Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Deep Learning-based Hotel Customer Behavior Prediction Model Construction and Management Strategy Optimization

Yajing Xi[1], Kun Liu[1,✉], Qiuhong Wang[1]

[1] *Caofeidian College of Technology, Tangshan, Hebei, 063200, China*

ABSTRACT

Accurately capturing the behavioral factors of different types of customer groups and adopting targeted service strategies is the key to business competition in the hotel industry. In this paper, we combine the variance Boston matrix and PSO-based K-means algorithm to achieve hotel customer attribute segmentation based on customer behavior, customer value and word-of-mouth reliability, and then use deep learning algorithms to construct a hotel customer behavior prediction model. The feature fusion layer and SENet are incorporated into the residual network in order to utilize the feature expression ability of different layers and the spatial coding ability between different channels to enhance the hotel customer behavior predictive ability. Downloading the public dataset from the online wine travel platform for example analysis, it is found that the classification of this paper's algorithm before customer segmentation has a correct rate of 83.75%, which is higher than the rest of the baseline models. After customer segmentation this paper's algorithm achieves the highest recall rate in all customer categories, and the recall rate is as high as 84% on category 1 customer groups, and the superiority of the designed algorithm is verified. This study facilitates hotel management to target customer service and retention according to different customer groups.

Keywords: boston matrix, PSO, K-means, SENet, hotel customer behavior

1. Introduction

The hotel industry is one of the fastest growing industries, and its competitiveness is increasing. In this competitive environment, hoteliers want to attract more customers, improve customer satisfaction, and establish stable customer loyalty, so as to increase the sales and profitability of the hotel [1-4]. Hotel customer behavior is an indicator closely related to customer relationship, which can help hoteliers predict customer consumption behavior and develop corresponding business strategies [5-8].

✉ Corresponding author.
E-mail address: lk12096@163.com (K. Liu).
Received 10 January 2024; Revised 15 February 2024; Accepted 25 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-494

Deep learning technology is one of the most widely used and effective techniques in the field of artificial intelligence. It adopts multi-layer neural network model for training, and realizes automatic analysis and prediction of data by continuously optimizing the weight parameters [9-11]. In the business field, deep learning technology is applied to marketing, customer behavior analysis, product recommendation and so on. Among them, customer behavior prediction is one of the important applications of deep learning technology [12-13]. By making full use of the Internet large-scale data and multi-source data, and combining deep learning technology for modeling and predictive analysis, it can help hotels better understand the changing law of customer behavior and provide scientific basis for related decision-making [14-16]. Customer behavior prediction is a key task in the process of hotel operation, through the analysis and prediction of customer consumption behavior, hotels can better develop marketing strategies and improve hotel performance [17-18].

Literature [19] uses random forest and logistic regression supervised machine learning algorithms based on frequency, RFM paradigm to predict hotel customer churn behavior. The results show that the random forest algorithm exhibits high prediction performance, and the research results have some reference value for hotel practitioners. Literature [20] aims to provide a predictive framework for customer churn through multiple stages so that churn can be accurately predicted and prevented. It is emphasized that the proposed model has a good performance in customer churn prediction, and it is pointed out that the enhancement model reflects more significant advantages in prediction. Literature [21] utilizes LSTMs to predict hotel occupancy rates based on online review texts. The results of review data collection and prediction comparison of multiple models show that. LSTMs outperforms other models in terms of prediction accuracy. It is emphasized that LSTMs are a feasible method to predict hotel occupancy. Literature [22] utilizes CRISP-DM technique and incorporates classification algorithms such as Random Forest. Based on the hotel booking dataset from systems such as C3Rewards, features such as merchantsaTM customer behavior, interests, etc. are investigated. Finally the accuracy of various algorithms in the experiment is presented. Literature [23] investigates customers' hotel revisit behavior based on customer review data. Customer data from online hotel bookings were analyzed and user feedback comments were sentiment analyzed. The results showed that there were differences in the content and sentiment of one-time visitors' and revisitors' feedback. Literature [24] utilizes machine learning techniques such as decision trees and analyzes the characteristics of customers in combination with their purchase records, and selects models with a high degree of promotion through promotion graphs in order to achieve precision marketing. The results show that decision trees have relatively better prediction effect with higher degree of promotion.

This paper is based on the Boston matrix theory, combined with the improved K-means algorithm based on customer value, customer behavior and word-of-mouth reliability to construct a hotel customer segmentation model of the Boston matrix. After pre-processing the data, the influencing factors are dimensionalized using principal component analysis. On this basis, the three first-level indicators of customer value, customer behavior and word-of-mouth reliability are combined with the Boston matrix to construct a three-dimensional variance Boston matrix, and the improved K-means algorithm with PSO is used to cluster the customer segmentation method, which integrates multi-level features and SENet to achieve accurate prediction of hotel customer behavior.

2. Method

2.1. Hotel Customer Segmentation Model Construction Based on Variational Boston Matrix

2.1.1. Three-dimensional stereo model construction based on the variational Boston matrix. The Boston Matrix is one of the most commonly used methods for product strategy combination, and this type of method is useful for the action and strategy selection of enterprises and the allocation of resources. In order to find out whether there are valuable customers that are worth recovering

strategically, valuable customers can be recovered by utilizing this analytical principle, treating customers as products of the Boston Matrix, and managing the customer segmentation portfolio. Customer value and customer behavior as the basis for measurement of two factors, so as to form the Boston matrix based on customer value and customer behavior as shown in Figure 1.

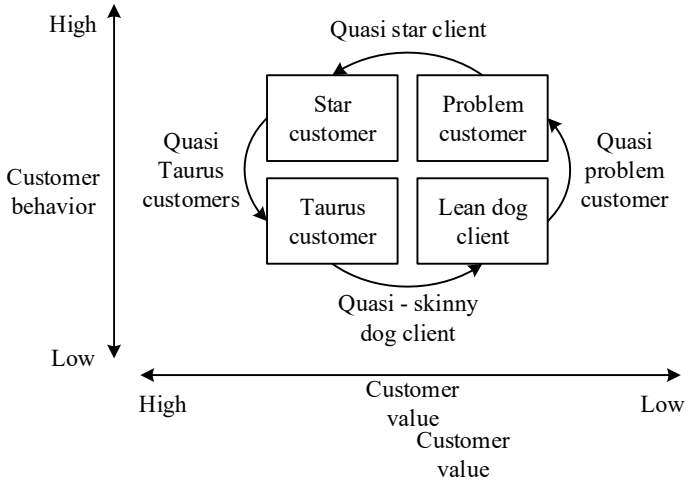

Fig. 1. Boston matrix based on capacity value and customer behavior

Under the orientation of word-of-mouth reliability, the Boston matrix based on customer value and customer behavior is converted into a three-dimensional three-dimensional model, and the three first-level indexes of customer value, customer behavior, and word-of-mouth reliability are used as three-dimensional axes to construct the three-dimensional variant Boston matrix in the variant Boston matrix [25]. On the basis of the original Boston matrix combined with Figure 1 to construct a three-dimensional variant Boston matrix, specifically shown in Figure 2.

