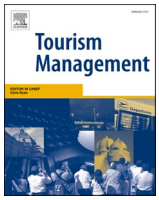
Tourism Management 87 (2021) 104357

Contents lists available at ScienceDirect

Tourism Management

journal homepage: www.elsevier.com/locate/tourman



Customer purchase forecasting for online tourism: A data-driven method with multiplex behavior data

Shui-xia Chen [a,b], Xiao-kang Wang [a], Hong-yu Zhang [a], Jian-qiang Wang [a,*], Juan-juan Peng [c]

[a] School of Business, Central South University, Changsha, 410083, PR China
[b] Business School, Sichuan University, Chengdu, 610064, PR China
[c] School of Information, Zhejiang University of Finance & Economics, Hangzhou, 310018, China

ARTICLE INFO

Keywords:
Online tourism purchase forecasting
Behavior data analysis
Machine learning
Result interpretation

ABSTRACT

Online tourism has received increasing attention from scholars and practitioners due to its growing contribution to the economy. While related issues have been studied, research on forecasting customer purchases and the influence of forecasting variables, online tourism is still in its infancy. Therefore, this paper aims to develop a data-driven method to achieve two objectives: (1) provide an accurate purchase forecasting model for online tourism and (2) analyze the influence of behavior variables as predictors of online tourism purchases. Based on the real-world multiplex behavior data, the proposed method can predict online tourism purchases accurately by machine learning algorithms. As for the practical implications, the influence of behavior variables is ranked according to the predictive marginal value, and how these important variables affect the final purchase is discussed with the help of partial dependence plots. This research contributes to the purchase forecasting literature and has significant practical implications.

1. Introduction

Tourism services have become one of the largest industries contributing to the economy (Hu et al., 2021). Taking Chinese market as an example, the total tourism revenue reached 6.63 trillion yuan in 2019. The Internet is becoming an important distribution channel for tourism, generating the transaction volume of China's online tourism market of approximately 1005.9 billion yuan in 2019. However, online tourism is not a panacea for the sustained economic prosperity because its viability depends on the uncertainty in a purchase. First, the uncertainty of tourism purchases stems from the characteristics of services: invisibility, heterogeneity, inseparability, and perishability, making it difficult for customers to evaluate quality when purchasing (Moeller, 2010). Studies also signify that purchase uncertainty was exaggerated in the e-marketplace, as customers had no physical interaction with the products before their purchases. In addition, there is a vast amount of information on the Internet, and as a result online consumer is easily confused by over-choice and not sure what to purchase.

In this regard, the analysis on purchase forecasting is required for online tourism managers to understand customer behavior and reduce the impact of uncertainty in purchases. For one thing, the results of purchase forecasting help online tourism managers reduce the gap between businesses and customers (İçer, Parmaksız & Ç, 2021), as it can provide valuable information to target customers in time to avoid uncertainty when shopping online. For another, in addition to the accurate forecasting results, Van Nguyen et al. (2020) argued that the further analysis on the interrelationship between influencing variables and prediction could reveal some valuable marketing strategies to conduct business activities. For example, customers are often puzzled by the high degree of similarities between tourism services on the Internet. The analysis of purchase variables can provide understandable information that customers focus on, thereby avoiding confusion caused by the similar information. Considering these, this paper proposes a customer purchase forecasting model for online tourism to obtain the accurate prediction result and the influence of variables as predictors.

In the existing marketing research, empirical and data-driven methods are widely used for customer purchase forecasting. Empirical studies aim to reveal factors that affect customer purchases by methods such as surveys. However, rigorous design and proper procedures are required in the empirical analysis (Xiao et al., 2016). In addition, these empirical methods are usually based on descriptive analyses such as the summary statistics. Further processing of purchase forecasting uses data-driven methods to obtain final results, as there are rich customer behavioral data in e-marketplaces. These online behavioral data contain

* Corresponding author.
E-mail address: jqwang@csu.edu.cn (J.-q. Wang).

https://doi.org/10.1016/j.tourman.2021.104357
Received 25 August 2020; Received in revised form 2 April 2021; Accepted 18 May 2021
Available online 6 June 2021

valuable information related to customer purchases and are readily available. Therefore, data-driven methods have become mainstream in recent years. This paper concludes recent data-driven studies in forecasting objectives, behavioral data type, and forecasting model, as shown in Table 1.

From Table 1, regarding the objective of forecasting, most researches studied the purchase forecasting of online products, while only a few studies focused on online tourism purchases. However, the tourism has unique features that differentiate it from products and other services. First, unlike online products, customers have difficulty evaluating tourism services with a uniform standard. Furthermore, tourism consumption usually takes longer to plan and costs more than many other types of services (Xu & Gursoy, 2015). When purchasing tourism services, online consumers try to gather as much information as possible to make final decisions. That's to say, the purchase of online tourism would not be the same as general online products and services. Exploring the impact of online behavioral data on online tourism purchases is essential. Therefore, based on the existing research, this paper proposes a data-driven purchase forecasting model for online tourism.

Data-driven purchasing forecasting focuses on analyzing the purchase determinant from online behavioral data. In terms of online tourism, Amaro and Duarte (2015) emphasized the importance of purchase determinants such as customer attitude, perceived risk, and trust. And these determinants can be delivered by analyzing the historical customer behavior data with the help of information technology. As shown in Table 1, some structured data, including operational behavior data and demographic information, are widely used to extract purchase determinants in existing studies. However, the available behavioral data are limited in online tourism, partly because tourism services are browsed and purchased relatively infrequently (Chen, Wu, et al., 2020). Moreover, some purchase data involving customer privacy are difficult to obtain. The availability of unstructured behavioral data such as online reviews offers the promise of an alternative means to extract purchase determinants (Chen, Zhang, et al., 2020). Especially for online tourism, as an experiential product, reviewers rely more on the review content to state their experience. As shown in Table 1, however, few studies focused on the analysis of unstructured data. Therefore, it is necessary to explore the effect of multiplex behavior data (including structured and unstructured data) on customer purchases to obtain valuable purchase determinants for online tourism.

The next step in terms of customer purchase forecasting is to determine the forecasting model. Existing forecasting methods can be classified into two categories, as shown in Table 1: methods with and without hypotheses. The statistical method is the representative forecasting method with the model hypothesis. This method achieves forecasting by determining the relationship assumptions between input-output variables in advance (Van Nguyen et al., 2020). However, in the real business e-marketplace, the relationships between variables are often difficult to describe in advance. In addition, rich information in the e-marketplace remains to be extracted, which may be too complex to be represented by prior simulation. Therefore, some forecasting methods without hypotheses, such as machine learning, are introduced into purchase forecasting. Many studies have confirmed that machine learning methods could achieve high purchase forecasting accuracy (Dou, 2020). However, it is well known that machine learning has shortcomings in result interpretability. Models lacking interpretability have difficulty convincing people and generating understandable managerial implications. As shown in Table 1, most existing studies ignored the interpretation of machine learning forecasting results. Therefore, based on the high accuracy of machine learning, it is necessary to analyze the interpretability of the results to provide some practicable marketing strategy.

To fill the existing research gaps, this paper proposes a data-driven method with multiplex behavioral data to obtain purchase forecasting for online tourism. First, this paper collects multiplex behavioral data, including customer information, operational behavior data, and online reviews. Corresponding methods are then proposed for the variable analysis of behavioral data. Second, considering the efficiency of machine learning algorithms, this study adopts several different machine learning algorithms to verify the extracted variable's validity. Finally, two influencing factor analysis methods, SHapley Additive exPlanations (SHAP) and partial dependence plot (PDP), are adopted to analyze the results of online tourism purchasing forecasting.

The remainder of this paper is organized as follows: Section 2 introduces the background and related work of this study. Section 3 introduces the data description and variable analysis. Section 4 develops the architecture of the proposed forecasting model. Then, this paper describes some experiments in section 5, and the conclusions are given in section 6.

2. Background and related work

As we mentioned above, only a few studies have dealt with purchase forecasting towards online tourism. Still, a data-driven method based on multiplex behavioral data with interpretable machine learning forecasting results has not been proposed. Therefore, this section introduces the background and related work from customer behavioral data analysis and customer purchase forecasting based on machine learning.

2.1. Customer behavioral data analysis

This subsection reviews existing analysis methods for the multiplex

Table 1
Recent studies on data-driven purchase forecasting.

Literature	Objective of forecasting	Behavioral data type		Forecasting model	
		Structured	Unstructured	Model hypothesis	Result interpretation
Baumann et al. (2018)	Purchase probability	✓			
Nishimura et al. (2018)	Purchase probability	✓		✓	
Shapoval and Setzer (2018)	Purchase	✓			
Zhu et al. (2018)	Tourism purchase	✓			
Wen et al. (2018)	Purchase category	✓		✓	
Zhu et al. (2019)	Tourism products	✓			✓
Bag et al. (2019)	Purchase intention	✓	✓	✓	
Dong and Jiang (2019)	Brand purchase	✓		✓	
Kytö et al. (2019)	Purchase	✓			
Park et al. (2019)	Purchase	✓			
Dou (2020)	Product purchase	✓			
Hu et al. (2020)	Buying behavior	✓			
Chu et al. (2020)	Purchase	✓		✓	
Park et al. (2020)	Purchase	✓		✓	
Martinez et al. (2020)	Product purchase	✓			
Chen et al. (2021)	Purchase	✓	✓		

behavioral data including operational behavior data and online reviews.

