scientific reports

OPEN

Forecast and analysis of aircraft passenger satisfaction based on RF-RFE-LR model

Xuchu Jiang[1], Ying Zhang[1], Ying Li[2] & Biao Zhang[3]✉

Airplanes have always been one of the first choices for people to travel because of their convenience and safety. However, due to the outbreak of the new coronavirus epidemic in 2020, the civil aviation industry of various countries in the world has encountered severe challenges. Predicting aircraft passenger satisfaction and excavating the main influencing factors can help airlines improve their services and gain advantages in difficult situations and competition. This paper proposes a RF-RFE-Logistic feature selection model to extract the influencing factors of passenger satisfaction. First, preliminary feature selection is performed using recursive feature elimination based on random forest (RF-RFE). Second, based on different classification models, KNN, logistic regression, random forest, Gaussian Naive Bayes, and BP neural network, the classification performance of the models before and after feature selection is compared, and the prediction model with the best classification performance is selected. Finally, based on the RF-RFE feature selection, combined with the logistic model, the factors affecting customer satisfaction are further extracted. The experimental results show that the RF-RFE model selects a feature subset containing 17 variables. In the classification prediction model, the random forest after RF-RFE feature selection shows the best classification performance. Finally, combined with the four important variables extracted by RF-RFE and logistic regression, further discussion is carried out, and suggestions are given for airlines to improve passenger satisfaction.

With the continuous improvement of people's living standards, civil aviation industry customer groups are growing, and people have put forward higher requirements for aviation service quality. In addition, the COVID-19 outbreak has hit most industries around the world, bringing many to a standstill. Movement restrictions and travel bans have had a serious impact on the transport sector, especially the airline sector, and the quality of airline service has become a more important factor for people's choice[1]. In a fierce competitive environment, the aviation industry should also develop from a simple transport role to service. Improving service quality is an important part of competitiveness and an important guarantee for sustainable and healthy development[2,3]. Therefore, airlines should investigate passengers' satisfaction with various services and overall satisfaction in a timely manner and accurately understand the service quality of existing services. In addition, airlines should accurately grasp the main factors affecting passenger satisfaction, and formulate corresponding strategies to improve service quality, to maximize the overall passenger satisfaction with the airline and improve passenger loyalty.

Based on the above challenges, this study uses the full-service passenger information and satisfaction survey results as the research object. While using machine learning algorithms to predict passenger satisfaction, the main factors affecting passenger satisfaction are further studied, and the priority of the factors are given.

This study provides a reference for airlines to accurately predict the overall satisfaction of passengers with the company's services and understand the main factors by using passenger personal and service satisfaction survey information. This study serves as a reference for airlines to use customer evaluation-driven service evaluation methods, and provides support for airlines to improve their competitiveness.

Airline passenger satisfaction. Customer satisfaction is increasingly recognized as a determinant of business performance and a strategic tool for gaining competitive advantage. Cardozo first put forward the viewpoint of customer satisfaction in 1965 and introduced it into the field of marketing for the first time[4],

[1]School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, China. [2]Department of Scientific Research, Zhongnan University of Economics and Law, Wuhan 430073, China. [3]School of Computer Science, Liaocheng University, Liaocheng 252059, China. ✉email: zhangbiao1218@gmail.com

and the theory of customer satisfaction has also made continuous development. High and stable customer satisfaction is considered an important determinant of an organization's long-term profitability. Yeung[5] found a positive relationship between customer satisfaction and a range of financial performance indicators by using the American Consumer Satisfaction Index. Research has also shown a significant moderate-to-strong association between satisfaction and a company's financial and market performance. More specifically, customer satisfaction is strongly linked to retention, revenue, earnings per share, and stock price[6].

For the aviation industry, some studies have used aviation services to build an index model for passenger satisfaction. Based on the combination of China's customer satisfaction index model and the actual situation of China Southern Airlines' satisfaction management, Zhang[7] designed the China Southern Airlines customer satisfaction evaluation index and proposed nine secondary indicators with air transport characteristics: flight operation quality satisfaction degree, ticketing service satisfaction, ground service satisfaction, air service satisfaction, arrival station service satisfaction, irregular flight service satisfaction, consumption value perception, overall satisfaction, and customer loyalty.

There are also studies using flight data or text reviews to predict passenger satisfaction. Sankaranarayanan et al.[8] used a logistic model tree (LMT) machine learning approach to predict passenger satisfaction levels based on factors such as airport punctuality, number of flights, punctuality rankings, average delays, and queue times for inferring passenger perceptions of punctuality and delay-related event satisfaction. Kumar et al.[9] predicted passengers' positive or negative attitudes by textually evaluating passengers' flight reviews.

Many studies have discussed the impact of passenger satisfaction on modern businesses, and how passenger satisfaction affects business performance and development in the airline industry, so it is crucial to obtain timely and accurate information on passenger satisfaction with airlines.

Service quality and passenger satisfaction. To achieve high levels of customer satisfaction, service providers should provide high levels of service quality, as service quality is often considered a prerequisite for customer satisfaction[10.] Since passengers are the direct recipients of services, service quality indirectly affects enterprise development by affecting passenger satisfaction. Therefore, airlines can understand the quality of the services provided by passengers' satisfaction with each service, check the services, and then improve the service quality. Park et al.[11] show that airlines that provide services that meet customer expectations enjoy higher levels of passenger satisfaction and value perception. Jiang et al.[12] took China Eastern Airlines (CEA) as a case study to investigate the domestic passengers of China Eastern Airlines at Wuhan Tianhe International Airport, China. Hu et al. [13]found that poor service quality can lead to customer dissatisfaction by using the Kano model to design a quality risk assessment model. Chow et al.[14] studied the relationship between customer satisfaction measured by customer complaints and service quality using fixed-effect Tobit analysis, discussed the seasonality of customer satisfaction, and compared customer satisfaction between state-controlled and nonstate-controlled operators.

In many studies, it is proposed to combine passenger satisfaction information to construct a service quality evaluation framework[15]. Therefore, the influence mechanism between service quality and satisfaction should be that service quality affects satisfaction, and satisfaction is a direct and effective indicator of service quality. Therefore, this study combines service quality with passenger satisfaction, predicts overall satisfaction with passengers' ratings of service quality and uses satisfaction information to analyze airline service quality.

Important service attributes. For the aviation industry, it is more efficient to improve customer satisfaction by accurately understanding the main factors affecting passenger satisfaction and making improvements based on service priorities. Several studies have investigated the main factors influencing passenger satisfaction. By constructing a nested logit model of airport-airline choice in the "two-step" decision-making process of air passengers, Suzuki[16] determined that the factors that play an important role in airline choice are ticket price, frequency of flight service provided to desired destinations and frequent flyer membership. Tsafarakis et al.[17] proposed that the improvement of onboard entertainment onboard Wi-Fi services can improve airline passenger satisfaction according to the multi-standard satisfaction analysis method. Hess[18] looked at these factors separately for several market segments and concluded that visit times, flight times and airfares are important for both business and holidaymakers. With the development of text mining technology, some researchers have mined customer comment text on web pages to analyze passenger satisfaction and the main influencing factors. Lucini et al.[19] used a text mining method to analyze online customer comments and predict passengers' attitudes and concluded that first-class passengers should be provided with customer service for different reasons. Provide comfort for premium economy passengers; Conclusions on checked baggage and wait times for economy passengers Cabin staff, in-flight service and cost performance were found to be the three most important dimensions in the prediction of airlines recommended by passengers. According to an online review study by Brochado[20], on-board service, airport operation, ground service and other factors have a significant impact on service quality assessment by using Leximancer to perform quantitative content analysis of airline passenger web reviews.

Many previous studies evaluated passenger satisfaction by statistical tests, building a satisfaction index system[7] or text analysis[20]. Then, this study uses machine learning to determine the main factors affecting satisfaction based on the data of airline passenger satisfaction surveys and gives priority to the services that airlines need to pay attention to, providing strategic support for companies committed to improving passenger satisfaction.