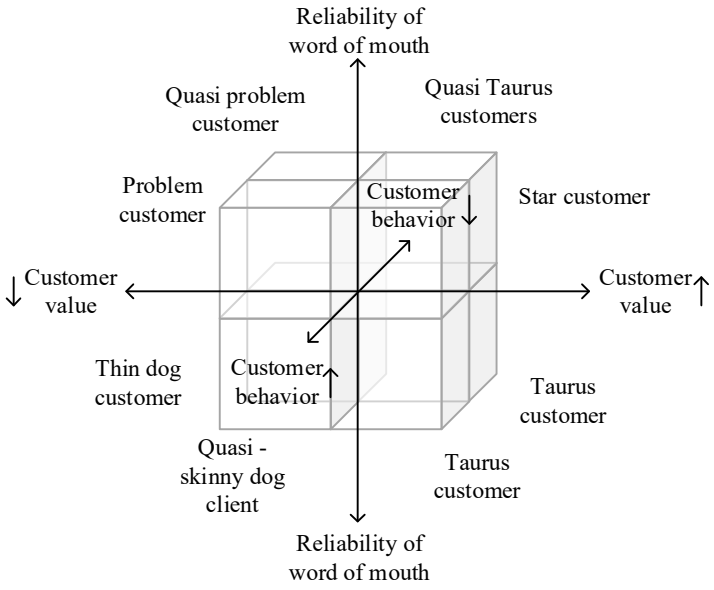

Fig. 2. 3D variation Boston matrix

2.1.2. K-means based hotel customer segmentation. K-Means clustering algorithm is one of the most widely used clustering algorithms currently, also known as K-means clustering algorithm.The basic idea of K-Means algorithm is: firstly, to find K clustering centers in the data set, and then divide the data set into K subsets, use the distance formula to find the distance from each subset to each

clustering center, and at the same time to make the sum of squares of distances from each subset to clustering center Minimize. The original dataset is classified by finding the optimal cluster center through a loop to get the optimal clustering. The distance formula used in this study is the Euclidean distance, and the formula is as follows:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2} \tag{1}$$

where k is the number of cluster centers of clusters, d is the number of variables contained in the data set, $d(x_i, x_j)$ represents the distance of similarity between two samples x_i and x_j, the larger the Euclidean distance between two samples, the more similar the two samples are; the smaller the Euclidean distance between two samples, the less similar the two samples are; the steps of K-Means are elaborated as follows:

Step 1: First determine a value of K, i.e., you wish to put the data set through the clustering algorithm to get K sets. Then K data points are arbitrarily selected from the feature space as the original clustering centers, called the centers of mass.

Step 2: On the dataset, the distance of each attribute point from each center of mass is calculated and each attribute value is assigned to the center of mass with the smallest distance, respectively.

Step 3: The data set corresponding to each center of mass is obtained, which is a total of K sets to complete the first round of clustering.

Step 4: After each round of clustering, the center of mass of each collection is recalculated and the clustering center is updated according to the set method. Afterwards, the clusters are re-clustered according to the updated center of mass, resulting in a new set of data, and the center of mass is updated again. This process is repeated until the updated center of mass position is not changed much and tends to be stable, then the clustering is completed, and finally a K categories with the latest center of mass as the center point are obtained.

2.2. Improved K-means algorithm for hotel customer segmentation

2.2.1. Principles of Particle Swarm Algorithm. Kennedy and Eberhart observed the behavioral patterns of bird flocks searching for food and proposed the particle swarm (PSO) algorithm. Each particle in the swarm searches for the solution of the problem by simulating the behavior of the bird flock searching for food, and each particle optimizes the direction of searching through its own searched optimal position (pbest) and other particles searched optimal position (gbest), i.e., updating its own speed and position towards the globally optimal solution through Eqs. (1) and (2), and continuously improves the fitness value of the objective function of each member. value to find the final objective through many selected generations. This algorithm is widely used because of its excellent performance in various optimization problems.

$$\begin{aligned} v_i(t+1) = v_i(t) + c_1 \times rand() \times (pbest_i - x_i(t)) \\ + c_2 \times rand() \times (gbest, -x_i(t) \end{aligned} \tag{2}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{3}$$

where: $i = 1, 2, \ldots, N$, N is the total number of particles v_i is the velocity of the particle, x_i is the current position of the particle; $rand()$ is a random number between (0, 1), and $c_1 c_2$ is the learning factor.

2.2.2. Optimized K-means algorithm. In this paper, based on the advantage of high computational efficiency of Kmeans algorithm, combining two methods to optimize the two defects mentioned above, an improved K-means clustering scheme is proposed. Firstly, the validity metric in the relative metric is used to cross-validate the K -value to select the optimal K -value [26]. Secondly, the PSO

algorithm is used to select the optimal initial center point of clustering, and then K-means is used for clustering to ensure the effectiveness of the clustering algorithm.

1) Effectiveness metrics for selecting the value of K. The validity indicator used in this paper to select the value of K is defined as follows: $SD(C) = a, xScat(C) + Dis(C)$. This is an aggregation indicator, a is a weight adjustment factor used to balance between the leveling dispersion of the clusters $Scat(C)$ and the overall separation between the clusters $Dis(C)$, and the dispersion and separation are defined as follows:

$$Scat(C) = \frac{1}{k}\sum_{c_k \in c} N_k \frac{|\sigma(C_k)|}{\sigma(C)} \tag{4}$$

$$Dis(C) = \frac{D_{\max}}{D_{\min}} \sum_{C_k \in C} \frac{1}{\sum_{c_1 \in c} d(\mu_k, \mu_1)} \tag{5}$$

where $\frac{D_{\max}}{D_{\min}}$ denotes the ratio of the maximum and minimum distances in C; $\sigma(C)$ denotes the standard deviation of the data items; $\sigma(C_k)$ denotes the standard deviation of cluster C_k; N_k denotes the number of samples divided into C_k clusters; and $d(\mu_k, \mu_1)$ denotes the Euclidean distance between μ_k, μ_1.

In the definition of the validity index, $Scat(C)$ indicates the degree of closeness between clusters, and its value is inversely proportional to the degree of closeness within clusters. $Dis(C)$ measures the degree of separation between clusters, which is affected by the clustering center of clusters and is proportional to the number of clusters; Let α take α as the starting point and 0.5 as the step change, calculate the value of cluster validity index of $K = 4.5.6.7.8$ according to the above validity index, and conclude that the average validity index of $\alpha = 0.5$ is the smallest.

2) Adaptation degree. In the improved algorithm, the PSO fitness is calculated using the basic idea of Kmeans division, where each sample is divided into initial step classes according to the nearest neighbor law, and then the fitness value is calculated for each sample under that division. The nearest neighbor law is defined as follows:

If X_i, j satisfies $\| X_j - Z_j \| = \min \| X_i - Z_k \|$, then X_i belongs to category j. The fitness function is as follows:

$$f(z_i) = \frac{1}{\sum_{i=1}^{k} \sum_{\forall x_i \in c_{ij}} d(x_i, c_{ij})} \tag{6}$$

Equation (5) is a commonly used fitness function in PSO algorithms that measures the sum of intra-cluster distances under this division. However, better clustering results need to take into account both smaller intra-cluster distances and larger inter-cluster distances, which makes the strategy of considering only intra-cluster distances insufficient. Therefore, in this paper, we choose the following fitness function proposed by Shachuf et al:

$$f(z_i) = w_1 d_{\min}(Z_i) + w_2 d_{\max}(z_i) \tag{7}$$

Among them:

$$d_{\min}(Z_i) = \max\left\{ \sum_{\forall x_i \in c_{ij}} \frac{d(x_i, c_{ij})}{|C_{ij}|} \right\} \tag{8}$$

$|C_{ij}|$ The sample size of the class C_{ij} being divided, $d_{\max}(Z_i)$ can be approximated to measure z_i the intra-cluster distance. $d_{\min}(Z_i) = \min\{d(c_{il}, c_{ip})\}(\forall l, p, l \neq p)$ measures the inter-cluster distance under that division. W_1 and W_2 are the weights that balance the importance of the two components, and when the appropriate weights are obtained, the minimum value of f balances the categorization scheme in satisfying the two criteria of small intraclass distances and large interclass distances.