2.1.1. Operational behavior data analysis method

Operational behavior data refer to the records generated by customers' online behavior, such as browsing, purchasing, and paying. Commonly used operational behavior data analysis methods include machine learning feature extraction methods and statistical description methods. The former adopts some machine learning methods to extract purchase features from operational behavior data. For example, Huang et al. (2019) developed a network architecture to exploit the underlying factors for customer purchase forecasting. Park et al. (2019) developed an encoder-decoder architecture to obtain features related to purchase forecasting. However, it should be noted that the features constructed by machine learning methods are difficult to understand directly. Therefore, statistical description methods that construct features manually are widely used (Martinez et al., 2020). Features obtained in this way can be explained strongly and tend to describe behavioral data more accurately. However, owing to the manual feature construction, the statistical description method is not suitable for the large-scale dataset and dynamic environments.

Therefore, to perform an effective statistical description of operational behavior data, this paper introduces probabilistic linguistic term sets (PLTSs) (Chen et al., 2019) and Newton's cooling law (Cai & White, 2011) to represent and aggregate operational behavior data. PLTS is a probabilistic language description model by which the data dimension can be reduced. The information aggregation method based on Newton's cooling law can achieve information aggregation considering weight difference.

2.1.2. Online review analysis method

Online reviews are the subjective comments given by customers and include numerical ratings and online textual reviews. The existing literatures have developed different analysis methods to address online reviews. For numerical ratings, some studies focused on analyzing rating content (Gavilan et al., 2018) without considering the rating bias of different customers. Rating bias refers to the fact that each customer has his/her rating tendency, with some customers rating strictly and others not. Therefore, to better understand customer ratings, further studies analyzed the rating bias between different customers, i.e., inconsistent scores (Toledo et al., 2015; Xiao et al., 2016). In addition to inconsistent scores, a growing number of studies have found that there were some correlations between ratings and customer psychology (Li et al., 2020). Therefore, to improve the effectiveness of rating-related analysis, the rating feature extraction method should take the psychological influence into consideration.

In terms of online textual reviews, Shao et al. (2014) adopted regression and correlation analysis to study the influence of online reviews on consumer purchase decisions by questionnaire. Chen et al. (2018) adopted a numerical simulation to study how customers determined their purchase decisions under the influence of positive and negative reviews. Most of the existing studies adopted qualitative methods or hypothesis models to analyze the impact of online reviews. Few studies focused on extracting purchase features from online reviews to improve the accuracy of purchase forecasting. In addition, excessive data dimensions occur when adopting the commonly used textual vector describing methods such as word segmentation and word frequency statistics. Textual vectors obtained by these methods may result in complex computation as well as damage the classification and clustering accuracy.

Hence, unlike the existing qualitative studies, this paper proposes a variable analysis method to extract purchase features from online reviews. This paper first eliminates inconsistent scores and proposes a model that considers customer psychology to obtain purchase features related to ratings. Then a customer preference analysis method based on the term frequency-inverse document frequency (TF-IDF) algorithm (Spärck Jones, 2004) is adopted to extract features from online textual reviews.

2.2. Online purchase forecasting based on machine learning

In this subsection, online purchase forecasting based on machine learning, including forecasting models and two methods for result interpretation, is introduced.

2.2.1. Machine learning models for customer purchase forecasting

Machine learning methods have been widely used in purchase prediction due to their high prediction performance. Kagan and Bekkerman (2018) trained a tree-based forecasting model and then applied it to forecast customer purchases. Zhou et al. (2019) established a two-layer model for purchase prediction, in which the second layer was combined with XGBoost (XGB) algorithm. In addition to adopting a single algorithm, Martinez et al. (2020) adopted different machine learning algorithms to obtain purchase prediction results. However, different machine learning methods have different applicable backgrounds. It is difficult to directly determine which machine learning method is optimal for online tourism purchase forecasting. In addition, as discussed above, machine learning methods have shortcomings in result interpretability, and most of the existing studies ignored further analysis.

Therefore, based on the analysis of behavioral data, this paper selects several different types of machine learning methods to predict the purchase of online tourism. Two models for forecasting result interpretation, SHAP and PDP, are introduced to explain the prediction results.

2.2.2. Forecasting result interpretation models: SHAP and PDP

SHAP is a feature analysis method that focuses on the importance of each prediction-related feature (Lundberg & Lee, 2017). In this method, the Shapley value of a feature is regarded as the average marginal contribution of this feature in all feature sequences. SHAP method can well solve the multicollinearity problem. In addition, the feature contributions obtained by SHAP take into account both the influence of individual features and the possible synergies among features. This method has been used to interpret forecasting results and the importance of features in many studies (Parsa et al., 2020).

PDP is a method to analyze the marginal effect of features on the forecasting result. In this method, keep all other variables unchanged, and then change the feature to be analyzed to observe how it affects the forecasting result. Unlike other feature analysis methods that only consider the importance of features, PDP shows whether the relationship between forecasting result and the variable is linear, monotonous or more complex. As an effective method of explaining forecasting results, PDP has been widely used in many studies (Van Nguyen et al., 2020).

3. Data description and variable analysis

3.1. Data description

The dataset used in this paper is provided by a tourism app (https://www.huangbaoche.com), including three types of data: 1) demographic information, such as age, gender, and province; 2) operational behavior data, such as browsing and purchasing; and 3) online review, including numerical ratings and textual reviews. There are two types of tourism services in this app: normal and boutique service. The boutique services are the critical items of the online tourism platform, while the normal services refer to the general services in addition to boutique services. Compared with normal services, the boutique services can bring higher income. Therefore, this study aims to predict whether customers purchase boutique services. A mathematical description of the research problem is given as follows:

Suppose that there are n customers, denoted by $i = \{1, \cdots, n\}$. The operational behavior variables of customer i are represented by actions

$A_{i,j}^q$ and the corresponding time $t_{i,j}^q$, for $q = 1, \cdots, q_i$. Here, q_i signifies the total number of action periods and $j = 1, \cdots, j_q$ indicates the total number of actions in action period q. An action period is defined as all actions from the appearance of the wake-up app (denoted as action 1) to the presence of the next action 1 when the time interval between adjacent actions does not exceed a specific value (such as 3 h). If the time interval exceeds this value, we treat it as missing data and then supplement action 1 between these two actions. The corresponding time for the supplementary action 1 is consistent with the time of the next adjacent action. Numerical ratings and textual reviews are represented by $S_{i,k}$ and $R_{i,k}$. The corresponding order number and order time are denoted as $O_{i,k}$ and $T_{i,k}$, respectively, for $k = 1, \cdots, m_i$. Here, m_i denotes the total order number of *i-th* customer. All the notations below have the same meaning. The research problem can be described as follows:

Given

$U_i = \{(A_{i,j}^q, t_{i,j}^q, S_{i,k}, R_{i,k}, O_{i,k}, T_{i,k}) \big| q = 1, \cdots, q_i, j = 1, \cdots, j_q, k = 1, \cdots, m_i\}$ for each customer i between time τ_a and τ_b, predict whether a given customer purchases the boutique services in the next time. This research adopts the machine learning algorithms to address this problem. The first step is to construct a feature vector $x_{i,\tau}(\tau \le \tau_b)$ to characterize customer i at time τ based on the behavior data. Then, the forecasting value y_{i,τ_b+1} for the upcoming time $\tau_b + 1$ can be obtained based on the actual values $y_{i,\tau}$ and $x_{i,\tau}$. Therefore, the research problem can be reformulated as follows:

Estimate the value y_{i,τ_b+1} from the feature vector x_{i,τ_b+1}, given information for all $i = \{1, \cdots, n\}$ and $\tau \in [\tau_a, \tau_b]$. The solution for this problem requires the help of feature variable extraction $x_{i,\tau}^1, \cdots, x_{i,\tau}^M$. Here, M is the number of variables. Therefore, the following section introduces the analysis of variables related to purchase forecasting.

3.2. Variable analysis related to purchase forecasting

This section introduces variable analysis of the collected multiplex behavioral data including operational behavior data and online reviews.

3.2.1. Variable analysis of operational behavior data

In the collected dataset, the operational behavior data are represented by 9 action types, among which 1 represents the wake-up of the app, 2–4 represent browsing different types of online tourism services, and 5–9 represent actions from filling out to final paying. Since the collected dataset has been desensitized, detailed information about each action is unknown. An example of operational behavior data is shown in Table 2.

Considering that the amount of operational behavior data is too large to be analyzed by statistical description, this paper adopts PLTSs to describe these data. First, actions 2–4 and actions 7–9 are regarded as the same action for calculation due to the disorder of actions 2–4 and the similarity of actions 7–9. This paper then adopts the language scale function (Yu et al., 2018) to convert the acquired action into a simple form. The corresponding relationship between the action and subscript is shown in Table 3. Next, converting the actions to PLTSs can be illustrated by Example 1.

Table 2
Example of operational behavior data.

Period	Actions	Start timestamp	Finish timestamp	Timestamp intervals
1	1, 5, 2, 5, 5, 5, 2, 4	1,496,658,129	1,496,658,455	326
2	1, 5	1,497,821,443	1,497,821,452	9
3	1, 5, 6, 7, 8, 9, 3	1,498,751,517	1,498,752,459	942
4	1, 4, 5, 4, 2, 5, 6	1,498,866,689	1,498,867,027	338
5	1, 5, 6, 4, 4, 5, 6, 7	1,498,871,720	1,498,871,874	154

Table 3
Corresponding relationship between actions and subscripts.