Feature selection based on RF-RFE.

Feature selection refers to removing redundant or irrelevant features from a set of features. Whether the samples contain irrelevant or redundant information directly affects the performance of the classifier, so it is very important to choose an effective feature selection method. Guyon[21] reviewed the existing variable feature selection methods: filtering, embedding and packaging. The filtering method sorts the features in the preprocessing step and is independent of the learning algorithm, which means that the selected features can be transferred to any modeling algorithm. The filter can be further classified according to the filtering measures used, i.e., information, distance, dependence, consistency, similarity and statistical measures. The embedded method takes feature selection as part of the implementation of the modeling algorithm. The features of method selection depend on the learning algorithm. Two typical examples are lasso and various decision tree-based algorithms. Wrapper uses an independent algorithm to train the prediction model for candidate feature subsets and uses greedy strategies such as forward or backward to identify the optimal feature subset from all possible subsets in the learning process.

Recursive feature elimination (RFE) is a sequence backward selection algorithm belonging to the wrapper. This method uses a specific underlying algorithm to continuously reduce the scale of the feature set through recursion to select the required features. Guyon[21] proposed RFE based on the SVM model and proved that SVM-RFE achieved very good results in the process of gene selection, which has become a widely used method in gene selection research[22]. Marcelo MCA combined RFE with logistic regression (LR) and a support vector machine (SVM) to select the features of tobacco spectral information, which improved the prediction accuracy of the model[23]. Wei[24] and others proposed the SVM-RFE-SF method, which divides the original gene set into several gene subsets, and RFE divides the genome with the smallest score of the sorting criteria, which effectively solves the problem of heavy calculation caused by eliminating only one gene at a time.

The feature selection method using random forests in most studies is the wrapper[25]. In the research of feature selection based on RF-RFE, Wu combined RF with RFE and used the importance ranking output of the RF algorithm to select variables[26]. Chen tested two completely different molecular biology data sets and found that using RF-RFE feature selection improves the quality of the model and makes the model construction process more efficient than the model without any feature selection[27]. Shang et al.[28] used the RF-RFE algorithm to select important variables from the initial variables for evaluating traffic event detection and took important variables as input to prove that the model has better performance.

This study uses machine learning models to predict the overall satisfaction of passengers with the full service provided by the airline. Combined with feature selection, the prediction models before and after feature extraction are compared, and the best prediction model is selected. We aim to select the important factors that affect passenger satisfaction. The preliminary feature selection is combined with the logistic model for further selection, and the model results are used to prioritize the influencing factors.

Methodology

In this section, we present a detailed introduction to the preliminary feature selection method RF-RFE (random forest-based recursive feature elimination) and various classification models used for passenger satisfaction in this study.

Recursive feature elimination based on random forest.

RF. The RF proposed by Breiman[29] is a parallel integration algorithm based on a decision tree. Because of its relatively good precision, robustness and ease of use, it has become one of the most popular machine learning methods. The decision tree may be completely different due to the small change in data, so it is not stable enough. The RF reduces the variance brought by a single decision tree, improves the prediction performance of the general decision tree, and can give the importance measurement of variables, which brings substantial improvement to the decision tree model.

RF uses a decision tree as the base learner to construct a bagging ensemble. Bagging is a parallel integrated learning algorithm based on a self-help sampling method. Each sampling set is used to train a base learner, and then these base learners are combined. When combining the prediction output, the simple voting method is usually used for the classification task.

Let the training set $D = \left\{ \left(x_1, y_1\right), \left(x_2, y_2\right), \ldots, \left(x_n, y_n\right) \right\}$, and the prediction result of the new sample is Eq. (1):

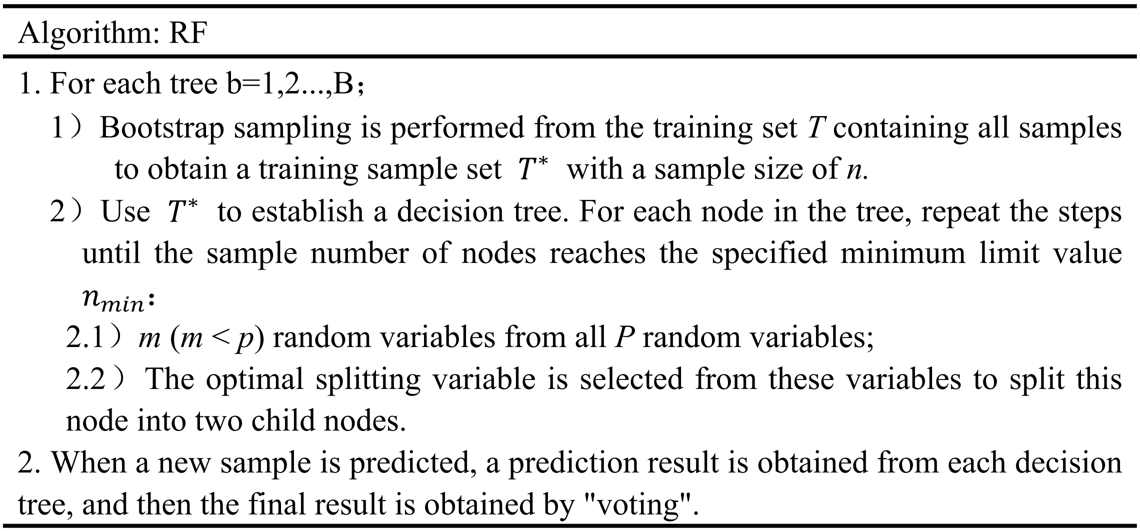
$$f(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{t=1}^{T} \mathbb{I}\left(h_t(x) = y\right) \tag{1}$$

where y is the output category set, $h_t(x)$ is the prediction result of the new sample x by the t-th learner, and y is the real category of the sample.

RF introduces the selection of random attributes on the basis of bagging integration. Different from selecting an optimal attribute when a single decision tree divides attributes, RF adopts the method of random selection for each node attribute set in the decision tree, first randomly selects an attribute subset from all attributes, and then selects an optimal attribute from the subset. Therefore, based on the sample disturbance brought by bagging, the RF further introduces attribute disturbance, which increases the generalization performance of the integration. The algorithm description of RF is shown in Table 1.

Importance of RF characteristics. The importance measurement indicators based on RF include the mean decrease impurity (MDI) based on the Gini index and the mean decrease accuracy (MDA) based on OOB data[30]. This method uses the frequency of attributes in the RF decision tree to reflect the importance of features. This paper chooses the MDI method based on the Gini index to measure the importance of features.

When constructing the CART decision tree, RF takes the attribute with the largest Gini gain as the splitting attribute by calculating the Gini gain of all attributes of the node. Gini represents the probability that a randomly selected sample in the sample set is misclassified, let p_k be the proportion of class k samples, and the calculation equation is Eq. (2):



```
Algorithm: RF
1. For each tree b=1,2...,B；
    1）Bootstrap sampling is performed from the training set $T$ containing all samples
       to obtain a training sample set $T^*$ with a sample size of $n$.
    2）Use $T^*$ to establish a decision tree. For each node in the tree, repeat the steps
       until the sample number of nodes reaches the specified minimum limit value $n_{min}$：
        2.1）$m$ ($m < p$) random variables from all $P$ random variables;
        2.2）The optimal splitting variable is selected from these variables to split this
             node into two child nodes.
2. When a new sample is predicted, a prediction result is obtained from each decision
   tree, and then the final result is obtained by "voting".
```

Table 1. RF algorithm.

$$Gini(p) = \sum_{k=1}^{K} p_k(1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2 \tag{2}$$

The Gini gain obtained by dividing the data set according to attribute a is Eq. (3):

$$Gini(D, a) = Gini(D) - \sum_{v}^{V} \frac{|D^v|}{|D|} Gini(D^v) \tag{3}$$

where V is the number of value categories of attribute a and $|D^v|$ is the number of value categories of attribute a.