2.3. Customer behavior prediction model based on IResNet

2.3.1. IResNet modeling ideas. Based on ResNet, this study proposes IResNet (IResNet), a customer behavior prediction model that incorporates multi-layer features and SENet, to enhance the recognition of customer behavior. In ResNet, the nonlinear expressive ability of the network improves with the superposition of layers, while its unique residual structure can effectively avoid the performance degradation problem caused by the deep network. However, it has been found that the ResNet model, after multiple convolutional operations, obtains advanced feature expression while losing some detail information, and it tends to use only the last layer of extracted features as the input to the classification function, and the loss of shallow features affects the prediction performance of the model. Therefore, in this paper, we use the principle of convolutional neural network in image processing, i.e., using high-level features and low-level features separately to obtain richer ennobling and detail information, and fusing the multi-layer features output from each residual block of ResNet to make full use of each layer will be characterized, so as to make the extracted feature maps express richer information, which can then be better mapped to the customer's behavior.

At the same time, in order to prevent redundant features from interfering with the prediction ability of the model, this study takes advantage of the idea of the attention mechanism of SENet to focus on the important information of the channels with high weights and filter the redundant channel information with low weights, respectively, so as to enable the model to achieve a better classification effect in terms of enhancing the spatial coding ability of the model.The working principle of SENet is to obtain the channel statistical information of the network through the global average pooling SENet works by obtaining the channel statistical information of the network through global average pooling, and using two nonlinear fully-connected layers to perform gate mechanism operations on the obtained channel statistical information in order to compute the degree of dependence of the model on each channel, so as to adjust each channel value according to the difference of the weight value. Where $\otimes$ denotes the element corresponding to the multiplication Reductionratio is the rate of descent, which aims to reduce the number of parameters [27].

The channel weighting formula is shown in equation (9):

$$W = \sigma\left(D_2\left(\delta\left(D_1(GAP(x))\right)\right)\right) \tag{9}$$

Finally, in this paper, the fused features are subjected to Global Average Pooling (GAP), i.e., the average of all pixels in each channel is calculated. The purpose is to reduce the model complexity while also effectively integrating the spatial information in order to further improve the prediction performance of the model.

In summary, the IResNet model proposed in this study has the following advantages: first, using ResNet as the base model can avoid network performance degradation while the depth is sufficiently large, and has stronger feature extraction capability, which in turn improves the model performance; second, by fusing the features of ResNet at each level, it can make the information expression richer, which in turn effectively reduces the customer behavior recognition Third, before feature fusion, the dependency relationship between channels is modeled by integrating the output vectors of each level of ResNet into SENet, which suppresses invalid channel features and enhances the expression ability of effective channel features; fourth, global average pooling is used at the end of the model, which can

effectively reduce the model parameters, reduce the amount of network computation and complexity, prevent the model from fitting, and thus enhance the generalization ability of the model.

2.3.2. IResNet model construction. Based on the above ideas, this study constructs the IResNet customer behavior prediction model to binary classify the borrower's default and advance prediction respectively, and the overall framework of this model is shown in Figure 3.

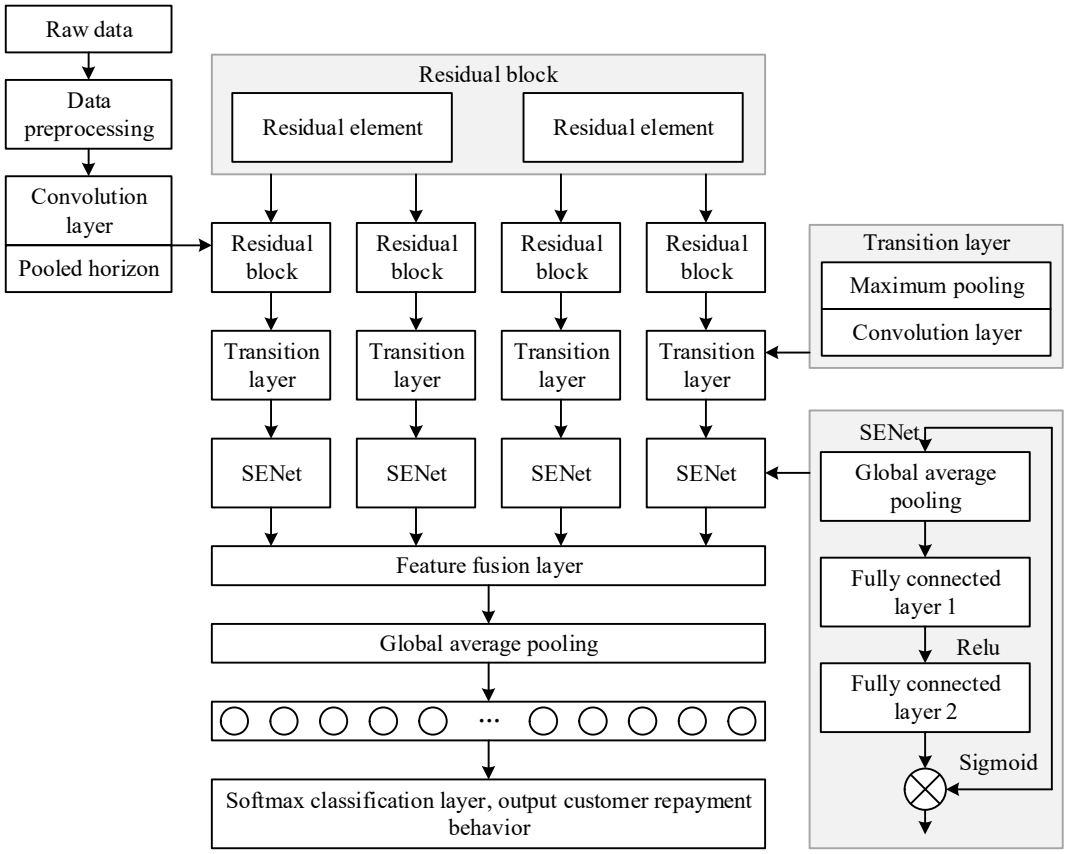

Fig. 3. Frame diagram of the model

When predicting whether a hotel guest will default, y can be expressed as:

$$y = \begin{cases} 0, & \text{Normal repayment} \\ 1, & \text{Default repayment} \end{cases} \tag{10}$$

When predicting whether advancement of hotel guests will occur, y can be expressed as:

$$y = \begin{cases} 0, & \text{Normal repayment} \\ 1, & \text{Repayment in advance} \end{cases} \tag{11}$$

