkets, and the time of ticket sharing can be adjusted according to the different ticket pre-booking preference of passengers in each submarket. If the loss of family tourism flow is more in the predicted passenger flow, the ticket sharing time of the corresponding ticket collection can be modified in advance according to the characteristics that the ticket pre-booking time of this kind of passengers is generally earlier, which could meet more passengers' demand for tickets. If different kinds of passengers are all lost in the predicted passenger flow, the ticket collection of the corresponding OD can be reduced in the ticket allocation stage, which could meet the travel needs of the other ODs. Therefore, based on the prediction of passengers' train selection, the train revenue can be maximized through reasonable arrangement of ticket selling policy.

V. CONCLUSION

In this paper, four types of passengers are classified by analyzing their ticket pre-booking behavior based on K-Means Clustering analysis, Naive Bayes Classification and Multi-Logit Model modeling using passenger survey data and ticket data. The accuracy of the model in describing passengers' pre-booking behavior is verified by real-world ticket data. The research shows that: 1) Among various factors that may affect the passenger choice behaviors, these factors are proved to have the most significant impact: the arrival and departure time, travel time, ticket price, seat type and whether the train is the originating train. 2) By analyzing the parameters of the choice model, it can be concluded that passengers from different sub markets have common preferences in choice behavior as well as specific characteristics. 3) Also, further prediction on passenger flow on trains can be conducted based on the proposed passenger choice model. By comparing the predicted data and actual data, different adjustment methods can be designed, such as the adjustment of train formation in the line planning stage, the adjustment of ticket allocation and ticket sharing in the ticket selling stage, etc., which can provide scientific decision basis for the design and sale of the high-speed railway products.

ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China [Nos.2016YFE0201700, 2018YFB1201402].

REFERENCES

[1] T. Chen, B. Mao, L. Gao, and F. Yue, "Research about passenger travel choice behavior of dedicated passenger railway line," Journal of the China Railway Society, vol. 3, 2007, pp.8-12.

[2] M. Su, W. Luan, Y. Ma, and R. Zhang, "Passenger travel choice influencing factors on Beijing-Shanghai high-speed corridor," Railway Transport and Economy, vol. 41, no. 1, 2019, pp.58-63.

[3] S. Wang, and P. Zhao, "Analysis of passengers' choice behavior for dedicated passenger railway lines based on logit model," Journal of the China Railway Society, vol. 31, no. 3, 2009, pp.6-10.

[4] L. Qiang, "A research on passengers' choice behavior in high-speed railway based on ticket data," Railway Transport and Economy, vol. 40, no. 4, 2018, pp.52-57.

[5] van Ryzin, Garrett, and G. Vulcano, "A market discovery algorithm to estimate a general class of nonparametric choice models," Management Science, vol. 61, no. 2, 2015, pp. 281-300.

[6] K. Qiao, P. Zhao, and J. Wen, "Passenger market segmentation of high-speed railway based on latent class model," Journal of Transportation Systems Engineering and Information Technology, vol. 17, no. 2, 2017, pp.28-34.

[7] B. Qian, B. Shuai, and C. Chen, J. Li, "Study on subdivision of DPL passenger market based on mixed regression model," Railway Transport and Economy, vol. 36, no. 1, 2014, pp.60-65.

[8] F. Shi, L. Deng, and L. Huo, "Boarding choice behavior and its utility of railway passengers," China Railway Science, vol. 6, 2007, pp.117-121.

[9] R. Xu, and L. Nie, "The impacts of HSR operation parameters on sharing ratio of passenger flow and their sensitivity analysis," Railway Transport and Economy, vol.39, no. 11, 2017, pp.21-27.