Based on the calculation of feature importance, the specific steps are as follows:

(1) For each decision tree, the node where feature $\propto$ appears is set A, and the change in the Gini index before and after node i branch is calculated as follows Eq. (4):

$$\Delta Gini = Gini(i) - Gini(l) - Gini(k) \tag{4}$$

where $Gini(l)$ and $Gini(k)$ are the Gini index of the new node after branching.

(2) The importance of feature $\propto$ in the tree is shown in Eq. (5):

$$IM_{\propto} = \sum_{a \in A} \Delta Gini \tag{5}$$

where a is the node where feature $\propto$ appears.

(3) Suppose n is the number of decision trees, and the importance of feature $\propto$ is Eq. (6):

$$IMPORTANCE(\propto) = \sum_{N} IM_{\propto} \tag{6}$$

Then, normalize the importance of all features in Eq. (7):

$$IM(\propto) = \frac{IMPORTANCE(\propto)}{\sum_{i}^{c} IMPORTANCE(i)} \tag{7}$$

where c is the number of features.

(4) The larger the $IM(\propto)$ value is, the more important the feature is to the result prediction, that is, the higher the importance of the feature.

Recursive feature elimination based on RF. RF-RFE uses RF as an external learning algorithm for feature selection, calculates the importance of features in each round of feature subset, and removes the features corresponding to the lowest feature importance to recursively reduce the scale of the feature set, and the feature importance is constantly updated in each round of model training. Based on the selected feature set, this study uses cross validation to determine the feature set with the highest average score based on classification accuracy. The algorithm flow chart is shown in Fig. 1.

The RF-RFE flow is as follows:

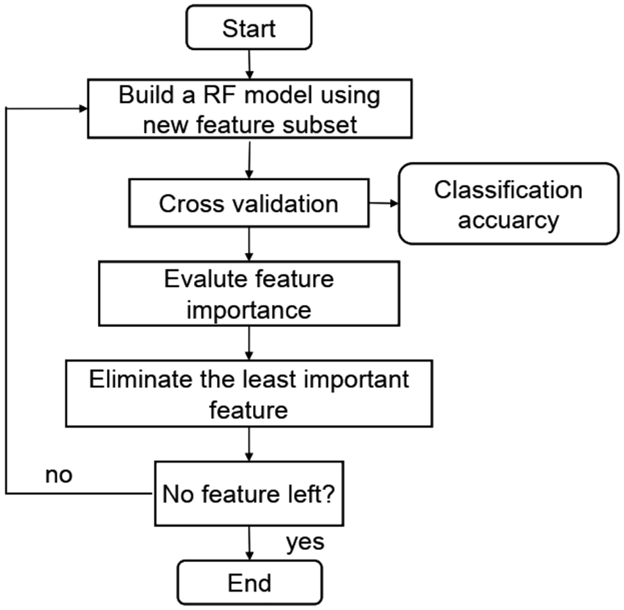

Start
Build a RF model using new feature subset
Cross validation
Classification accuarcy
Evalute feature importance
Eliminate the least important feature
no
No feature left?
yes
End

Figure 1. The RF-RFE flow.

(1) Bootstrap sampling is carried out from the training set T containing all samples to obtain a training sample set T^* with a sample size of n. The decision tree is established by using T^*, and a total of b decision trees are generated by repeating this process;
(2) The prediction results of each decision tree are combined by "voting", and the effect of the RF regression model is evaluated based on classification accuracy by using the fivefold cross validation method;
(3) Calculate and sort the importance $IM(\propto)$ of each feature $\propto$ in the feature set based on MDI;
(4) According to the backward selection of the sequence, delete the feature with the lowest feature importance, and repeat steps 1–3 for the remaining feature subset until the feature subset is empty. According to the cross-validation results of each feature subset, the feature subset with the highest classification accuracy is determined.

Satisfaction prediction based on machine learning algorithm. According to whether the processed data are marked artificially, machine learning can be generally divided into supervised learning and unsupervised learning. Supervised learning data sets include initial training data and manually labeled objects. The machine learning algorithm learns from labeled training data sets, tries to find the pattern of object division, and takes labeled data as the final learning goal. Generally, the learning effect is good, but the acquisition cost of labeled data is high. Unsupervised learning processes unclassified and unlabeled sample set data without prior training, hoping to find the internal rules between the data through learning to obtain the structural characteristics of the sample data, but the learning efficiency is often low. The satisfaction status in this study is the data set label. In the training process, the supervised machine learning algorithm learns the corresponding relationship between features and labels and applies this relationship to the test set for prediction.

k-nearest neighbors (KNN). KNN is a supervised learning algorithm. Because the training time overhead is zero, it is also representative of "lazy learning"[31]. K-nearest neighbor has been used as a nonparametric technique in statistical estimation and pattern recognition. The working principle is as follows: for a given new sample, find the K samples closest to the sample in the training set based on a certain distance measurement and take the number of categories with the largest number of K samples as the category of the new sample. The samples are not processed in the training stage, so it belongs to "lazy learning". As shown in Fig. 2, if there are 3 squares, 2 circles and 1 triangle around a data point, it is considered that the data point may be square. The parameter K in KNN is the number of nearest neighbors in majority voting.

LR. LR is used to evaluate the relationship between dependent variables and one or more independent variables, and the classification probability is obtained by using logical functions[32]. It is a learning algorithm with a logistic function as the core. A logistic function is used to compress the output of the linear equation to (0, 1). The logistic function is defined as Eq. (8):

$$Logistic(z) = \frac{1}{1 + e^{-z}} \tag{8}$$

Consider the binary classification problem, given the data set $D = (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N), x_I \subseteq R^n, y_i \in 0, 1, i = 1, 2, \cdots, N$.

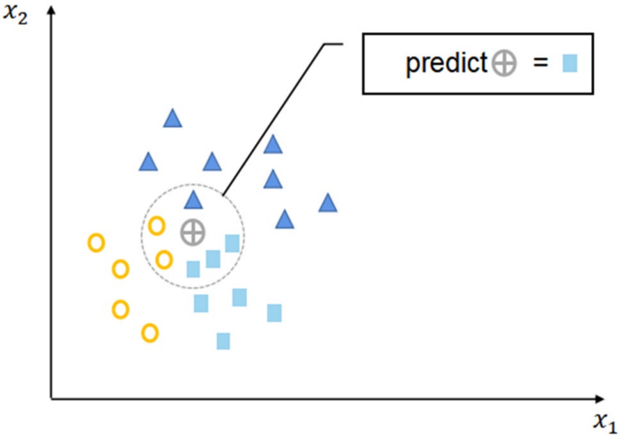


Figure 2. KNN.

P is the probability that the sample is a positive example, and the coefficient in the following formula is determined by LR through the maximum likelihood method $\beta_0, \beta_1, \cdots, \beta_k$ to make an estimate [Eqs. (9) and (10)]:

$$logit\left(p\right) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{9}$$

$$p = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\right)} \tag{10}$$

When *P* is greater than the preset threshold, the sample is divided into positive examples, and vice versa.

$\frac{p}{1-p}$ is called the odds ratio (odds), which refers to the ratio of the probability of event occurrence to the probability of event nonoccurrence. The logarithm of the winning rate is linear with the coefficient of the variable. When the features have been standardized, the greater the absolute value of the coefficient, the more important the feature is. If the coefficient is positive, this characteristic is positively correlated with the probability that the target value is 1; if the coefficient is negative, this characteristic is positively correlated with the probability that the target value is 0.