First, the features of each layer are extracted on the basis of the original ResNet, and the output features after each residual block are input to the transition layer. The transition layer consists of a pooling layer and a convolution layer, which first performs maximum pooling to reduce the number of parameters, and then undergoes 1×1 convolution to change the number of channels and increase the feature expression ability; second, SENet is incorporated into the feature vector after mapping in the transition layer in order to obtain the weights of all the channels of each residual block, and then adaptively adjusts the feature response value of each channel; and then the output features of each residual block, which obtains the channel weights, are merged into the feature fusion layer, and the feature fusion layer indicates the pairwise vector splicing. Then, the output features of each residual

block with obtained channel weights are merged into the feature fusion layer, and the feature fusion layer indicates the pairwise vector splicing. At the same time, in order to reduce the number of parameters, the fused features are pooled with global average. Finally, the customer behavior is output using softmax classifier:

$$y = soft\max(GAP(o)) \tag{12}$$

$$o = o_1 \oplus o_2 \oplus o_3 \oplus o_4 \tag{13}$$

In Eq. (13), o_i is the output feature after the ind residual block, transition layer and attention mechanism processing, and the expression is shown in (14):

$$o_i = SE\left(C\left(P(x_i, 2^{4-i})\right)\right) \tag{14}$$

where $P(x_i, 2^{4-i})$ represents the maximum pooling operation with step $2^4 - i$ for the output feature map x_i of the ind residual block, C for the 1×1 convolution, and SE for the process of multiplying the individual channels by the weight values computed according to SENet.

3. Hotel customer behavior prediction results and analysis

3.1. Sample data sources and descriptions

All customer statistics in this study come from a public dataset containing user behavior data from the Ctrip platform for one week during July 2023, the data has been desensitized to protect customer privacy, but it does not affect the analysis of the issue. The dataset contains a total of 831054 data and 50 variables, and the specific meaning of each variable is shown in Table 1. By observing the meanings of all variables, it can be seen that label is the target variable in this dataset, which represents the customer's situation.

Table 1. Variable meaning

N	Variable field	Variable interpretation	N	Variable field	Variable interpretation
1	sampleid	Sample id	27	firstorder _ bu	First bu
2	label	Target variable	28	customereval _ pre2	The 24-hour history of hotel customers scoring mean
3	d	Date of visit	29	delta _ price2	The user preferences price -24 hours for the hotel average price
4	arrival	Date of check-in	30	commentnums _ pre	The number of hotel reviews is most visited in 24 hours
5	iforderpv _ 24h	Fill in the page in 24 hours	31	customer _ value _ profit	Customer value _ nearly 1 year
6	decisionhabit _ user	Decision habits: use users to observe decision-making habits	32	commentnums _ pre2	The 24-hour history of the hotel reviews the mean
7	historyvisit _ 7ordernum	Nearly 7 days of user history booking	33	cancelrate _ pre	The number of hotels has been accessed in 24 hours
8	historyvisit _ totalordernum	Nearly 1 year user history booking	34	novoters _ pre2	The 24-hour history of the hotel scores mean
9	hotelcr	Current hotel history cr	35	novoters _ pre	The number of hotel scores is the highest number of hotels in 24 hours
10	ordercanceledpr ecent	The user cancels the order rate within a year	36	ctrip _ profits	Customer value
11	landhalfhours	Time in 24 hours	37	deltaprice _ pre2 _ t1	Within 24 hours, the price of the hotel and the price of the difference,T+ 1
12	ordercannceledn um	The user cancels orders within a year	38	lowestprice _ pre	The lowest price is allowed in 24 hours
13	commentnums	Current hotel reviews	39	uv _ pre	24 hours of historical browsing
14	starprefer	Star preference	40	uv _ pre2	The 24-hour history of the hotel's historical uv mean
15	novoters	Current hotel scores	41	lowestprice _ pre2	The lowest mean of the hotel has been visited within 24 hours
16	consuming _ capacity	Consumption index	42	lasthtlordergap	The last time of the year
17	historyvisit _ avghotelnum	The number of visitors to the hotel in nearly three months	43	businessrate _ pre2	Within 24 hours, the business properties index mean of the hotel has been accessed
18	cancelrate	Current hotel history cancellation rate	44	cityuvs	Yesterday visited the current city with the application uv number of the check-in date
19	historyvisit _ visit _ detailpagenum	Visit the hotel details of the hotel in 7 days	45	cityorders	Yesterday submitted the application of the current city to the date of the check-in date
20	delta _ price1	Customer preferences price -24 hours for the maximum hotel price	46	lastpvgap	The last time we visited
21	price _ sensitive	Price index	47	cr	User conversion rate
22	hoteluv	Current hotel history uv	48	sid	Session id,sid= 1 can be considered a new visit
23	businessrate _ pre	24 hours of historical browsing number of hotel business properties index	49	visitnum _ oneyear	Annual access times
24	ordernum _ oneyear	User year order	50	h	Access time point
25	cr _ pre	The 24 hour history of the hotel history cr			
26	avgprice	Average price			

3.2. Description of indicators

An analysis of the metrics reveals that after removing the id and label, the remaining metrics can be roughly divided into two categories: customer behavior metrics, and hotel-related metrics to correspond to the reasons for customer behavior.

3.2.1. Customer behavior indicators. The Customer Behavior metrics include 26 variables that primarily reflect the hotel's historical customer spending. Some of these variables are derived from other metrics used by the hotel in its daily customer management to reduce the dimensionality of the dataset. For example, the spending power index is derived by analyzing the customer's historical order data, the single consumption amount and the total amount of the orders transacted compared to all customers of the same kind to give different weights to add up, and the same is also the price sensitivity index, the user conversion rate and so on.

3.2.2. Hotel-related indicators. Hotel-related indicators include 23 variables. It mainly focuses on hotels, through customers' ratings, number of visits, popularity value, etc. Hotel booking, as one of the main products of Ctrip.com, allows customers to score each hotel they have stayed at after the order is completed, mainly based on the quality of hotel services. The average of all the customers' scores on the hotels is averaged to get a rating average, generally speaking, the higher the rating average, the higher the customers' satisfaction with the hotel, and the customers will be inclined to book the hotel. Conversely, the lower the mean value of customer ratings for a hotel, the higher the probability that customers will book from that site.

3.2.3. Data pre-processing. Due to the large amount of data in this dataset and the presence of missing values in some of the data, as well as possible outliers in human data collection, preprocessing of the original dataset is essential. After careful observation, the following data preprocessing steps were roughly required in order to make the dataset meet the requirements of the study: filling of missing values, derivation of new variables, correction of outliers, and normalization. The processing of the dataset mainly relies on sklearn in python, and the data available for the constructed model is obtained after scientific and reasonable processing.

1) Statistical descriptive analysis of data. First of all, the original dataset is read into python software, the toolkit used is numpy and pandas, these two toolkits are the mainstream toolkit for the current data analysis process, with the advantages of fast analysis speed and beautiful output results. For example, to do descriptive statistics on all variables the results are shown in Table 2, observe the average value, extreme value, quartiles, etc., you can find that the data set exists in the character type of data, the data cleaning process needs to change the data type.