Action	1	2–4	5	6	7–9
Corresponding subscript	1	2	5	6	7

Example 1. Suppose that the $i-th$ customer's operational behavior data are shown in Table 2. Actions in the $q-th$ action period can be denoted as $A_{i,j}^q$ and represented as:

$$A_i^1 = \{S_1(0.125), S_2(0.375), S_5(0.500)\},$$

$$A_i^2 = \{S_1(0.500), S_5(0.500)\},$$

$$A_i^3 = \{S_1(0.143), S_2(0.143), S_5(0.143), S_6(0.143), S_7(0.428)\},$$

$$A_i^4 = \{S_1(0.143), S_2(0.429), S_5(0.286), S_6(0.143)\},$$

$$A_i^5 = \{S_1(0.125), S_2(0.250), S_5(0.250), S_6(0.250), S_7(0.125)\},$$

After obtaining the PLTS representation of operational behavior data, it is necessary to aggregate the information that belongs to the same customer. Considering that the relevance of historical behavior to future purchases decreases over time, this paper utilizes a time decay function called Newton's cooling law mathematical model to describe these decreases. The aggregation method is expressed as follows:

$$\omega_q = \frac{e^{\left(-\alpha\left(T - t_i^q\right)\right)}}{\sum_{q=1}^{q^m} e^{\left(-\alpha\left(T - t_i^q\right)\right)}}, \tag{1}$$

$$O_i = \omega_1 O_i^1 + \omega_2 O_i^2 + \cdots + \omega_{m_q} O_i^{m_q}. \tag{2}$$

where ω_q represents the weight of the action in the $q-th$ action period. Here α is the attenuation constant, which can be calculated by regression analysis. T and t_i^q represent the prediction and the starting time, respectively. This model aggregates information with different weights, and the weight difference is obtained according to the length of the time interval. Therefore, information aggregation considering the time decay can be achieved. After obtaining the representation of operational behavior data, this paper analyzes the variables by referring to the existing statistical description method. A detailed introduction of the obtained variables is as follows:

(1) Variables related to customer profile.
Gender, province, and age of customers.

(2) Variables related to orders.
1) The number of total orders, normal service orders, and boutique service orders;
2) The last order type;
3) The time of the last order;
4) The mean value of the time interval between the adjacent orders.

(3) Variables related to the operational behavior content.
1) The number of total actions, last action type, and the ratio of each action;
2) The statistic value of different combination ratios of any two actions, including the value of the mean, standard deviation, maximum, and minimum;
3) The action number's statistical value after each action in the corresponding action period, including the value of the mean, standard deviation, maximum, and minimum.

(4) Variables related to the operational behavior time.
1) The time of each action in the last action period;

2) The mean time interval between the order and each action of this order period;
3) The time interval statistic of any two adjacent actions, including the value of the mean, standard deviation, maximum, and minimum;
4) The statistical value of the time interval between each action and the last action in each action period: including the value of the mean, standard deviation, maximum, and minimum.
5) The time interval between the last 10 actions.

3.2.2. Variable analysis of online reviews

(1) Variable analysis of numerical rating.

In the tourism app, customers can feedback on their satisfaction level with the tourism services based on a Likert scale, usually on a 5-point scale. Considering the inconsistent score in the numerical rating, this paper adopts the following model to eliminate rating bias:

$$\mu_i = \frac{\sum_{k=1}^{m_i} S_{i,k}}{m_i}, \tag{3}$$

$$s_{i,k} = S_{i,k} + \lambda \frac{\sum_{k=1}^{m_i} (S_{i,k} - \mu_i)}{m_i}. \tag{4}$$

where $S_{i,k}$ is the raw rating and $s_{i,k}$ is the standardized rating. The parameter λ is related to the customer's rating bias. Selecting a part of the existing ratings as the training set, the parameter λ can be determined by a linear regression model.

After obtaining the standardized rating, this paper further analyzes the rating considering the psychological influence. When making decision, customers usually presuppose a reference point and then measure whether each result is above or below this reference point. According to the risk perception theory, customers tend to show risk aversion for the gain results and risk preferences for loss results. The risk perception difference can be defined by the psychological drive of risk, called prospect theory (Tversky, 1979). A prospect value function $u(x)$ to determine the gain $x \geq 0$ and loss $x < 0$ is established as follows:

$$u(x) = \begin{cases} x^{\alpha} & x \geq 0, \\ -\delta(-x)^{\beta} & x < 0. \end{cases} \tag{5}$$

where α and β represent the risk preference and risk aversion parameters of gains and losses, respectively. Here, δ represents the coefficient of loss aversion. The prospect value function is shown in Fig. 1.

Generally, on a 5-point scale, a rating greater than 3 points is considered favorable, a rating less than 3 points is considered a loss, and a rating of 3 points is considered neutral. According to the prospect theory, there is risk perception difference when customers give different ratings, where they show risk aversion for favorable ratings and risk preferences for loss ratings. Considering this, we therefore introduce the prospect theory into the variable analysis of numerical ratings to capture the risk perception difference. According to Eq. (5), the customer risk aversion and risk preference for different ratings can be calculated by

$$u(s_{i,k}) = \begin{cases} (s_{i,k} - 3)^{\alpha} & s_{i,k} \geq 3, \\ -\delta(3 - s_{i,k})^{\beta} & s_{i,k} < 3. \end{cases} \tag{6}$$

Based on nonlinear regression, Kahneman (1992) analyzed these parameters and found that the median values of α and β were 0.88, while the median value of δ was 2.25. Therefore, $\alpha = \beta = 0.88$ and $\delta = 2.25$ are set in this paper. Then, the rating variable that considers the customer's risk perception can be obtained.

(2) Variable analysis of online textual review.

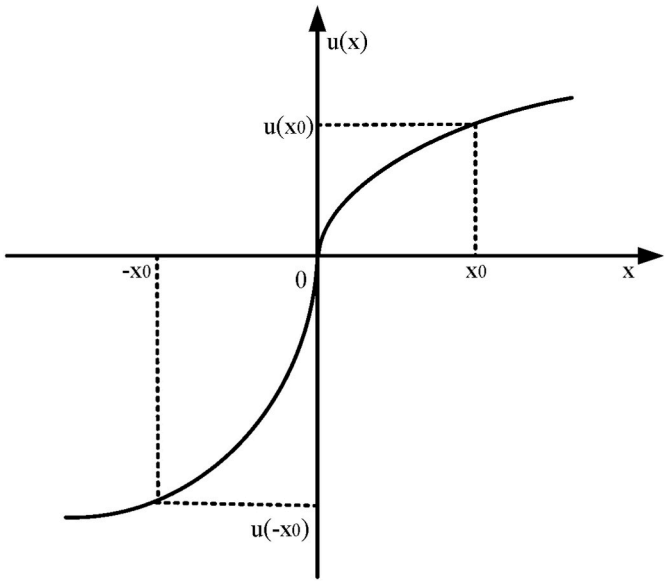


Fig. 1. Prospect value function.

Considering the excessive data dimensions caused by the existing textual analysis methods, a textual preference extraction model is used in this paper to obtain purchase-related variables (Chen et al., 2021). The steps of the preference analysis method are as follows:

Step 1 Text preprocessing.

Since there are only Chinese reviews in the collected dataset, data cleaning, tokenization, and stop word removal are adopted for text preprocessing. An open-source Python package (Jieba, https://github.com/fxsjy/jieba) and a frequently used stop word dictionary are adopted to complete word segmentation and stop word removal, respectively.

Step 2 Textual features extraction.

The commonly used TF-IDF algorithm is adopted to extract the service features, and the obtained information includes the keywords of the service features and the corresponding weights.

Step 3 Textual features preference analysis.

First, establish the item feature matrix W of different services, including the feature keywords and the corresponding weights. W_{kj} is the relationship strength between item a_k and feature x_j, that is, the opinion of most customers on the item a_k.

Second, construct the customer feature matrix I based on the features of the services. I_{ij} is the preference relationship strength of the customer for feature x_j. To construct matrix I, we first calculate the probability of these selected feature keywords in each review. Then, n customer feature matrices including the feature keywords and probability can be obtained, where n is the number of reviews. The comprehensive customer feature matrix can be obtained by aggregating each feature matrix of the same customer.

Finally, customer preference can be obtained by calculating the similarity between matrices W and I.