Gaussian Naive Bayes (GNB). Naive Bayes (NB) is a direct supervised machine learning algorithm[33]. The NB classifier is based on the Bayesian probability theorem and predicts future opportunities according to previous experience. NB assumes that the input variables are conditionally independent [Eq. (11)].

$$P\left(Y = y_k | X_1, \ldots, X_n\right) = \frac{P\left(Y = y_k\right) P\left(X_1, \ldots, X_n | Y = y_k\right)}{\sum_j P\left(Y = y_j\right) P\left(X_1, \ldots, X_n | Y = y_k\right)} = \frac{P\left(Y = y_k\right) \prod_i P\left(X_i | Y = y_k\right)}{\sum_j P\left(Y = y_j\right) \prod_j P\left(X_i | Y = y_j\right)} \tag{11}$$

where *X* is the input vector $(X_1, X_2, \ldots, X_n)$ and *Y* is the output category.

On the basis of NB, GNB further assumes that the prior probability of the feature is a Gaussian distribution, that is, the probability density function is as follows in Eq. (12):

$$P\left(x_i = x | Y = y_k\right) = \frac{1}{\sqrt{2\pi\delta_{ik}^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_{ik}}{\delta_{ik}}\right)^2} \tag{12}$$

For a given test set sample $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{n})$, calculate *P* [Eq. (13)]:

$$P\left(Y = y_k\right) \prod_i P\left(X_i | Y = y_k\right), \quad k = 1, 2, \ldots, K \tag{13}$$

To determine the class of the sample *y* [Eq. (14)]:

$$y = \underset{y_k}{argmax}\, P\left(Y = y_k\right) \prod_i P\left(X_i | Y = y_k\right) \tag{14}$$

RF. The working principle of RF[34] is to combine the results of each decision tree, as shown in Fig. 3. This strategy has better estimation performance than a single random tree: the estimation of each decision tree has low deviation but high variance, but clustering realizes the trade-off between overall deviation and variance and provides the importance of prediction variables to the prediction of result variables. RF has good prediction

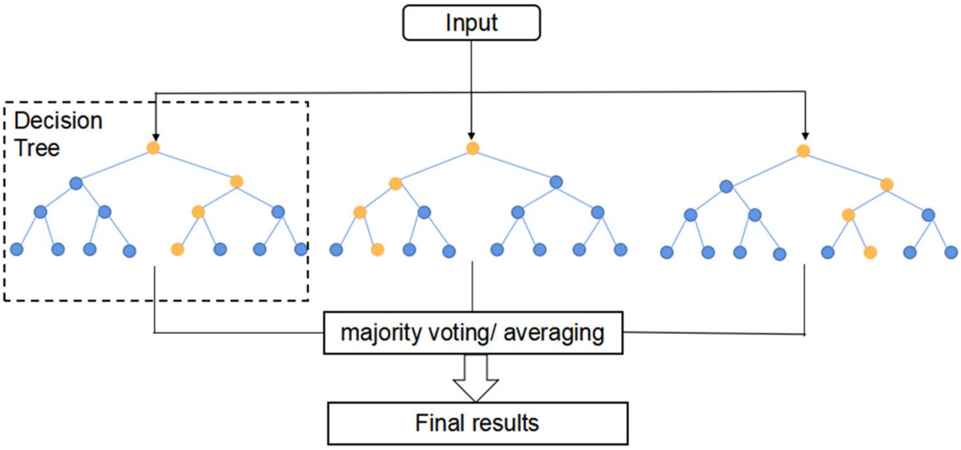


Figure 3. Working principle of RF.

performance in practical applications and can be used to address multiclass classification problems, category variables and sample imbalance problems.

Backpropagation neural network (BPNN). BPNN is one of the most widely used neural network models and is a typical error backpropagation algorithm[35]. Since the emergence of BPNNs, much research has been done on the selection of activation functions, the design of structural parameters and the improvement of network defects. The main idea of the BP algorithm is to divide the learning process into two stages: forward transmission and reverse feedback. In the forward transmission stage, the input sample reaches the output layer from the input layer through the hidden layer, and the output end forms an output signal. In the backpropagation stage, the error signals that do not meet the precision requirements are spread forward step by step, and the weight matrix between neurons is corrected through the pre-adjustment and post-adjustment cycles. When the iteration termination condition is met, the learning stops.

(1) Forward transmission

First, the input vector of the sample is X, T is the corresponding output vector, m is the number of neural units in the input layer, and P is the number of nodes in the output layer:

$$X = (x_1, \ldots, x_m)$$

$$T = \left(T_1, \ldots, T_p\right)$$

The calculation process equation of the forward transmission output layer is Eq. (15):

$$I_j = \sum_{i=1}^{m} w_{ij}x_i + \theta_j \tag{15}$$

where j represents the node of the hidden layer, w is the weight matrix between the input layer node and the hidden layer node, θ_j is the threshold of node j, and the output value of node j is Eq. (16):

$$O_j = f\left(I_j\right) \tag{16}$$

where f is called the activation function, which is the processing of the input vector. The function can be linear or nonlinear.

(2) Reverse feedback

Calculate the error between the true value of the sample and the output value of the sample. For the problem of second classification, two neural units are often used as the output layer. If the output value of the first neural unit of the output layer is greater than that of the second neural unit, it is considered that the sample belongs to the first category (Eq. (17)):

$$E_i = O_i(1 - O_i)(T_i - O_i) \tag{17}$$

The error of the middle hidden layer is accumulated by weight through the node error of the next layer (Eq. (18)):

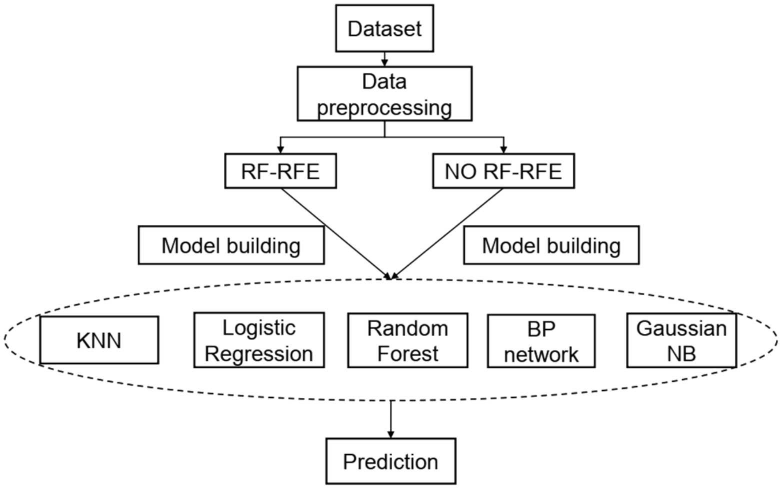


Figure 4. Research framework.

$$E_j = O_i(1 - O_i)\sum_k E_k W_{jk} \tag{18}$$

where E_k is the error of the k-th node of the next layer and W_{jk} is the weight from the j-th node of the current layer to the k-th node of the next layer.

Update the weights and offsets, respectively (Eq. (19)):

$$\begin{aligned} W_{ij} &= W_{ij} + \Delta W_{ij} = W_{ij} + \lambda E_j O_i \\ \theta_j &= \theta_j + \triangle\, \theta_j = \theta_j + \lambda E_j \end{aligned} \tag{19}$$

where λ is the learning rate, with a value of 0–1. When the training reaches a certain number of iterations or the accuracy is higher than a certain value, the training is stopped.

Passenger satisfaction prediction

Based on the preliminary selection of features, this study establishes models to predict passenger satisfaction. After comparing the prediction performance of various classification models before and after feature selection, we select the model with the best prediction performance. Figure 4 shows the research framework of the prediction analysis process.

Data source. The data used in this study are the passenger satisfaction data set of an American airline on Kaggle (https://www.kaggle.com/binaryjoker/airline-passenger-satisfaction). Through the survey of passengers who arrived at the airport in 2015, a sample of 129,880 passengers using the full service of the airline was collected. There are 23 attributes in the data set, of which the input variables include 4 numerical continuous variables, 4 class discrete variables and 14 qualitative sequential variables, indicating the customer's satisfaction with relevant services (0–5 points). The data used in this study mainly contain three dimensions of information, including basic information, flight information and satisfaction information. The constituent elements and the specific variable names and variable attributes are shown in Table 2. The output variables are category variables, that is, passengers' final satisfaction and dissatisfaction or neutral attitude.