Table 2. Descriptive statistics of some variables

	sampleid	lowestprice_pre2	lasthtlordergap	businessrate_pre2	cityuvs	cityorders	lastpvgap
Nonnull value	703800	673877	456788	615019.2	695919.48	664288.26	604674
mean	641580	325	103868	0.3756	10.8612	2.2983	12290
Standard deviation	423300	359	125240	0.2243	16.0106	3.6092	26113
Minimum value	25092	1	0	0	0.0071	0.0071	0
25% quantile	318240	148	15299	0.1734	0.8435	0.1295	562
50% quantile	612000	238	47828	0.408	3.5975	0.6395	2905
75%	904740	396	141732	0.561	13.5935	2.8019	10941

Maximum value	2284800	44574	537567	1.0098	68.4828	14.7971	198274

2) Missing value filling and handling of abnormal values. Taking into account the date of visit (d), the date of stay (rival) is a character type type and has no practical meaning, is not suitable for subsequent modeling, it can be derived from the new variable advance booking = check-in time - visit time, so that after the transformation of the new variable has a strong information value.

A total of 6 columns in the dataset have negative values, in which there is no abnormality in the negative value of the average value of the price of the visited hotels and the price difference between the opponents within 24 hours, and the negative value of the remaining 5 columns should be an input error in data collection, respectively, for the value of the customer, the customer value of the last year, the user's preference for the price of the most hotels viewed in the last 1 year, the price of the hotel with the most users preferred price -24 hours browsing, the average price of the hotel with the user's preferred price -24 hours browsing, the lowest price of hotels can be booked for that year. After referring to the relevant knowledge in the field of hotel operation, customer value refers to the size of the profit contributed by the customer for the enterprise, so the negative value of customer_value_profit, ctrip_profits two variables in the dataset is replaced by 0, and delta_price1, delta_price2, lowestprice three hotel price related variables are treated as median.

3) Correlation test. By observing the names and meanings of the variables, it is easy to find that there should be multicollinearity between these characteristic variables, the correlation coefficients between the variables should be found out, and according to the size of the correlation coefficients, some of the variables that have a high degree of correlation should be excluded. The correlation coefficients of the data variables are shown in Fig. 4, due to the fact that too many variables will make the picture look crowded, so only part of the correlation coefficients of the variables are put out in the figure, in order to display the results more intuitively, the correlation coefficient graph is gradient processed, the darker the color in the graph that is, the bigger the correlation coefficient of the two variables is, and if the color is lighter, then the correlation coefficient between the variables is smaller as well. In statistics, if the Spearman's correlation coefficient between the variables is in the range of 0.8 to 1, then it can be said that the two variables are highly correlated. The highest correlation coefficient of the variables is 0.72, which is within the permissible range.

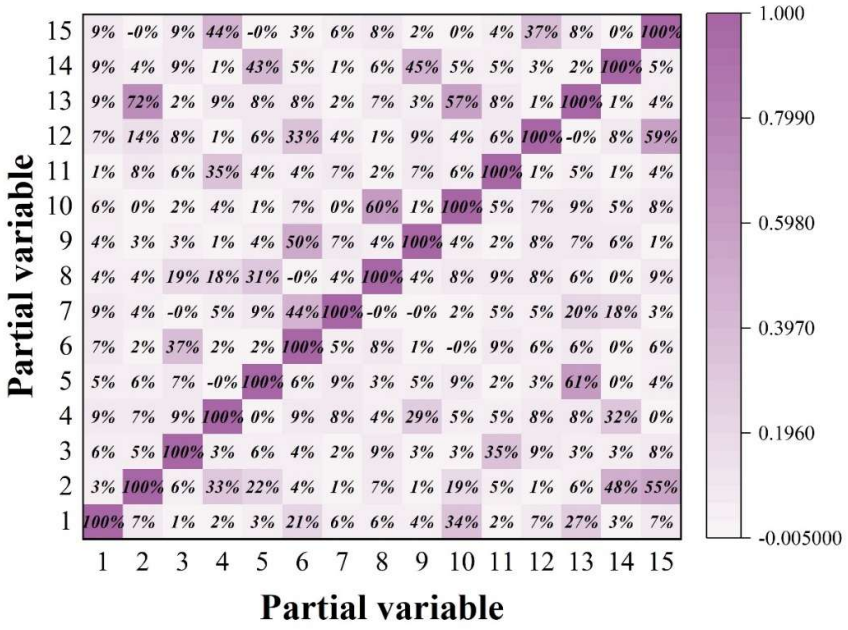

Fig. 4. Partial variable correlation coefficient

3.3. Hotel Customer Behavior Segmentation Model Construction

Using Means-K cluster analysis method, according to the selected 22 data indicators for clustering, according to the inflection point method to determine the K value to take 3 or 4 can be, and combined with the practical significance of this paper according to the customer value of the segmentation, and ultimately determined that K = 3, that is, clustered into 3 classes, each class of customer groups first with label 0, label 1, label 2, said. Among them, there are 241568 labels 0, accounting for 34.9%, 132079 labels 1, accounting for 19.08%, and 318372 labels 2, accounting for 46%.The relationship between the change of K value and SSE is shown in Figure 5.

Customer segmentation is to subdivide customers with the same value characteristics into the same category, which is able to divide customers into different categories, but the judgment of the high and low value of customers in different categories is limited, so it is necessary to further judge the value of each category of customers. After customer segmentation based on Means-K cluster analysis, it is necessary to identify the high and low value of each class of customers. In this paper, principal component analysis is performed based on the customer lifetime value evaluation system, from which the principal components representing value are identified, and the average score of each customer group in each category on the principal components representing value is calculated, so as to quantitatively judge the high or low value of customers.

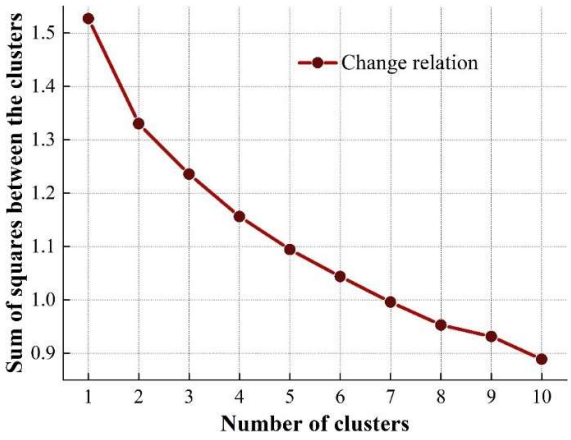

Fig. 5. The change between K values and SSE

The customer lifetime value evaluation system containing 22 indicators is subjected to principal component analysis, and in order to ensure the comparability of the principal component scores, the negative indicators in the evaluation system are taken as the opposite number to be positively oriented, and the cumulative contribution curve is shown in Figure 6.

The cumulative contribution of the first 14 principal components reaches 88.97%, which leads to the magnitude of the loadings corresponding to the first 14 principal components. The first principal component only corresponds to negative loadings on precancelrate and lastpvgap, which are -0.1163 and -0.0759, respectively, and positive loadings on all other indicators, and the first principal component is the one with the largest variance, i.e., carries the largest amount of information, among all linear combinations of the original variables, and the evaluation system of customer lifetime value constructed in this paper is itself based on the The customer value evaluation system constructed in this paper is itself based on customer value, so the first principal component can be interpreted as the value difference factor, and the average score of the first principal component of each type of customer group can be calculated, so as to quantitatively determine the level of customer value. The average score of the first principal component of different customer groups is shown in Table 3, the average score of the first principal component of the customer group corresponding to label 0 is -1.7709, accounting for 35%, which is a low-value customer, the average score of the first principal component

of label 1 is 3.3032, accounting for 19%, which is a high-value customer, and the average score of the first principal component of label 2 is -0.0089, accounting for 46%, which is a medium-value customer.