4. Architecture of the proposed forecasting model

This section introduces the architecture of the proposed online tourism purchase forecasting model, as shown in Fig. 2. This architecture is designed following Cross Industry Standard Process and Data

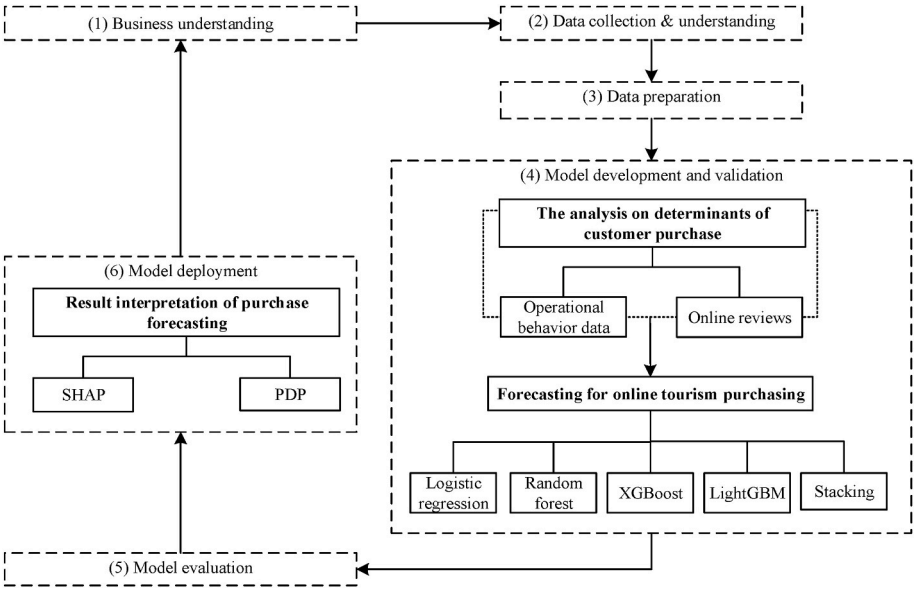


Fig. 2. Architecture of the proposed forecasting model.

Mining framework (Oztekin, 2016). First, business understanding represents proposing the research objective, i.e., research on online tourism purchase forecasting in this paper. Second, data collection and understanding refer to collecting and understanding the dataset related to the research objective. Third, data preparation is adopted to obtain the normalized data. In this paper, missing values of demographic information are filled in with 0 for the calculation. Additionally, the provinces are divided into three tiers according to Chinese city classification for the ease of calculation. The preparation of the operational behavior data and online reviews are conducted according to the methods in section 3.2.

Then, model development and validation, including the analysis of determinants and the forecasting of online tourism purchasing, are conducted. The steps of the analysis on determinants of customer purchase are as follows:

Step 1 This study adopts the variable analysis method proposed in section 3.2.1 to obtain the operational behavior data variables.

Step 2 Eqs. (3) and (4) are adopted to eliminate the rating bias. In addition, rating-related variables that consider customer risk perception are obtained by Eq. (6).

Step 3 With regard to the textual data, this paper sets the parameter topK of TF-IDF algorithm to 20. The results, including the feature keywords and corresponding weights, can be obtained as listed in Table S1 of Supplementary_A. Customer preference features can be obtained by the variable analysis method of online textual review in section 3.2.2.

The obtained variables are then introduced into prediction models to obtain the final results. Some commonly used machine learning algorithms are adopted to perform the predictions, including logistic regression (LR), random forest (RF), XGB, LightGBM (LGB), and stacking.

- Logistic regression.

LR (Conklin, 2002) is a generalized linear regression analysis algorithm. By adopting the sigmoid function $g()$, the actual value of the classification can be associated with the predicted value, and the prediction function of LR is:

$$h_\theta(x) = g\left(\theta^T x\right) = 1 \Big/ \left(1 + e^{-\theta^T x}\right), \tag{7}$$

where θ represents the combination of parameters remaining to be calculated. Then, the cost functions of LR are derived based on the maximum likelihood estimation:

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & if \ \ y = 1 \\ -\log(1 - h_\theta(x)) & if \ \ y = 0 \end{cases} \tag{8}$$

$$J(\theta) = 1/m \sum_{i=1}^{m} Cost\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) = -1/m \sum_{i=1}^{m} \left(y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right) \tag{9}$$

where $J(\theta)$ represents the final cost function. And the forecasting result can then be obtained by parameter optimization algorithms such as the gradient descent method and conjugate gradient method.

- Random forest.

The basic unit of RF (Biau & Scornet, 2016) is the decision tree, and this algorithm belongs to a branch of machine learning-integrated methods. The final output of RF is obtained by integrating the output of multiple trees. Unlike the normal decision tree model, RF first obtains the base decision tree node and randomly selects k elements. Then, the classification can be conducted based on the obtained subset of elements. Supporting that there are N training samples with M features, the steps of RF for classification are as follows:

First, select m features to determine the results of decision tree nodes, where m is much less than M. Second, a training set is obtained from N training samples with N return sampling, and the rest are taken as the test sets. Finally, m features are selected randomly,

and the decision value of each node is determined based on the selected features. Then, the best splitting method is calculated according to these m features.

- **XGBoost and LightGBM.**

XGB (Chen & Wang, 2016) is a scalable end-to-end tree boosting system. In XGB, the algorithm continually adds trees and constantly morphs features to grow a tree. Each time a tree is added, a new function is learned to fit the last predicted residuals. The objective function of XGB is:

$$L = \sum_i l\left(\hat{y}_i - y_i\right) + \sum_k \Omega(f_k). \tag{10}$$

The objective function consists of two parts: the error of the model $l(\hat{y}_i - y_i)$ and the regularization term $\Omega(f_k)$. The first part is used to measure the difference between the prediction value $\hat{y}_i$ and the true value y_i. The other part characterizes the complexity function of the tree. In $\Omega(f_k)$, k is the number of trees and f_k represents the model of the k-*th* tree. The calculation expression of $\Omega(f_k)$ is:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\left\|\omega^2\right\|. \tag{11}$$

The regularization term also contains two parts, where T is the number of leaf nodes and ω denotes the leaf node score. γ controls the number of leaf nodes and λ regulates the leaf nodes scores to prevent overfitting. The XGB algorithm is widely used because of its advantages, such as supporting parallelization and adding sparse data processing.

LGB (Wang et al., 2017) is a distributed gradient lifting framework based on a decision tree. Unlike XGB, LGB uses a leafwise growth strategy for a decision tree, each time finding a leaf with the highest split gain from the current leaf to split to achieve the loop's growth. The comparisons between XGB and LGB are shown in Table 4. From this table, it can be obtained that the tree split search of LGB is based on the histogram algorithm. Therefore, LGB occupies less memory and has lower complexity of data separation. In addition, LGB supports cache hit ratio optimization and categorical features.

- Stacking

Stacking (Wagner et al., 2019) is a process of model hierarchical fusion. Taking the two-level stacking as an example, the basic flow is shown in Fig. 3. In this method, multiple classifiers (classifiers 1 of Fig. 3) are initially trained to obtain the output of each classifier. Then a new classifier (classifier 2 of Fig. 3) that adopts the former output results as the input is trained. Moreover, the final result is outputted. LR model is often used as the classifier 2 in Fig. 3. Stacking has been widely used in various algorithm games and has achieved good prediction performance.

Next, as shown in Fig. 2, the step of model evaluation measures the performance of the proposed forecasting model. Finally, the step of model deployment refers to the result interpretation of the purchase forecasting. In turn, the valuable result interpretation can help to understand the business objective.

Table 4
Comparisons between XGB and LGB.

	XGB	LGB
Tree growth	Level-wise	Leafwise with max depth limitation
Split search	Presorted algorithm	Histogram algorithm
Memory cost	2*#feature*#data*4Bytes	#feature*#data*1Bytes
Calculation of split gain	O (#data*#features)	O (#bin*#features)
Cache-line aware optimization	N/A	40% speed-up on Higgs data
Categorical feature support	N/A	8 speed-up

5. Experiments

This section introduces the proposed model into the collected dataset to forecast online tourism purchases. There are over 40,000 customers in the collected database with a period from September 2016 to September 2017. During this period, 20,653 orders and more than 1 million actions of these customers are recorded. There are 9863 ratings and 3752 textual reviews in the dataset.

Commonly used evaluation methods, including mean absolute error (MAE), accuracy (ACC), precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC), are used to evaluate the forecasting performance of models. During the experiment, the dataset is randomly divided into a training set and test set according to the commonly used 7:3 ratio. The *GridSearchCV* method is adopted to adjust the parameters of RF, XGB, and LGB. Stacking method takes XGB and LGB as the first level classifiers and LR as the second classifier. All of the following experiments are conducted in Python 3.7.

5.1. Results analysis

Online tourism forecasting results obtained by the proposed data-driven methods are shown in Table 5.

First, Table 5 shows that LR is the worst-performing model. This is because there is a significant difference in the number of two samples (i.e., the purchases of boutique service and normal service), with a ratio of 1:7. In this case, the prediction performance of LR is usually poor. In addition to LR, other models can achieve better prediction. ACCs of the other models exceed 90%, which indicates that most models can predict 90% of the samples accurately. AUC of all the forecasting methods except for LR exceed 0.9. The value of AUC is generally between 0.5 and 1, and it is generally believed that the method has high accuracy when AUC is greater than 0.9. These results illustrate the feasibility and effectiveness of the proposed prediction framework. Fig. 4 shows the comparison results of these methods to compare the forecasting performance of the selected methods in addition to LR.

First, it can be seen from Fig. 4 that MAE of stacking is the smallest, followed by XGB, LGB, and RF. Based on the definition of MAE, this result indicates that the average prediction error of stacking is the smallest, and the prediction performance is the best. Secondly, the forecasting accuracy of these four models is more than 93% for ACC, and the best model in terms of ACC is stacking. As seen in Fig. 4, the values of the precision and recall appear to be irregular. The highest precision value is XGB, but LGB exhibits the best performance in the recall. This is because the precision and recall are conflicting performance measures. F1-score that considers the balance of precision and recall can be used to evaluate the final forecasting. The optimal model under F1-score is still stacking. Finally, AUC obtained by the four models are all over 0.94, and the best prediction model is stacking with an AUC of 0.97.