Data preprocessing. *Data cleaning and standardization.* There are 393 missing values in the data set. After deleting the missing values, there are 129,487 samples. There are large differences in the value range of the four numerical variables: age, flight distance, departure delay (minutes) and arrival delay (minutes). The data are standardized and transformed into dimensionless values for comparison and weighting between different variables.

Correlation test. According to the correlation matrix results, the correlation coefficient between departure delay (departure_delay_in_minutes) and arrival delay (arrival_delay_in_minutes) is as high as 0.964, so the variable of arrival delay is removed.

Feature selection based on RF-RFE. In this study, the combination of RFE and the cross validation method is used to calculate the selected feature set in each RFE stage for cross validation. Taking the accuracy as the evaluation criterion, the number of features with the highest accuracy and the corresponding feature subset are finally determined. The RF-RFE feature selection results are shown in Fig. 5. The broken line diagram can intuitively judge the accuracy results obtained by the number of different feature subsets, in which the number

Variable properties	Variable name	Satisfaction (0–5)	Flight operation quality	Departure arrival time convenient
Numerical type	Age		Ticketing service	Ease of online booking
	Flight distance			online boarding
			Ground service	Gate location
	Departure delay in minutes			Baggage handling
				Checking service
	Arrival delay in minutes		Air service	Inflight Wi-Fi service
				Food and drink
Category type	Gender			Seat comfort
	Type of travel			Inflight entertainment
	Customer type			Onboard service
				Leg room service
	Customer class			Inflight service
				Cleanliness

Table 2. Variable name and attribute.

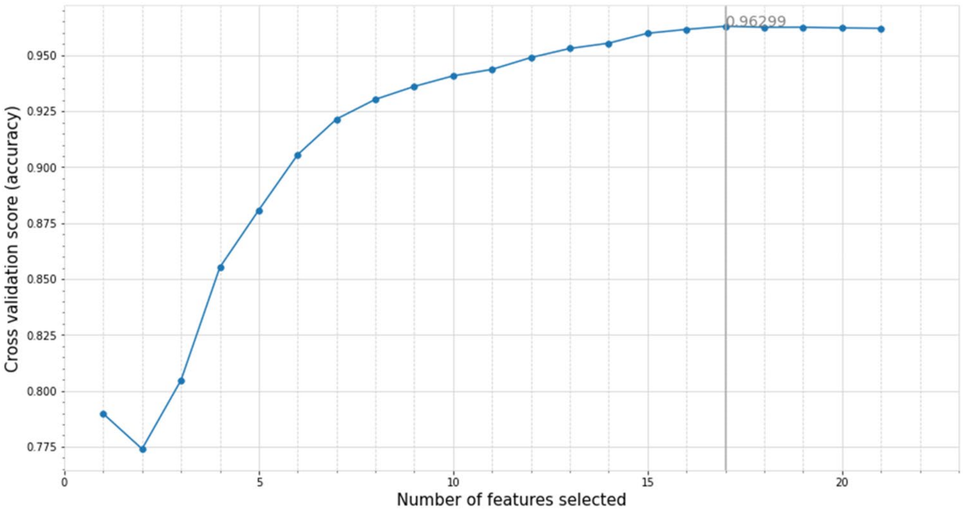


Figure 5. Feature selection results based on RF-RFE.

of features in the feature subset with the highest accuracy is 17, and the classification accuracy of the test set is 0.963.

Model prediction.

Model evaluation index. To evaluate the classification performance of the classifier used in the research, five evaluation indexes are introduced: accuracy, precision, recall, F value and AUC value. For the binary classification problem, the sample can be divided into four cases according to the combination of its real category and the category predicted by the classifier: true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Obviously, TP + FP + TN + FN = the total number of test sets.

Accuracy is the most commonly used performance measure in classification tasks. It refers to the proportion of the number of correctly classified samples to the total number of samples for a given test set. Its equation is Eq. (20):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

Precision refers to the proportion of samples whose real situation is positive in the samples determined as positive by the classifier for a given test set. Its equation is Eq. (21):

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

The recall rate refers to the proportion of samples determined as positive by the classifier in all positive samples for a given test set. The equation is Eq. (22):

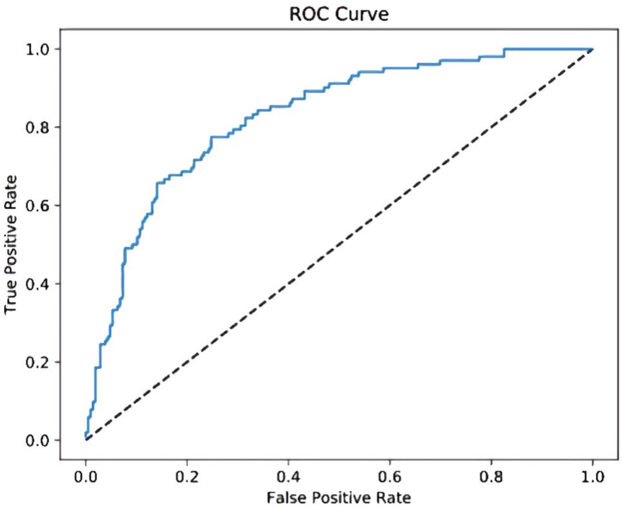


Figure 6. ROC curve.

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

In general, the precision and recall contradict each other. The higher the precision is, the lower the recall; when the recall is high, the precision is often low. To comprehensively consider the precision and recall, the F value is introduced to more comprehensively evaluate the classification performance of a classifier. F is the weighted harmonic average based on precision and recall, and the equation is as follows Eq. (23):

$$\begin{aligned} \frac{1}{F} &= \frac{1}{1+\beta^2} \cdot \left(\frac{1}{Precision} + \frac{\beta^2}{Recall} \right) \\ F &= \frac{\left(1+\beta^2\right) \times Precision \times Recall}{\left(\beta^2 \times Precison\right) + Recall} \end{aligned} \tag{23}$$

where β reflects the relative importance of precision to recall. When $\beta = 1$, that is, the precision is as important as the recall, the F value is the commonly used F1 value. Its equation is Eq. (24):

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{24}$$

The ROC curve sorts the samples according to the prediction results of the classifier, forecasts the samples one by one as positive examples in this order, and draws the ROC curve as shown in Fig. 6 with "FP rate" as the abscissa and "TP rate" as the ordinate, and the AUC value is the area under the ROC curve. The AUC value can directly evaluate the quality of the classifier. The larger the AUC value is, the better the classifier performance.

Model prediction results. Using the feature subset (including 17 features) after RF-RFE feature selection, a classification model is constructed to predict the satisfaction of passengers in the test set, including KNN, LR, GNB and RF. During BPNN training, the activation function is set as "ReLU", the L2 penalty (regularization term) parameter is 0.0001, and the solver is the default "Adam", which can work well in terms of training time and verification score in the face of relatively large data sets.

This study uses five evaluation indexes, accuracy, precision, recall, F1 value and AUC value, to compare the classification performance of the classifier vertically and horizontally. Table 3 shows the classification performance of each classification model before and after RF-RFE feature selection.

Among all classifiers, the RF model based on RF-RFE feature selection performs best, and the five indexes (accuracy, precision, recall, F1 value and AUC value) are greater than those of the other classifiers. The indexes (except precision) of the KNN model increased after RF-RFE. The five indexes of the logistic model decreased slightly after RF-RFE. After RF-RFE feature selection, the two indexes (accuracy and precision) of the GNB model increase, and the other indexes are slightly lower than those of the model without feature selection. After feature selection, the indexes (except recall) of the BPNN model are slightly reduced. In general, in the overall comparison of the five classification models, RF is better than BPNN, followed by KNN and logistics, and GNB is the worst model.