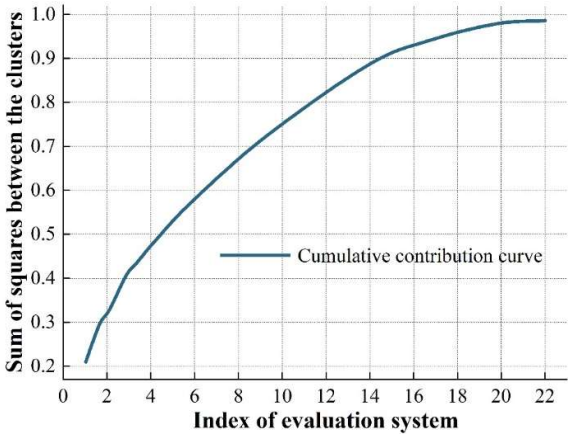

Fig. 6. Cumulative contribution curve

Table 3. The first main component of the different customer base is the average

Different customer base	Quantity	Proportion	The first average score	Customer segmentation
Label 0	241568	0.35	-1.7723	Low
Label 1	132079	0.19	3.3032	Height
Label 2	318372	0.46	-0.0089	Medium

In order to further verify the reasonableness of the customer segmentation results and the distinction between high and low customer value, three indicators, namely, talue_proficustomer_v, starprefer and neyearordernum_o, are selected to draw the kernel density curve, and the results are shown in Fig. 7 to Fig. 9.

It can be seen that for the customer value metric, label 0 is concentrated around 0, label 1 is concentrated between 0 and 30, and label 2 is concentrated between -5 and 5; for the star preference metric, label 0 is concentrated around 2, label 1 is concentrated between 4, and label 2 has peaks at 2, 3, and 4, and all of them have lower peaks than label 1; and for the number of users' annual orders, label 0 is concentrated between -2 and 0, label 1 is concentrated between -20 and 40, and label 2 is concentrated between -10 and 30. According to the distribution of the chart, it can be seen that the high and low customer value is consistent with the above judgment.

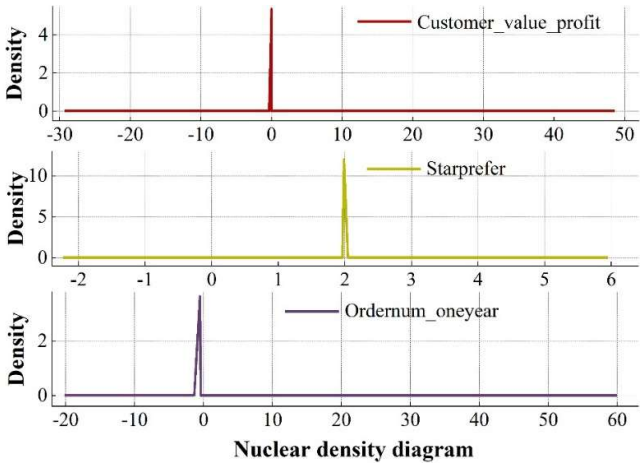

Fig. 7. Label 0 kernel density diagram

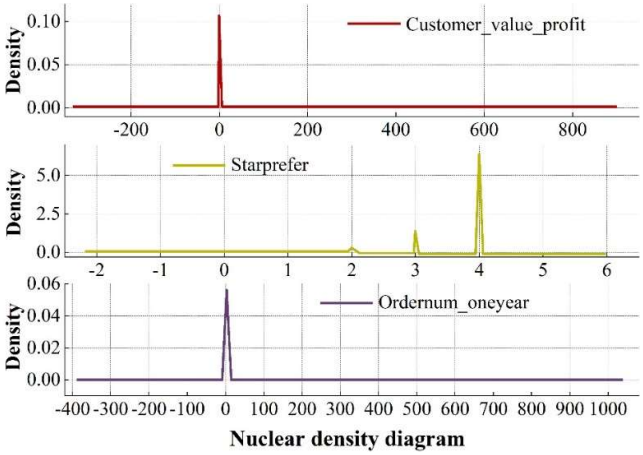

Fig. 8. Label 1 kernel density diagram

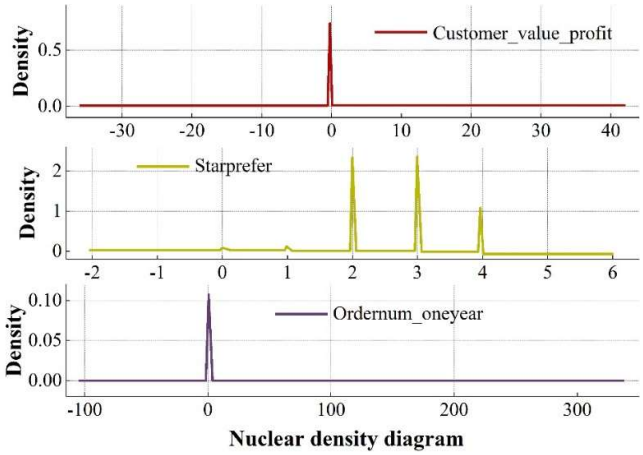

Fig. 9. Label 2 kernel density diagram

3.4. Analysis of Experimental Results and Management Recommendations

3.4.1. Results of pre-customer segmentation experiments. In this section, the algorithm of this paper is trained with the three baseline models on the preprocessed overall dataset, and the models are evaluated on the test set to verify the advantages and disadvantages of the prediction effect of each classification model. Table 4 shows the results of each evaluation index for each model.

It can be seen that this paper's algorithm performs best in all four evaluation metrics, and its overall prediction performance is significantly better than the other three models. By comparing the two indicators of precision rate and recall rate, it can be seen that this paper's algorithm is far superior to the other algorithms in terms of precision rate, indicating that this paper's algorithm has a higher accuracy rate among all the samples of predicting customers. In terms of model accuracy, this paper's algorithm achieved the highest value of 0.84, which is 10% higher than the second-ranked decision tree algorithm. Finally, it is worth noting that the model in this paper also far exceeds the other models in the comprehensive metric of AUC value.

Table 4. The prediction effect of each model before the customer segmentation

Algorithm	Accuracy	Precision	Recall	F1-score	AUC
LR	0.62	0.62	0.7	0.66	0.61
DT	0.68	0.71	0.66	0.69	0.68
DNN	0.66	0.66	0.74	0.7	0.66

This model	0.78	0.84	0.77	0.81	0.82

The ROC curves for each model are shown in Figure 10. Overall, the algorithm in this paper performs the best in terms of comprehensive performance, followed by the decision tree algorithm, while the DNN model performs the best in the important metric of recall.