Overall, the above models have good prediction performances from different evaluation measures, which indicates that the proposed prediction model can obtain accurate prediction results. In addition, stacking performs the best among these models.

5.2. Result interpretation and managerial insight discussions

As we discussed above, the interpretation of the model results is as important as the accuracy of forecasting. Therefore, this subsection explores the interpretation of the forecasting results to obtain data-driven managerial implications for online tourism services. Taking the forecasting results of XGB as an example, SHAP is used to obtain variable contributions based on their marginal value. Then, PDP is adopted to identify how these important variables affect the final purchase.

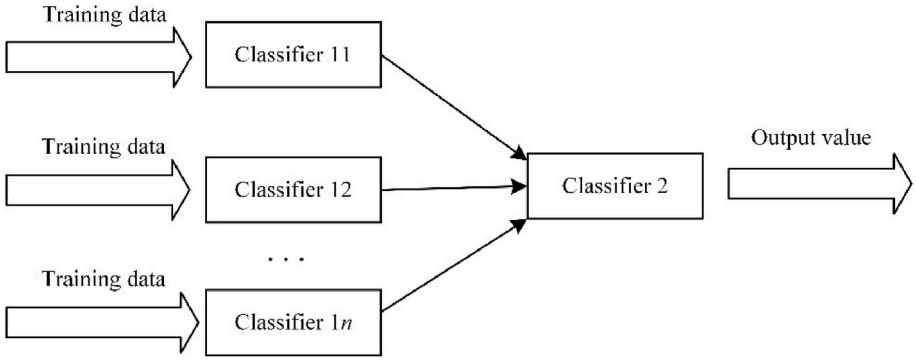


Fig. 3. Example of a two-level stacking model.

Table 5
Forecasting results obtained by the proposed data-driven methods.

Prediction model	MAE	ACC	Precision	Recall	F1-score	AUC
LR	0.16	83.71%	0.519 4	0.067 7	0.119 7	0.527 7
RF	0.07	93.09%	0.882 9	0.666 2	0.759 4	0.944 8
LGB	0.06	94.41%	0.868 4	0.776 3	0.819 7	0.969 2
XGB	0.06	94.19%	0.887 3	0.739 4	0.806 6	0.969 6
Stacking	0.05	94.50%	0.879 4	0.769 7	0.820 9	0.970 0

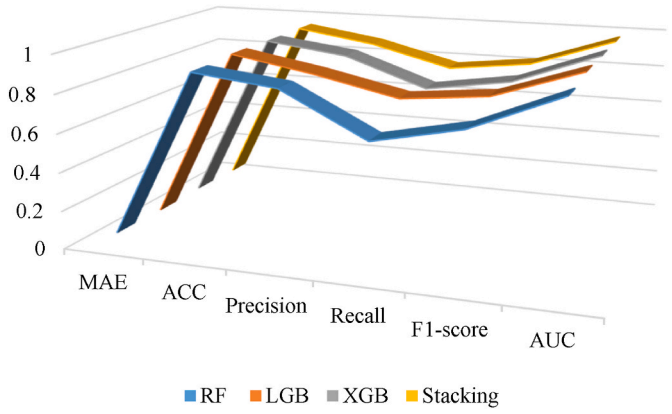


Fig. 4. Comparison results of the prediction models in addition to LR.

5.2.1. *Results and discussions of SHAP*

Based on the open-source package of Python, i.e., *shap*, the importance of the top 20 variables can be calculated by the mean SHAP value, as shown in Fig. 5. In this figure, normal service preference, boutique order number, rating, and last action type are the most vital variables. Fig. 5 also illustrates that other variables, such as last order time and boutique service preference, are moderately important. Some variables not shown in Fig. 5, such as gender and province, show weak forecasting contributions to customer purchases of boutique tourism services.

After obtaining the importance of variables, Fig. 6 shows the summary of SHAP values for the top 20 variables to understand the effect of each variable. In Fig. 6, each point represents a sample, the red point represents a larger value, and the blue point represents a smaller value. The ordinate represents the top 20 variables, and the abscissa represents SHAP value. It can be concluded that there are three types of variables in Fig. 6. The first type's characteristic is that the larger the variable value is, the greater the SHAP value, i.e., the greater the contribution to the forecasting. Correspondingly, the smaller the variable value is, the smaller the contribution to the forecasting. Representative variables include boutique service number and last action type. The second type of variable shows that the larger the variable value is, the smaller the SHAP value, i.e., the smaller the forecasting contribution. Correspondingly, the smaller the variable value is, the greater the forecasting contribution. This type of variable includes normal service preference. Third, the variable value change has little effect on SHAP value, such as the minimum of the time interval between action 5 and the last action.

Then, considering that how these variables affect the forecasting results remains to be analyzed, this paper adopts PDP to analyze the result interpretation and discuss some managerial insights in the following subsection.

5.2.2. *Results and discussions of PDP*

In this part, the top four important variables obtained by SHAP are selected for PDP analysis to identify the effects of each variable. Based on the *pdpbox* package in Python, PDP of these four variables can be obtained, as presented in Fig. 7 to Fig. 10. In PDP, the x-axis represents the variable value, while the y-axis represents the change in forecasting results compared to the baseline. The blue shaded parts in these figures represent the confidence interval, and the curves represent the influence of each variable on the final forecasting. Next, the influences of the selected four variables are discussed.

- Normal service preference

This feature is calculated by the variable analysis of online textual reviews, representing customers' preference for normal tourism service. According to SHAP results in Fig. 5, normal service preference is the primary variable contributing to the final forecasting. That is, compared to other variables, valuable information extracted in online reviews has a greater impact on boutique tourism purchases. This result confirms the fact we discussed above that for experiential products such as online tourism, customers rely more on review content to state their experience. In addition, according to the summary of SHAP value in Fig. 6, this variable negatively affects the purchase of boutique tourism services. In PDP of this variable shown in Fig. 7, the curve below the baseline also confirms this finding. In particular, Fig. 7 shows that as the value of normal service preference increases, this variable's negative effect on forecasting results remains constant. In other words, once customers show a preference for normal service, they are less likely to purchase boutique tourism service, and this likelihood does not change with the increase of normal service preference. This finding can be explained by bounded rationality theory, which indicates that customers' cognition limits their rationality. When customers prefer normal tourism service, they are therefore more likely to avoid browsing information about boutique tourism service that requires many extra cognitive efforts.

Notably, as displayed in Fig. 5, compared with the normal service preference, the boutique service preference has no significant impact on the final result (14th rank). This is because in the boutique tourism purchases, the normal service preference can be regarded as the negative feedback. While according to the negativity bias, the unfavorable feedback has a greater impact than favorable ones. This finding can serve as

Fig. 5. The importance of the top 20 variables.

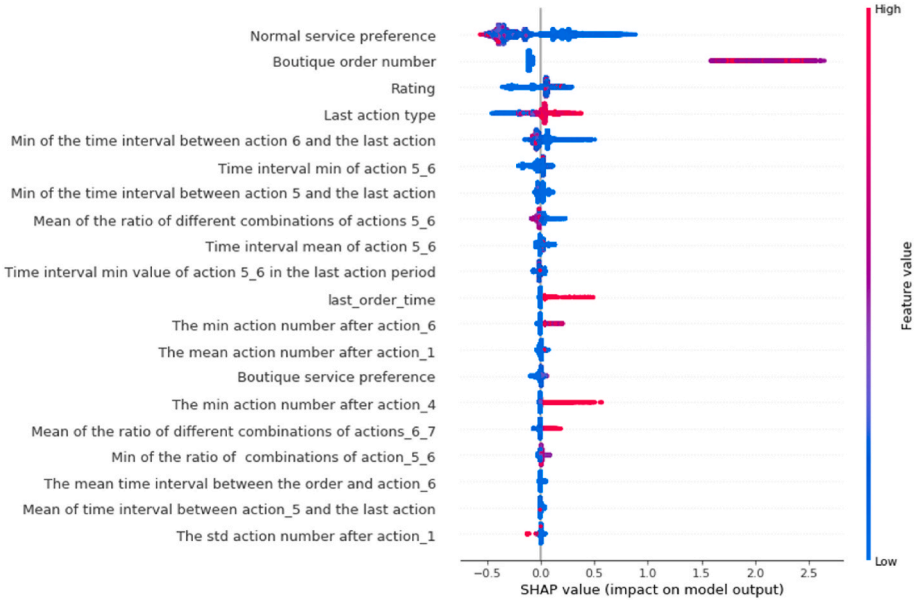


Fig. 6. Summary of SHAP value for the top 20 variables.

a reminder that managers cannot only rely on direct preferences to predict future customer purchases. The indirect negative preferences of customers can provide more valuable information. In this paper, for customers who prefer normal tourism service, we think they are less likely to purchase boutique tourism service in the future. This finding may help managers identify some irrelevant customers.