	Models	Accuracy	Precision	Recall	F1 score	AUC
No RF-RFE	KNN	0.930	0.944	0.890	0.917	0.925
	LR	0.873	0.867	0.835	0.851	0.869
	GNB	0.865	0.861	0.821	0.841	0.860
	RF	0.962	0.972	0.940	0.955	0.960
	BP	0.959	0.964	0.940	0.952	0.957
RF-RFE	KNN	0.934	0.942	0.903	0.922	0.930
	LR	0.872	0.865	0.834	0.849	0.867
	GNB	0.866	0.863	0.819	0.840	0.860
	RF	0.963	0.973	0.942	0.957	0.961
	BP	0.954	0.936	0.960	0.948	0.955

Table 3. Model evaluation results.

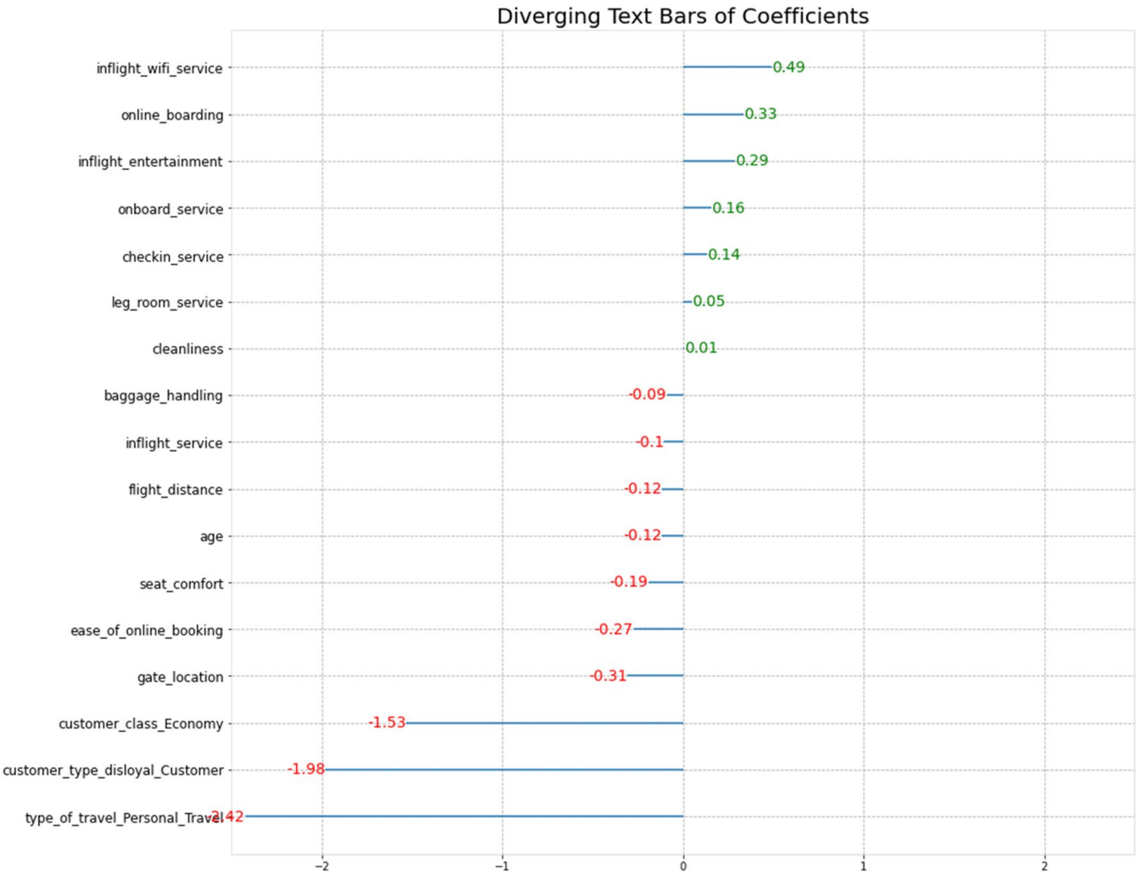


Figure 7. LR coefficient.

Discussion on important variables

In this study, the most important factors affecting passenger satisfaction were selected by the feature selection method. In "Model prediction" section, we used RF-RFE to make a preliminary selection of the feature set and initially selected 17-dimensional features from 22-dimensional features, but the number of features was still large. Therefore, we used the logistic regression model to further select the features in 4.1, and the features with the largest coefficients were further analyzed.

Extracting variables from logistic model. The LR model is constructed based on RF-RFE feature selection, and further feature extraction is carried out through LR. Figure 7 shows the coefficient of the LR variable. The results show that except for the cleanliness of the variable, all other variables passed the significance test. In LR, when the features are standardized, the greater the absolute value of the coefficient, the more important the feature is. If the coefficient is positive, the characteristic is positively correlated with the probability that the output is 1; in contrast, it is positively correlated with the probability that the output is 0. Among the 17 variables

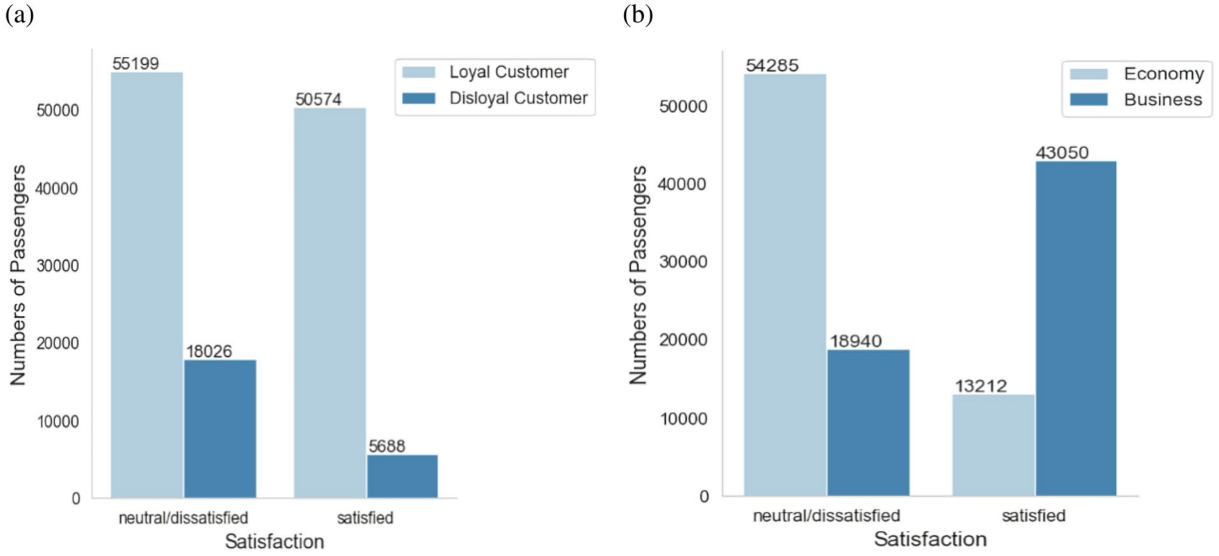


Figure 8. Distribution of people by customer type: (**a**) by passenger type; (**b**) by customer class.

after RF-RFE feature extraction, the first 5 variables with the largest absolute value of the coefficient are type_ of_ travel_ personal travel, customer_ type_ disloyal customer, customer_ class_ economy, inlight_ Wi-Fi_ service, and online_ boarding.

Analysis and suggestions. In airline competition, service quality is the key to winning the choice of passengers. Airlines should adopt a customer-oriented service evaluation method to improve service by understanding which service strategies are the most effective strategies to win more passengers. Based on the big data of airline passengers, the research results of this study give the priority of services that airlines need to pay attention to and give suggestions for services with higher priority through the extraction of important variables.