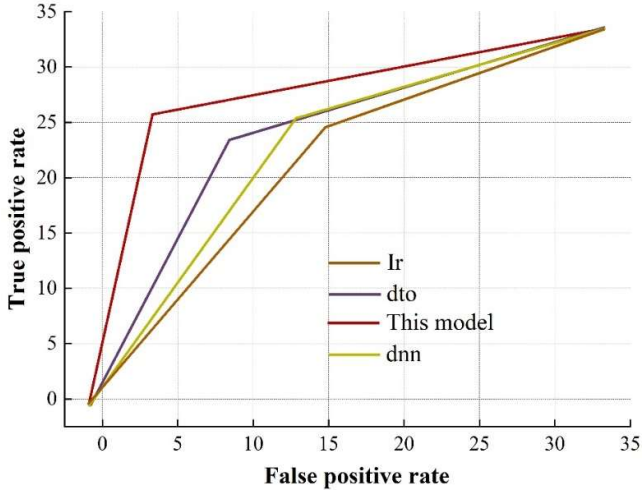

Fig. 10. The ROC curve of each model of the client

In the following, the algorithms of this paper that performed best in the above metrics are further analyzed using the confusion matrix. The confusion matrix for the algorithms of this paper used to create the model is given in Table 5, where cells on the horizontal line of the table indicate correct predictions and cells on the vertical line indicate incorrect predictions.

The correct rate of classification is 83.75%, which means that the model can correctly classify 8 out of 10 customers. However, out of 108,152 clients, only 85,016 can be recognized, which means the recall rate is 78.6%. Therefore the focus should be on the metric of recall rate. In summary, the comprehensive performance of this paper's algorithm before customer segmentation performs well, but there is still room for improvement.

Table 5. The confusion matrix of this algorithm

Observed Predicted	0	1	Recall
0	81842	10242	88.90%
1	23136	85016	78.60%
Precision	78%	89.20%	83.75%

3.4.2. Experimental results after customer segmentation. This section evaluates each model on each of the three types of customer groups after segmentation. Tables 6 and 7 show the results obtained for recall and AUC values for each model predicted on each type of customer group, respectively.

When the customers are segmented, the DNN model does not get further improvement in recall, and some of the data even decreases to a certain extent, and the two algorithms, LR and DT, also have relatively small recall improvement on all types of customers, which indicates that the improvement of customer segmentation on the recall of the above three models is very limited. Instead, it is observed that the algorithm in this paper achieves the best recall rate in all customer categories, with the highest recall rate of 84% on the 1st customer group.

On the AUC value index shown in Table 7, the algorithm in this paper performs much better, not only obtaining the highest AUC value on all three types of customers, but also far exceeding the decision tree model with the next best performance. Moreover, except for the DNN model, which still has no

improvement, the AUC values of all the other models are improved to some extent after customer segmentation. In summary, after customer segmentation, the model in this paper performs best in both recall rate and AUC value.

Table 6. The recall rate of each model in four groups of customers

	LR	DT	DNN	This model
Pre-segmentation	0.7	0.66	0.74	0.83
Class I	0.66	0.67	0.71	0.84
Class ii	0.63	0.65	0.66	0.79
Class iii	0.7	0.66	0.69	0.80

Table 7. The AUC of each model in four groups of customers

	LR	DT	DNN	This model
Pre-segmentation	0.61	0.68	0.65	0.88
Class I	0.6	0.69	0.62	0.81
Class ii	0.62	0.69	0.59	0.85
Class iii	0.65	0.67	0.61	0.86

In order to more conveniently and intuitively compare the prediction performance of each model, the author based on each model, each indicator, the values on its four types of customer groups are calculated the average results, as a way to represent the overall performance of each model in each indicator after customer segmentation, the results are shown in Figure 11. As can be seen from the figure, after customer segmentation, the model in this paper maintains the previous advantages, still performs well in all indicators, especially in the accuracy rate and AUC value of the two indicators, far more than the other three algorithms.

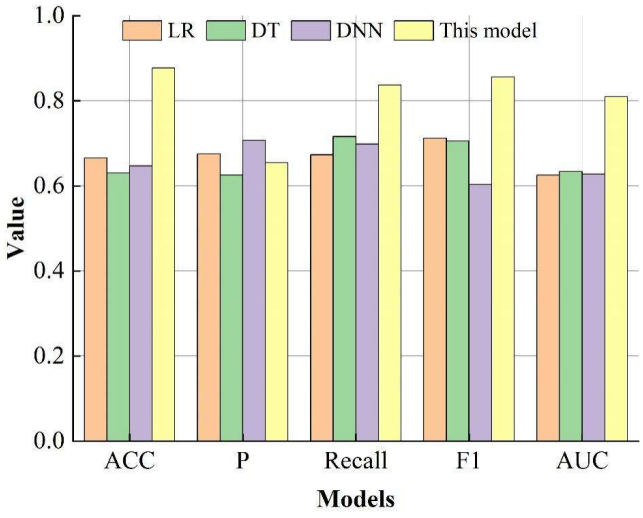

Fig. 11. The average index of each model on each customer's base

The following further shows the comparison of the prediction effect of the model before and after customer segmentation for this paper's algorithm, as shown in Figure 12.

Compared with the pre-customer segmentation, in addition to the three types of customers, the evaluation indexes predicted by using this paper's algorithm on the other two types of customer data have been improved to different degrees. Among them, they are mainly reflected in the accuracy rate,

precision rate and AUC value, and the highest values of these three metrics after customer segmentation are all increased by 6% compared with those before customer segmentation.

In summary, the above two sets of comparison experiments not only verify the obvious advantages of this paper's model over the other three models in the field of customer behavior prediction, but also prove that customer segmentation can improve the prediction performance of this paper's algorithm to a greater extent.

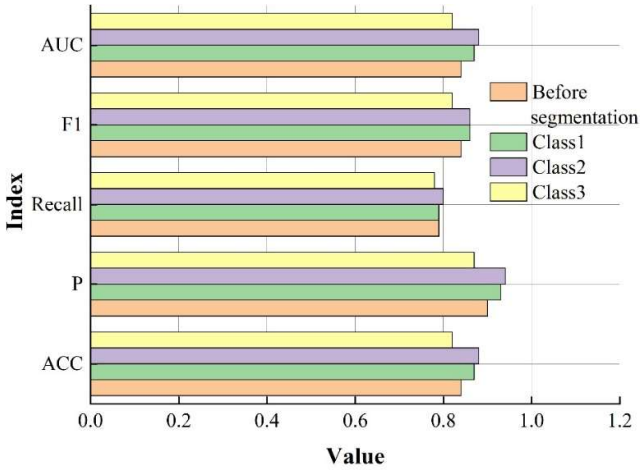

Fig. 12. The algorithm is a target of customer segmentation

3.5. Optimization of management strategies based on hotel customer behavior

The following is a specific analysis of the various customer groups and the corresponding management recommendations for their characteristics.

1) Customers seeking high quality. Although the proportion of such customers is not large, they are mainly characterized by the pursuit of higher cost performance. Therefore, we should focus on recommending high-quality services for them, and at the same time, we can cooperate with local scenic spots to attract them with neighboring attractions and special trips. Moreover, in order to ensure the quality of travel, such customers usually make travel plans in advance, so the hotel should push relevant information for them at least one week in advance.

2) Low-value customers. Low-value customers make up the largest portion of an organization. They have a very low demand for products, but due to their large base, an organization can't just give them up. Therefore, this part of the group, can be combined with specific behavioral conditions for further analysis, and strive to transform them into long-term customers. Marketing strategy is mainly to provide low-priced, discounted hotels, marketing efforts do not have to be too large, and do not do personalized marketing for the time being.