More importantly, considering the importance of customer preference in online reviews, managers should pay more attention to extracting customer preference from online tourism reviews. In other words, it is necessary to consider various ways to capture customer preferences when designing tourism online review pages. Considering that customers are sometimes reluctant to give comments, some easy ways, such as designing multiple-choice questions similar to questionnaires, can be adopted to attract customers to give reviews. These multiple-choice questions should include customer focus, current tourism issues, and distinct tourism service features. For example, managers can set up questions related to tourism keywords in Table S1 of Supplementary_A. Since these questions are something customers caring about, they are more likely to give comments. Current tourism issue-related questions can also attract customers giving reviews, and as a result attract potential customers to browse and heed. Then, questions containing distinct tourism service features can provide more specific tourism descriptions. Therefore, potential customer confusion caused by similar tourism information can be avoided. Additionally, some functions to promote customers' interactions can be attached to tourism online review pages, allowing customers to supplement content not mentioned in the multiple-choice questions. Trust transference theory confirms that social interactions and information exchange contribute to increased online trust and eventually lead to higher demand. The customer preferences obtained in this way are more in line with the actual preference and are of high quality and easy to analyze. When customers perceive the positive quality of online tourism platforms, the

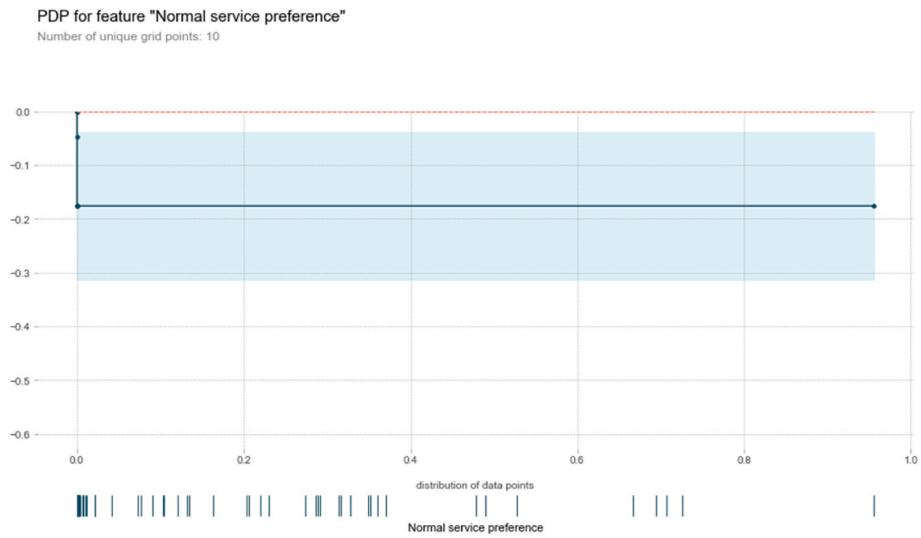


Fig. 7. PDP of the normal service preference.

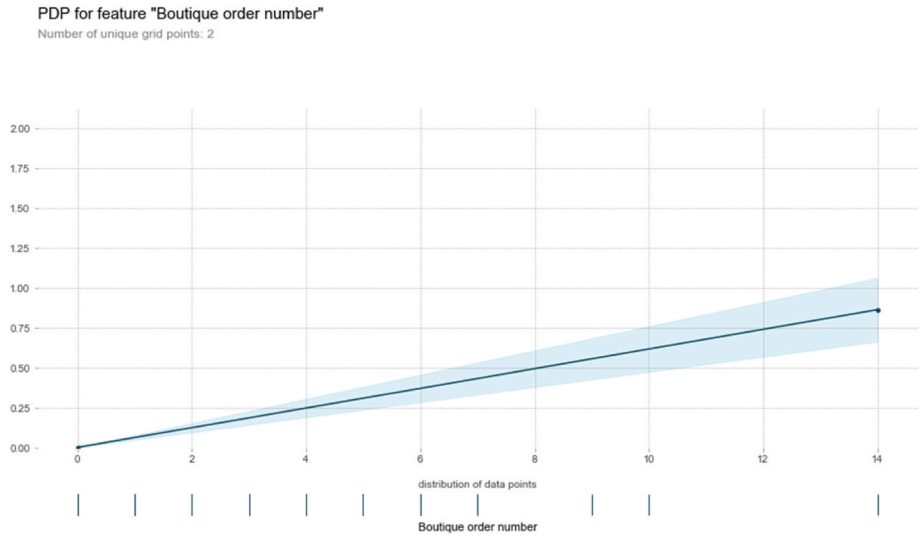


Fig. 8. PDP of the boutique order number.

possibility that they use them to purchase also increases.

- Boutique order number

Fig. 5 suggests that the boutique order number shows great importance in boutique tourism service purchases. This result is consistent with actual online purchases; that is, historical selection affects future purchases. In addition, PDP in Fig. 8 illustrates that the boutique order number has a monotonic positive association with boutique tourism service purchases. That's to say, the purchases of boutique tourism services tend to increase as the number of historical boutique orders increases, indicating the positive reputation and goodwill accumulated over a long period. Therefore, this variable can be regarded as the basis for a business to establish a lasting relationship with its customers and is closely related to the customer trust.

In addition, Fig. 5 also shows that the last order time is also the key variable among the order-related variables. PDP of the last order time is shown in Fig. S1, demonstrating this variable's monotonic positive effect on the final result. In other words, customers who have recently purchased tourism services are more likely to purchase boutique tourism services. This finding contradicts some previous studies, which confirmed that tourism services were purchased relatively infrequently. Therefore, customers are less likely to purchase again when they have recent travel experience. However, as mentioned above, some studies suggested that the relevance of historical information to future purchases decreased over time. In other words, it is reasonable that recent purchases have positive effects. These two results seem to contradict each other, and we believe that there should be a balanced time point. When the last order time is closer to the time point, customers are more likely to purchase boutique tourism services. Moreover, as shown in Fig. 5, in terms of order-related variables, the effect of boutique order number is more important than that of last order time. In this case, the boutique order number may influence the effect of the last order time when forecasting, thereby generating PDP result in Fig. S1. However, since the last order time shows relatively little impact on the final result, we do not carry out a detailed discussion. Further analysis can be conducted to discover the particular effect of the last order time.

Combined with the above findings, some management implications can be concluded to guide the business activities of online tourism. On the Internet, repeat customers, those who order more than one, can be more profitable than new customers (Kim & Gupta, 2009). Therefore, combining with PDP results in Fig. 8, managers should put forward some

methods to increase customer trust and encourage consumers to make repeat purchases. For example, different combinations of tourism boutique services can be recommended to different customers according to their purchase possibilities in Fig. 8. In addition, managers should keep in touch with customers who show a high possibility of purchasing repeatedly and offer discount activities to them in time to attract purchasing.

- Rating

Rating is the representative of crowd intelligence and shows great importance to the purchases of boutique tourism services, according to Fig. 5. In line with this finding, many studies have emphasized the importance of ratings on customer purchases (Fang et al., 2016). PDP of this variable is shown in Fig. 9. From this figure, PDP values of ratings higher than the neutral level (i.e., rating between 3 and 5) are higher. That is, customers with ratings between 3 and 5 are more likely to purchase boutique tourism services. While PDP values of ratings between 1 and 3 are relatively low, indicating that these customers are less likely to purchase boutique tourism services. This finding is consistent with the existing researches and can easily be explained: since the rating represents the customer's satisfaction with the tourism service, a rating below the neutral level indicates that the provided tourism service does not satisfy the customer's desired expectations. Therefore, customers are less likely to purchase boutique tourism services in the future. Accordingly, customers with higher satisfaction levels are more likely to purchase boutique tourism again.

While in terms of the dynamic effects of this variable, it can be found in Fig. 9 that the impact continues to increase linearly when the rating increases from the lowest level to the neutral level. However, when the rating is higher than 3 points, PDP values remain unchanged with the increase of rating. In summary, the customer purchases will increase as the variable increases within the certain scope. When the it equals a certain value, increasing the variable has no effect on the final result. This finding can be explained by the marginal benefits decrease theory of consumer utility, which indicates that rational consumers seek maximum total utility, but they do not increase consumption when the marginal utility is reduced to zero. That is to say, although improving the variable value can increase the revenue (for variables such as the rating), the continuous improvement does not bring new income when reaching a certain level (where marginal utility is zero). This result is consistent with a previous study (Van Nguyen et al., 2020), which confirmed that positive polarity reviews positively affected sales, but such an effect would level off when there was a strong positive polarity in the reviews. For this type of variable, managers should avoid the continuous resource investment when it reaches the certain level. Instead, resources should be redirected to areas where they can be used more productively to avoid wasting. This conclusion can also be extended to the analysis of other variables with the similar characteristics.

Therefore, it should be noted that tourism managers should pay more attention to customers with ratings lower than 3 points, as the resource investment in these customers may bring considerable incomes. This result is contradictory with some traditional views that managers should focus on customers with ratings from 3 to 5. Although these customers are more likely to purchase boutique tourism service, the stable effect of ratings from 3 to 5 shows few contributions when increasing resource investment in these customers. In this regard, considering the roles of unfavorable ratings, i.e., informants (customer feedback) and recommendations (service popularity) (Park & Lee, 2008), managers can take measures on ratings to attract customers. For example, managers should first understand why these customers give unfavorable ratings by textual online review analysis and following up customers. Then, analyze the aspects that need to be improved to reach the neutral level of ratings.

- Last action type

The last action type represents what customers did for the last time they used the app, so it is easy to associate this variable with customer purchase. Fig. 5 suggests that this variable has a vital contribution to boutique tourism service purchases. And PDP in Fig. 10 shows the specific effects of this variable on the final result. First, it can be seen from Fig. 10 that actions before filling out (action 1 to action 5) do not affect the purchase of boutique tourism services. This is because online customers often spend more time and effort comparing and judging tourism services they want to purchase. According to rationality boundary theory, they therefore tend to have uncertainty when browsing online before filling out, triggering great debate and requiring greater cognitive effort to evaluate. Therefore, actions before filling out show an unclear effect on the boutique tourism service purchases.