The five important variables extracted from RF-RFE-logistics are travel type (personal travel and business travel), customer type (loyal and not-loyal), customer class (economy class and business class), inflight Wi-Fi service and online boarding. The travel type is uncontrollable, so we further analyze the three aspects of customer type and customer class, in-flight wi-fi, and online boarding.

Customer type and customer class. Figure 8 shows the distribution of satisfaction according to different passenger types and different cabin types. As seen from the figure, loyal customers account for more than 80%. Among loyal customers, the number of satisfied passengers is close to that of neutral or dissatisfied passengers. In the 1980s, an airline first proposed the Frequent Flyer Program (FFP). In 1994, Air China first implemented it in China. It is a plan proposed by airlines to reduce the risk of losing business and the loss of an existing customer base. It usually includes selling points and mileage to project partners to offset air tickets, upgrades or other rewards. For more advanced members, there are additional points as an incentive for high-value passengers. Since the cost of developing a new customer is often higher than that of maintaining a loyal old customer, FFP members of airlines are rapidly becoming the core competitiveness of business development. Therefore, airlines should invest more energy to improve frequent flyer plans, improve passenger satisfaction of loyal users and increase user stickiness.

When the satisfaction level is divided according to customer class, it can be found that the overall satisfaction of economy class passengers is significantly lower than that of business class passengers. Most business class passengers are satisfied with the service, while most economy class travelers tend to be neutral or dissatisfied with the service.

According to the customer class, 14 variables related to satisfaction, such as seat comfort and online boarding service, are compared to determine the main factors affecting the difference in passenger satisfaction of different class types. Figure 9 radar chart shows the average satisfaction score of 14 variables for passengers in different classes. The average scores of business class passengers and economy class passengers on seat comfort, leg extension space, on-board entertainment and online boarding are very different, but there is no significant difference in food and drinks.

Inflight Wi-Fi service. The result of RF-RFE-LR shows that the most important type of service is inflight Wi-Fi service, and its satisfaction score is the lowest among all services. Mobile Internet has become indispensable in daily life. People's diversified lifestyle is in sharp contrast to traditional monotonous entertainment. Passengers' demand for connecting personal electronic devices to the Internet is becoming increasingly stronger. If airlines can provide Wi-Fi services and provide free internet access or reasonable Wi-Fi prices, they can attract more passengers. According to the survey of INMARSAT in 2016, more than half (54%) of passengers will choose Wi-Fi instead of on-board meals. According to the survey in 2018, onboard Wi-Fi is listed as the fourth

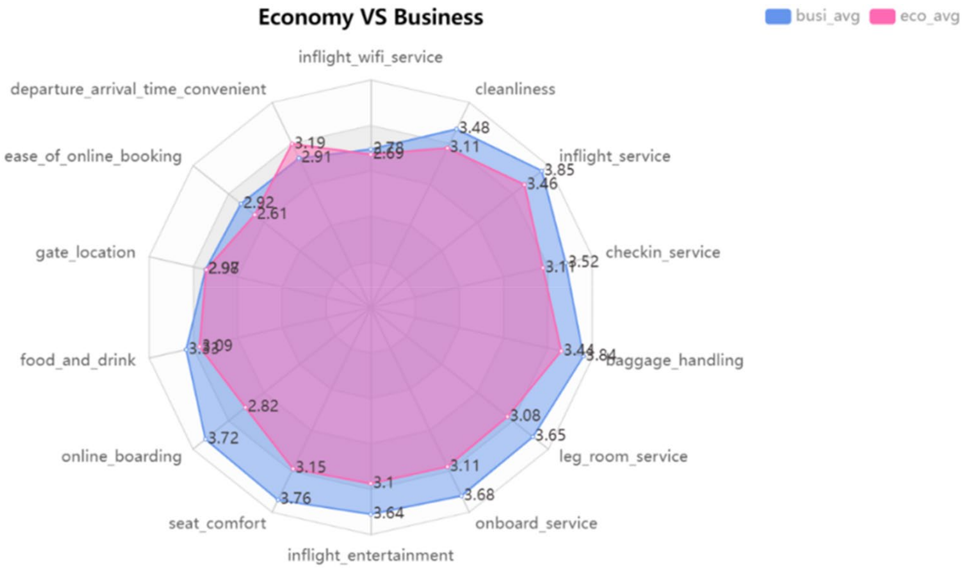


Figure 9. Average score of satisfaction.

most important factor when passengers choose airlines around the world after airline reputation, freely checked baggage and additional leg space. If the flight provides high-quality Wi-Fi, 67% of passengers said they would rebook the airline ticket. In addition, passengers carrying their own personal electronic equipment can also reduce the investment and maintenance of cabin entertainment equipment such as electronic displays and other hardware. The improvement of cabin facilities should be combined with the development of the times. Increasing the popularity of cabin Internet is the general trend. Therefore, from the perspective of improving passenger satisfaction, airlines have good reasons to provide and upgrade onboard Wi-Fi services.

Online boarding. Furthermore, the service that needs to be improved is online boarding. CAAC proposed the concept of "paperless travel" at the 2018 national civil aviation work conference, replacing paper boarding passes with paperless forms such as electronic boarding passes. With ID cards and e-tickets, you can check in directly on the Internet platform without printing boarding passes. It can not only realize the paperless boarding system but also save passengers waiting in line at the counter for check-in, save passengers' time and energy, and make the boarding process faster and more convenient. Therefore, in addition to providing ticket booking, seat reservation, change reservation and meal ordering services, the airline travel service app can add the service of check-in and printing electronic boarding passes in the design, optimize the user boarding experience, improve the competitiveness of the airline travel service app and increase user stickiness. Currently, due to the development of artificial intelligence, the "one card mode" combined with face recognition is possible. Passengers only need to use the second-generation ID card and face recognition camera to realize the unification of ticket information to make each process of passengers at the airport more efficient and improve their travel experience.

Conclusion

We proposed a RF-RFE-Logistic feature selection model to extract the influencing factors of passenger satisfaction based on airline passenger satisfaction survey information. We first use the RF-RFE algorithm to preliminarily extract a feature subset containing 17 variables and use a variety of classification learning algorithms to predict passenger satisfaction. RF on this feature subset shows the best classification performance (accuracy: 0.963, precision: 0.973, recall: 0.942, F1 value: 0.957, AUC value: 0.961). Then, we use a logistic model trained on the feature subset selected by RF-RFE to further extract the important variables affecting passenger satisfaction. Finally, the satisfaction of different passenger types and class types are compared and analyzed, and suggestions are given from the perspectives of online boarding and onboard Wi-Fi services.

There are also some deficiencies in this study. The evaluation indicators of passenger satisfaction surveys in the data set used in this study are not sufficient. In addition, we used the default parameters in the prediction model and did not consider the prediction results of different parameters. Based on the above limitations, we suggest that: (1) the ground service can also include aspects such as baggage claim speed, transfer service, etc. In addition, satisfaction level of related services under abnormal flight conditions can also be added; (2) pay attention to the optimized model parameters; (3) several other variables also affect passenger satisfaction to a certain extent, which should not be completely ignored.

Data availability

Data and methods used in the research have been presented in sufficient detail in the paper.