3) High-value customers. This type of customers not only respond better to the services provided by the enterprise, but also have the highest consumption frequency and the strongest consumption ability, so they are classified as high-value customers. Most of the customers in this category are business people, mainly characterized by short advance booking time, high number of orders, and high number of visits. Since this type of customer is usually the group of people who bring 80% of the revenue for the enterprise, it should be the key object of attention of the enterprise and implement personalized marketing for it.

4. Conclusion

In this paper, we use the variational Boston matrix and the improved particle swarm algorithm to realize hotel customer segmentation and construct a hotel customer behavior prediction model integrating multilevel features and SENet, which helps to improve the operation and management level of the hotel. The public dataset is downloaded from the online wine travel platform for simulation experiments, and the following conclusions are drawn:

1) According to the results of customer principal component analysis, labels 0, 1, and 2 represent low-, high-, and medium-value customers, accounting for 35%, 19%, and 46%, respectively.

2) The algorithm in this paper performs best on all four evaluation metrics, and the overall prediction performance is significantly better than the other three baseline models. The highest accuracy is 0.84, which is 10% higher than the second-ranked decision tree algorithm.

3) After customer segmentation, the model in this paper maintains the previous advantages and still performs well in all indicators, especially in the two indicators of accuracy rate and AUC value, which are far better than the other three algorithms.

Accordingly, it proposes management strategies for the three types of customer groups, realizing more detailed resource management based on the multidimensional attributes of different customer groups in hotel operation.

References

[1] Napierała, T., & Birdir, K. (2020). Competition in Hotel Industry: Theory, Evidence and Business Practice. European Journal of Tourism, Hospitality and Recreation, 10(3), 200-202.

[2] El-Adly, M. I. (2019). Modelling the relationship between hotel perceived value, customer satisfaction, and customer loyalty. Journal of Retailing and Consumer Services, 50, 322-332.

[3] Nobar, H. B. K., & Rostamzadeh, R. (2018). The impact of customer satisfaction, customer experience and customer loyalty on brand power: empirical evidence from hotel industry. Journal of Business Economics and Management, 19(2), 417-430.

[4] Lai, I. K. W. (2019). Hotel image and reputation on building customer loyalty: An empirical study in Macau. Journal of Hospitality and tourism Management, 38, 111-121.

[5] Nisa, C., Varum, C., & Botelho, A. (2017). Promoting sustainable hotel guest behavior: A systematic review and meta-analysis. Cornell Hospitality Quarterly, 58(4), 354-363.

[6] Uslu, A., & Caber, M. (2022). The role of hotel customer–employee bond in generating positive customer behavioral outcomes. Tourism Analysis, 27(3), 299-315.

[7] Liu, M. T., Liu, Y., Mo, Z., Zhao, Z., & Zhu, Z. (2020). How CSR influences customer behavioural loyalty in the Chinese hotel industry. Asia Pacific Journal of Marketing and Logistics, 32(1), 1-22.

[8] Izogo, E. E., & Mpinganjira, M. (2022). Social media customer behavioral engagement and loyalty among hotel patrons: does customer involvement matter?. International Journal of Tourism Cities, 8(3), 636-657.

[9] Deng, L., & Yu, D. (2014). Deep learning: methods and applications. Foundations and trends® in signal processing, 7(3–4), 197-387.

[10] Mu, R., & Zeng, X. (2019). A review of deep learning research. KSII Transactions on Internet and Information Systems (TIIS), 13(4), 1738-1764.

[11] Sharifani, K., & Amini, M. (2023). Machine learning and deep learning: A review of methods and applications. World Information Technology and Engineering Journal, 10(07), 3897-3904.

[12] Schmitt, M. (2023). Deep learning in business analytics: A clash of expectations and reality. International Journal of Information Management Data Insights, 3(1), 100146.

[13] Rane, N. L., Paramesha, M., Choudhary, S. P., & Rane, J. (2024). Artificial intelligence, machine learning, and deep learning for advanced business strategies: a review. Partners Universal International Innovation Journal, 2(3), 147-171.

[14] Maheswari, K., & Priya, P. P. A. (2017, March). Predicting customer behavior in online shopping using SVM classifier. In 2017 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS) (pp. 1-5). IEEE.

[15] Dadashnia, S., Niesen, T., Hake, P., Fettke, P., Mehdiyev, N., & Evermann, J. (2016). Identification of distinct usage patterns and prediction of customer behavior. BPI Challenge.

[16] Rajesh, G., Preethi, S. P., Priya, R. S., Rajesh, L., & Raajini, X. M. (2021). Customer Behavior Prediction for E-Commerce Sites Using Machine Learning Techniques: An Investigation. Reinventing Manufacturing and Business Processes Through Artificial Intelligence, 79-93.

[17] Reutterer, T., Platzer, M., & Schröder, N. (2021). Leveraging purchase regularity for predicting customer behavior the easy way. International Journal of Research in Marketing, 38(1), 194-215.

[18] Leung, X. Y., Bai, B., & Erdem, M. (2017). Hotel social media marketing: a study on message strategy and its effectiveness. Journal of Hospitality and Tourism Technology, 8(2), 239-255.

[19] Dursun-Cengizci, A., & Caber, M. (2024). Using machine learning methods to predict future churners: an analysis of repeat hotel customers. International Journal of Contemporary Hospitality Management.

[20] Khodabandehlou, S., & Zivari Rahman, M. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. Journal of Systems and Information Technology, 19(1/2), 65-93.

[21] Chang, Y. M., Chen, C. H., Lai, J. P., Lin, Y. L., & Pai, P. F. (2021). Forecasting hotel room occupancy using long short-term memory networks with sentiment analysis and scores of customer online reviews. Applied Sciences, 11(21), 10291.

[22] Hamdan, I. Z. P., Othman, M., Hassim, Y. M. M., Marjudi, S., & Yusof, M. M. (2023). Customer Loyalty Prediction for Hotel Industry Using Machine Learning Approach. JOIV: International Journal on Informatics Visualization, 7(3), 695-703.

[23] Park, E., Kang, J., Choi, D., & Han, J. (2020). Understanding customers' hotel revisiting behaviour: a sentiment analysis of online feedback reviews. Current Issues in Tourism, 23(5), 605-611.

[24] Li, J., Pan, S., & Huang, L. (2019). A machine learning based method for customer behavior prediction. Tehnički vjesnik, 26(6), 1670-1676.

[25] Che Yu Hung & Chien Chih Wang.(2024).An Approach for Multi-Item Product Sales Forecasting Based on Advancing the BCG Matrix with Matrix-Clustering and Time Modeling Techniques. Systems(10),388-388.

[26] Shakhatreh Majd,Shakhatreh Hazim & Ababneh Ahmad. (2023). Efficient 3D Positioning of UAVs and User Association Based on Hybrid PSO-K-Means Clustering Algorithm in Future Wireless Networks. Mobile Information Systems

[27] C. S. Parvathy & J. P. Jayan. (2024). Automatic Lung Cancer Detection Using Computed Tomography Based on Chan Vese Segmentation and SENET. Optical Memory and Neural Networks(3),339-354.