In addition, it can be seen in Fig. 10 that when the last action type is filling out (action 5), this variable begins to have a positive effect on the purchase of boutique tourism services. However, when the last action is paying (action 6 or later), the effect of this variable remains constant. It can be concluded that the action 5 and action 6 are the crucial actions. This finding can also be confirmed by Fig. 5, in which many variables related to these two actions have major influences. This result can be explained by the fact that after customers accomplish the filling out, i.e.,

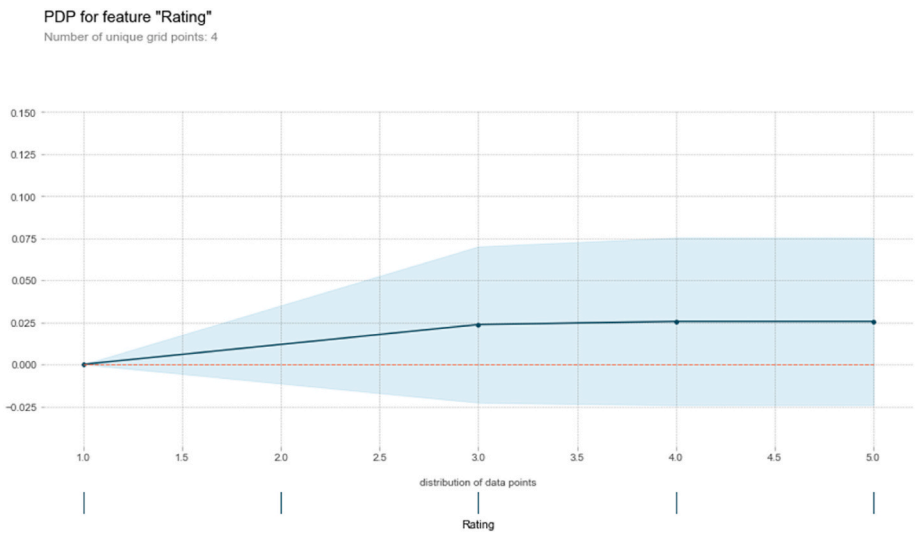


Fig. 9. PDP of the rating.

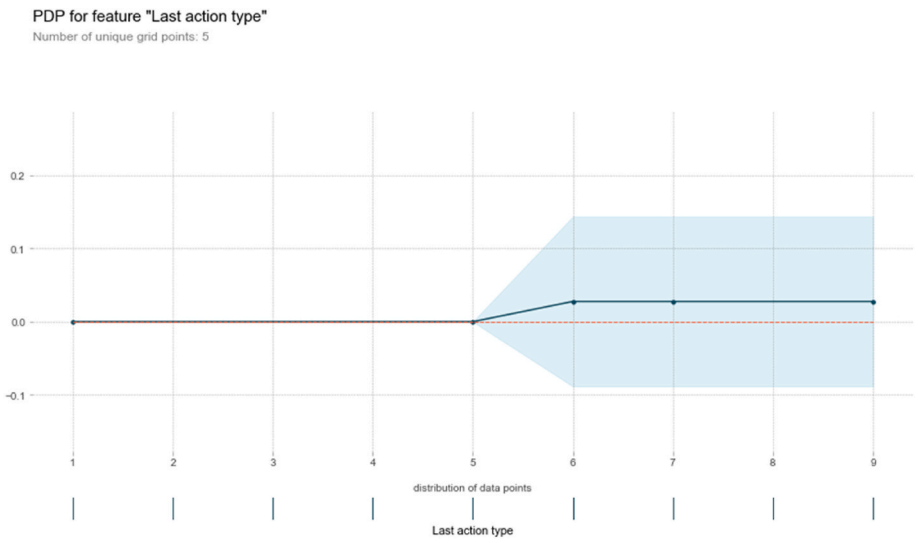


Fig. 10. PDP of the last action type.

action 5, they have found something they want among the numerous pieces of information. Hence, it is evident that these customers are more likely to purchase boutique tourism services. In addition, key actions obtained by the data-driven method help discover valuable information related to actions. Thus, the amount of data to be analyzed can be reduced, which is an excellent advantage for dealing with massive customer operational behavior data.

From the above findings, managers should pay attention to the process from action 5 to action 6, i.e., filling out to paying. First, managers can push information to customers staying in action 5 to remind them to pay. In addition, managers can optimize the patterns of filling out and payment in online tourism platforms to help customer accomplish paying in time. First, multiple payment methods, such as NFC, QR codes, and mobile wallets, can be offered for tourism payments. Second, the availability, ease of use, security, and compatibility of the tourism payment system should be highly valued. The safer the online tourism platform, the more likely customers utilize the platform (Lee et al., 2012). Third, if there is an international tourism business, the compatibility and ease with overseas customer payments should also be considered.

The above findings are obtained by the proposed data-driven method, which contributes in discovering the effect of variables that are difficult to obtain by the hypothesis or empirical analysis. In addition to the above four variables, the effects of other variables can also be analyzed. A summary of the findings in this work is provided in Supplementary Materials_B.

6. Conclusions

Modern tourism has become individual, accessible, and user-friendly thanks to the information presented on the Internet. This paper develops a data-driven method to obtain customer purchase forecasting results for online tourism and help managers provide better tourism services online. First, this paper proposes different variable analysis methods for multiplex behavior data. Second, this study adopts several different machine learning algorithms to verify the validity of the extracted variables. The experimental results show that stacking performs excellently. Finally, some market insights are discussed by two interpretation models. Compared with the existing research, this paper has the following advantages. First, this paper adopts PLTSs to transform a large amount of operational behavior data into a simpler form. Second, this paper introduces prospect theory to analyze customer ratings so that customer risk perception can be considered. The third advantage is that the proposed data-driven method contributes to discovering valuable findings to guide business activities of online tourism.

However, there are still some shortcomings in this paper. First, some critical information, such as airline tickets, service prices, and holiday packages, is not considered due to the limited data available. Second, in addition to tourism platforms, information from other websites, such as YouTube and Tik Tok, also impacts customer purchases. In future work, we will consider the multi-source information for a comprehensive analysis. Third, the findings of this paper are obtained based on one tourism platform, so it may be difficult to apply the results to other platforms. Future research can consider the commonalities and differences between different tourism platforms to share some effective implications.

Credit author statement

Shui-xia Chen: Conceptualization, Methodology, Data curation, Writing – original draft, Visualization. **Xiao-kang Wang**: Investigation, Methodology, Data curation, Validation, Visualization. **Hong-yu Zhang**: Investigation, Validation, Data curation. **Jian-qiang Wang**: Supervision, Writing – review & editing, Funding acquisition. **Juan-juan Peng**: Investigation, Validation, Data curation.

Impact statement

Using a real-world database on a travel service App, this paper proposes a data-driven method with multiplex behavior data to achieve two objectives: (1) to provide a highly accurate purchase forecasting model of online tourism; and (2) to analyze the influencing of behavior variables as predictors of customer purchase. On the one hand, this study can provide a highly accurate purchase forecasting model for online tourism. On the other hand, the findings of this study can analyze some practical market insights for related managers, from which the deployment of resource allocations and business strategies can be effectively conducted.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 71871228) and the Key Project of Hunan Social Science Achievement Evaluation Committee (XSP18ZDI021).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.

org/10.1016/j.tourman.2021.104357.

References

Amaro, S., & Duarte, P. (2015). An integrative model of consumers' intentions to purchase travel online. *Tourism Management, 46*, 64–79.

Bag, S., Tiwari, M. K., & Chan, F. T. S. (2019). Predicting the consumer's purchase intention of durable goods: An attribute-level analysis. *Journal of Business Research, 94*, 408–419.

Baumann, A., Haupt, J., Gebert, F., & Lessmann, S. (2018). Changing perspectives: Using graph metrics to predict purchase probabilities. *Expert Systems with Applications, 94*, 137–148.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test, 25*, 197–227.

Cai, L., & White, R. E. (2011). Mathematical modeling of a lithium ion battery with thermal effects in COMSOL Inc. Multiphysics (MP) software. *Journal of Power Sources, 196*, 5985–5989.

Chen, J., Kou, G., & Peng, Y. (2018). The dynamic effects of online product reviews on purchase decisions. *Technological and Economic Development of Economy, 24*, 2045–2064.

Chen, T. L., & Wang, C. C. (2016). Multi-objective simulation optimization for medical capacity allocation in emergency department. *Journal of Simulation, 10*, 50–68.

Chen, S.-x., Wang, J.-q., & Wang, T.-l. (2019). Cloud-based ERP system selection based on extended probabilistic linguistic MULTIMOORA method and Choquet integral operator. *Computational and Applied Mathematics, 38*, 88.

Chen, S.-x., Wang, X.-k., Zhang, H.-y., & Wang, J.-q. (2021). Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine. *Expert Systems with Applications, 173*, 114756.

Chen, L., Wu, Z., Cao, J., Zhu, G., & Ge, Y. (2020). Travel recommendation via fusing multi-auxiliary information into matrix factorization. *ACM Transactions on Intelligent Systems and Technology (TIST), 11*, 1–24.