Received: 5 December 2021; Accepted: 8 June 2022
Published online: 01 July 2022

References

1. Chen, S. *et al.* Airlines content recommendations based on passengers' choice using Bayesian belief networks. In *Bayesian Inference* (ed. Tejedor, J. P.) (IntechOpen, 2017). https://doi.org/10.5772/intechopen.70131.
2. Dolnicar, S., Grabler, K., Grün, B. & Kulnig, A. Key drivers of airline loyalty. *Tour. Manag.* **32**, 1020–1026. https://doi.org/10.1016/j.tourman.2010.08.014 (2011).
3. Jiang, H. & Zhang, Y. An investigation of service quality, customer satisfaction and loyalty in China's airline market. *J. Air Transp. Manag.* **57**, 80–88. https://doi.org/10.1016/j.jairtraman.2016.07.008 (2016).
4. Ok, S., Suy, R., Chhay, L. & Choun, C. Customer satisfaction and service quality in the marketing practice: Study on literature review. *Asian Themes Soc. Sci. Res.* **1**, 21–27. https://doi.org/10.33094/journal.139 (2018).
5. Yeung, M. C. & Ennew, C. T. From customer satisfaction to profitability. *J. Strateg. Market.* **8**, 313–326. https://doi.org/10.1080/09652540010003663 (2000).
6. Williams, P. & Naumann, E. Customer satisfaction and business performance: A firm-level analysis. *J. Serv. Market.* https://doi.org/10.1108/08876041111107032 (2011).
7. Zhang, W. Research on Customer Satisfaction of China Southern Airlines. Diss. Changsha: Hunan University. https://doi.org/10.7666/d.y2066030 (2011).
8. Sankaranarayanan, H. B., Vishwanath, B. V. & Rathod, V. An exploratory analysis for predicting passenger satisfaction at global hub airports using logistic model trees. In *2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*. 285–290. https://doi.org/10.1109/ICRCICN.2016.7813672 (IEEE, 2016).
9. Kumar, S. & Zymbler, M. A machine learning approach to analyze customer satisfaction from airline tweets. *J. Big Data* **6**, 1–16. https://doi.org/10.1186/s40537-019-0224-1 (2019).
10. Cronin, J. J. Jr., Brady, M. K. & Hult, G. T. M. Assessing the effects of quality, value, and customer satisfaction on consumer behavioral intentions in service environments. *J. Retail.* **76**, 193–218. https://doi.org/10.1016/S0022-4359(00)00028-2 (2000).
11. Park, J. W., Robertson, R. & Wu, C. L. The effect of airline service quality on passengers' behavioural intentions: A Korean case study. *J. Air Transp. Manag.* **10**, 435–439. https://doi.org/10.1016/j.jairtraman.2004.06.001 (2004).
12. Jiang, H. An investigation of airline service quality and passenger satisfaction–the case of China Eastern Airlines in Wuhan region. *Int. J. Aviat. Manag.* **2**, 54–65. https://doi.org/10.1504/IJAM.2013.053048 (2013).
13. Hu, K. C. & Hsiao, M. W. Quality risk assessment model for airline services concerning Taiwanese airlines. *J. Air Transp. Manag.* **53**, 177–185. https://doi.org/10.1016/j.jairtraman.2016.03.006 (2016).
14. Chow, C. K. W. Customer satisfaction and service quality in the Chinese airline industry. *J. Air Transp. Manag.* **35**, 102–107. https://doi.org/10.1016/j.jairtraman.2013.11.013 (2014).
15. Etemad-Sajadi, R., Way, S. A. & Bohrer, L. Airline passenger loyalty: The distinct effects of airline passenger perceived pre-flight and in-flight service quality. *Cornell Hosp. Q.* **57**, 219–225. https://doi.org/10.1177/1938965516630622 (2016).
16. Suzuki, Y. Modeling and testing the "two-step" decision process of travelers in airport and airline choices. *Transp. Res. Part E Logist. Transp. Rev.* **43**, 1–20. https://doi.org/10.1016/j.tre.2005.05.005 (2007).
17. Tsafarakis, S., Kokotas, T. & Pantouvakis, A. A multiple criteria approach for airline passenger satisfaction measurement and service quality improvement. *J. Air Transp. Manag.* **68**, 61–75. https://doi.org/10.1016/j.jairtraman.2017.09.010 (2018).
18. Hess, S., Adler, T. & Polak, J. W. Modelling airport and airline choice behaviour with the use of stated preference survey data. *Transp. Res. Part E Logist. Transp. Rev.* **43**, 221–233. https://doi.org/10.1016/j.tre.2006.10.002 (2007).
19. Lucini, F. R., Tonetto, L. M., Fogliatto, F. S. & Anzanello, M. J. Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews. *J. Air Transp. Manag.* **83**, 101760. https://doi.org/10.1016/j.jairtraman.2019.101760 (2020).
20. Brochado, A., Rita, P., Oliveira, C. & Oliveira, F. Airline passengers' perceptions of service quality: Themes in online reviews. *Int. J. Contemp. Hosp. Manag.* https://doi.org/10.1108/IJCHM-09-2017-0572 (2019).
21. Guyon, I. *et al.* (eds) *Feature Extraction: Foundations and Applications* 207 (Springer, 2008).
22. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422. https://doi.org/10.1023/A:1012487302797 (2002).
23. Marcelo, M. C. *et al.* Fast inline tobacco classification by near-infrared hyperspectral imaging and support vector machine-discriminant analysis. *Anal. Methods* **11**, 1966–1975. https://doi.org/10.1039/C9AY00413K (2019).
24. Wei, Y., Shutao, L. & Mingkui, T. Gene selection method based on SVM-RFE-SFS. *Chin. J. Biomed. Eng.* **29**, 93–99. https://doi.org/10.3969/j.issn.0258-8021.2010.01.015 (2010).
25. Gregorutti, B., Michel, B. & Saint-Pierre, P. Correlation and variable importance in random forests. *Stat. Comput.* **27**, 659–678. https://doi.org/10.1007/s11222-016-9646-1 (2017).
26. Wu, C., Liang, J., Wang, W. & Li, C. Random forest algorithm based on recursive feature elimination method. *Stat. Decis.* **21**, 60–63. https://doi.org/10.13546/j.cnki.tjyjc.2017.21.014 (2017).
27. Chen, Q., Meng, Z., Liu, X., Jin, Q. & Su, R. Decision variants for the automatic determination of optimal feature subset in RF-RFE. *Genes* **9**, 301. https://doi.org/10.3390/genes9060301 (2018).
28. Shang, Q. *et al.* Traffic incident detection based on variable selection and kernel extreme learning machine. *J. ZheJiang Univ. (Eng. Sci.)* **51**, 1339–1346 (2017).
29. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
30. Khoda Bakhshi, A. & Ahmed, M. M. Real-time crash prediction for a long low-traffic volume corridor using corrected-impurity importance and semi-parametric generalized additive model. *J. Transp. Saf. Secur.* https://doi.org/10.1080/19439962.2021.1898069 (2021).
31. Zhang, M. L. & Zhou, Z. H. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognit.* **40**, 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019 (2007).
32. Hosmer, D. W. Jr., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression* 398 (Wiley, 2013).
33. Ontivero-Ortega, *et al.* Fast Gaussian Naïve Bayes for searchlight classification analysis. *Neuroimage* **163**, 471–479. https://doi.org/10.1016/j.neuroimage.2017.09.001 (2017).
34. Shi, T. & Horvath, S. Unsupervised learning with random forest predictors. *J. Comput. Graph. Stat.* **15**, 118–138. https://doi.org/10.1198/106186006X94072 (2006).
35. Wang, L., Zeng, Y. & Chen, T. Back propagation neural network with adaptive differential evolution algorithm for time series forecasting. *Expert Syst. Appl.* **42**, 855–863. https://doi.org/10.1016/j.eswa.2014.08.018 (2015).

Author contributions

Conceptualization, X.J. and Y.Z.; methodology, X.J. and B.Z.; formal analysis, Y.Z.; data curation, X.J.; supervision, Y.Z. and Y.L.; writing—original draft preparation, Y.Z. and B.Z.; writing—review and editing, Y.L.

Funding

The research is supported by Hubei Province Education Science Planning Project (2021GB010), and by philosophy and social sciences research project of Hubei Provincial Department of Education (21G032), and by “the Fundamental Research Funds for the Central Universities”, Zhongnan University of Economics and Law (2722022DG002).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.Z.