Chen, L., Zhang, L., Cao, S., Wu, Z., & Cao, J. (2020). Personalized itinerary recommendation: Deep and collaborative learning with textual information. *Expert Systems with Applications, 144*, 113070.

Chu, S., Jiang, H., Xue, Z., & Deng, X. (2020). Adaptive convex clustering of generalized linear models with application in purchase likelihood prediction. *Technometrics*, 1–13.

Conklin, J. D. (2002). Applied logistic regression. *Technometrics, 44*, 81–82.

Dong, Y., & Jiang, W. (2019). Brand purchase prediction based on time-evolving user behaviors in e-commerce. *Concurrency and Computation: Practice and Experience, 31*.

Dou, X. (2020). *Online purchase behavior prediction and analysis using ensemble learning*. Chengdu, China: IEEE 5th International Conference on Cloud Computing and Big Data Analytics. https://doi.org/10.1109/ICCCBDA49378.2020.9095554

Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016). Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management, 52*, 498–506.

Gavilan, D., Avello, M., & Martinez-Navarro, G. (2018). The influence of online ratings and reviews on hotel booking consideration. *Tourism Management, 66*, 53–61.

Huang, C., Wu, X., Zhang, X., Zhang, C., Zhao, J., Yin, D., & Chawla, N. (2019). Online purchase prediction via multi-scale modeling of behavior dynamics. In *25th ACM SIGKDD international conference, anchorage, AK*. https://doi.org/10.1145/3292500.3330790

Hu, M., Qiu, R. T. R., Wu, D. C., & Song, H. (2021). Hierarchical pattern recognition for tourism demand forecasting. *Tourism Management, 84*, 104263.

Hu, X., Yang, Y., Zhu, S., & Chen, L. (2020). *Research on a hybrid prediction model for purchase behavior based on logistic regression and support vector machine*. Chengdu, China: International Conference on Artificial Intelligence and Big Data. https://doi.org/10.1109/ICAIBD49809.2020.9137484

İçer, Ç., Parmaksız, D., & Ç, A. (2021). Predicting likelihood to purchase of users for E-commerce. In C. Kahraman, S. Cevik Onar, B. Oztaysi, I. Sari, S. Cebi, & A. T (Eds.), *Intelligent and fuzzy techniques: Smart and innovative solutions*. Cham: Springer.

Kagan, S., & Bekkerman, R. (2018). Predicting purchase behavior of website audiences. *International Journal of Electronic Commerce, 22*, 510–539.

Kahneman, A. T. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*, 297–323.

Kim, H.-W., & Gupta, S. (2009). A comparison of purchase decision calculus between potential and repeat customers of an online store. *Decision Support Systems, 47*, 477–487.

Kytö, E., Virtanen, M., & Mustonen, S. (2019). From intention to action: Predicting purchase behavior with consumers' product expectations and perceptions, and their individual properties. *Food Quality and Preference, 75*, 1–9.

Lee, J.-E. R., Rao, S., Nass, C., Forssell, K., & John, J. M. (2012). When do online shoppers appreciate security enhancement efforts? Effects of financial risk and security level on evaluations of customer authentication. *International Journal of Human-Computer Studies, 70*, 364–376.

Li, Q., Zeng, D. D., Xu, D. J., Liu, R., & Yao, R. (2020). Understanding and predicting users' rating behavior: A cognitive perspective. *INFORMS Journal on Computing, 32*, 855–1186.

Lundberg, S., & Lee, S. I. (2017). *A unified approach to interpreting model predictions*. Long Beach, CA: Neural Information Processing Systems 2017.

Martinez, A., Schmuck, C., Pereverzyev, S., Pirker, C., & Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research, 281*, 588–596.

Moeller, S. (2010). Characteristics of services – a new approach uncovers their value. *Journal of Services Marketing, 24*, 359–368.

Nishimura, N., Sukegawa, N., Takano, Y., & Iwanaga, J. (2018). A latent-class model for estimating product-choice probabilities from clickstream data. *Information Sciences, 429*, 406–420.

Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. *Industrial Management & Data Systems, 116*, 1678–1699.

Park, C., Kim, D., Yang, M.-C., Lee, J.-T., & Yu, H. (2020). Click-aware purchase prediction with push at the top. *Information Sciences, 521*, 350–364.

Park, C., Kim, D., & Yu, H. (2019). An encoder–decoder switch network for purchase prediction. *Knowledge-Based Systems, 185*, 104932.

Park, D.-H., & Lee, J. (2008). eWOM overload and its effect on consumer behavioral intention depending on consumer involvement. *Electronic Commerce Research and Applications, 7*, 386–398.

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention, 136*, 105405.

Shao, J.-b., Li, Z.-z., & Hu, M.-y. (2014). The impact of online reviews on consumers' purchase decisions in online shopping. *International conference on management science and engineering*. Helsinki: Finland. https://doi.org/10.1109/ICMSE.2014.6930242

Shapoval, K., & Setzer, T. (2018). Next-purchase prediction using projections of discounted purchasing sequences. *Business & Information Systems Engineering, 60*, 151–166.

Spärck Jones, K. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 60*, 493–502.

Toledo, R. Y., Mota, Y. C., & Martínez, L. (2015). Correcting noisy ratings in collaborative recommender systems. *Knowledge-Based Systems, 76*, 96–108.

Tversky, K. A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–292.

Van Nguyen, T., Zhou, L., Chong, A. Y. L., Li, B., & Pu, X. (2020). Predicting customer demand for remanufactured products: A data-mining approach. *European Journal of Operational Research, 281*, 543–558.

Wagner, H. N. R., Köke, H., Dähne, S., Niemann, S., Hühne, C., & Khakimova, R. (2019). Decision tree-based machine learning to optimize the laminate stacking of composite cylinders for maximum buckling load and minimum imperfection sensitivity. *Composite Structures, 220*, 45–63.

Wang, D., Zhang, Y., & Zhao, Y. (2017). LightGBM: An effective miRNA classification method in breast cancer patients. In *Proceedings of the 2017 international conference on computational biology and bioinformatics, newark, NJ, USA*. https://doi.org/10.1145/3155077.3155079

Wen, Y.-T., Yeh, P.-W., Tsai, T.-H., Peng, W.-C., & Shuai, H.-H. (2018). Customer purchase behavior prediction from payment datasets. *Proceedings of the eleventh ACM international conference on web search and data mining.* , CA, USA: Marina Del Rey. https://doi.org/10.1145/3159652.3159707

Xiao, S. S., Wei, C. P., & Dong, M. (2016). Crowd intelligence: Analyzing online product reviews for preference measurement. *Information & Management, 53*, 169–182.

Xu, X., & Gursoy, D. (2015). A conceptual framework of sustainable hospitality supply chain management. *Journal of Hospitality Marketing & Management, 24*, 229–259.

Yu, S. M., Wang, J., & Wang, J. Q. (2018). An extended TODIM approach with intuitionistic linguistic numbers. *International Transactions in Operational Research, 25*, 781–805.

Zhou, A., Ren, K., Li, X., & Zhang, W. (2019). *Mmse: A multi-model stacking ensemble learning algorithm for purchase prediction*. Chongqing, China: 8th Joint International Information Technology and Artificial Intelligence Conference.

Zhu, G., Wu, Z., Cao, J., & Gu, J. (2018). *Online tourism purchase analysis and prediction*. Lanzhou, China: Sixth International Conference on Advanced Cloud and Big Data. https://doi.org/10.1109/CBD.2018.00039

Zhu, G., Wu, Z., Wang, Y., Cao, S., & Cao, J. (2019). Online purchase decisions for tourism e-commerce. *Electronic Commerce Research and Applications, 38*, 100887.

Shui-Xia Chen received the M.S. degree in Management Science and Engineering at the Business School of Central South University, Changsha, China, in 2020. She is currently working toward the Ph.D. degree in Management Science and Engineering from Sichuan University, Chengdu, China. Her current research interests include data analysis and forecasting, machine learning, decision-making theory and application.

Xiao-kang Wang received the B.S. degree in School of Information Science and Engineering from Central South University, China, in 2015. He is currently working toward the Ph.D. degree in Management Science and Engineering from Central South University, Changsha, China. His current research interests include decision-making theory and application, evaluation theory and method, and information management. E-mail:xkwang@csu.edu.cn.

Hong-yu Zhang is an associate professor of the Business School, Central South University (CSU). She was born in 1979 and received her PhD in Management Science and Engineering from CSU in 2009. She also holds an MS degree in Computer Software and Theory in School of Information Science and Engineering, CSU. Her research currently focuses on decision support system driven by big data and social commerce. E-mail: hyzhang@csu.edu.cn

Jian-qiang Wang is a professor in the Department of Management Science and Information Management at the Business School of Central South University. He was born in 1963. He holds a PhD in management science & engineering and he is also a PhD supervisor in this major. Over the past couple decades, his research interests are in the area of decision-making theory. His current research interests include (i) decision-making theory and application; (ii) risk management and control, and (iii) information management. E-mail: jqwang@csu.edu.cn.

Juan-juan Peng is an associated professor in School of Information Management and Engineering, Zhejiang University of Finance & Economics. She is also a postdoctoral researcher in Business School, Central South University. She received her M. Sc. degree in Computational Mathematics from Wuhan University of Technology and Ph.D. degree in Business School of Central South University. Her current research interests include (i) decision-making theory and application; (ii) risk management and control, and (iii) information management. E-mail: pengjj81@csu.edu.cn.