Mining the Value Characteristics of One-Way Car-Sharing Users Based on Empirical Data

Xue Mi\textsuperscript{1a}, Ying-Qi Wang\textsuperscript{1b}, Dong-Fan Xie\textsuperscript{1c},

\textsuperscript{1}Institute of System Science, School of Traffic and Transportation, Beijing Jiaotong University, Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Beijing Jiaotong University, Beijing, China

\textsuperscript{a}18120861@bjtu.edu.cn, \textsuperscript{b}18120754@bjtu.edu.cn, \textsuperscript{c}dfxie@bjtu.edu.cn

Abstract—Car-sharing has been emergent in many cities all over the world due to the environmental sustainability and the diverse service for users. However, the huge investment and the long payback period hinder the development of car-sharing. To this end, this study attempts to mine the value of car-sharing users in terms of empirical data, aiming to provide suggestions for companies to enhance user management. Based on the order data of a car-sharing company, several indicators are defined. On the basis, the users are divided into three categories, namely high-value users, low-value users and potential users. Furthermore, with the proposed indicators and the personal attributes, a prediction model is developed based on random forest method to predict the category of users in advance. The results show that with the observation period of 7 weeks, the long-term value type of a user can be predicted with accuracy of more than 80%.

1. INTRODUCTION

With the rapid development of the economy, the global automobile production and car ownership are increasing year by year, resulting in serious traffic congestion. In recent years, car-sharing is regarded as a new transportation mode, has attracted much attention. In general, car-sharing could contribute to urban traffic. For instance, car-sharing can effectively reduce the car ownership and thus can alleviate traffic congestion. In addition, electric cars become more and more prevalent for car-sharing systems. Consequently, the total greenhouse gas emissions can be effectively reduced [1,2]. All in all, it can be seen that car-sharing is a prospective transportation mode for urban traffic [3-5].

Car-sharing is a car rental mode that the use of a car for a particular trip, mainly for people to rent a car for a short period of time. Since users are the key component of car-sharing system which may impact on the operation of a car-sharing system, many studies have conducted on car-sharing users. For instance, some studies focus on users’ willingness to use car-sharing service [6-11]. Some studies addressed on travel mode selection [12-16]. In addition, more studies concentrate on the behavior characteristics of users [17-22].

Based on the order data of car-sharing companies, the travel behavior of users can be extracted and analyzed, which is expected to be helpful for companies to manage users with different needs. Qian et al. [17] used the clustering method to analyze the usage patterns of car-sharing users in Hangzhou, China. Ying et al. [18] attempted to mine the travel behavior of car-sharing users based on GPS data and order data in Hangzhou. Based on the order and user data, Chen et al. [20] analyzed the travel...
demand and the spatiotemporal distribution of demand, and analyzed the characteristics of high-frequency users and commuting users.

Particularly, the user value is one of the most essential factors for car-sharing companies. So far, several studies have been conducted on this topic. Sai et al. [21,22] employed the RFM model to access the user value. Furthermore, they used clustering method to investigate the value of car-sharing users, dividing users into high, medium, and low value categories.

There are still several problems to be solved in the existing literature. First, most studies focus on the users in the integrated level. Therefore, the diversity of car-sharing users is not completely considered. Second, several works carried out the clustering based on existing variables. The decision variables are not well defined and selected. In addition, few works have been done to deeply understand the value of car-sharing users, although this is an important topic for car-sharing company.

To address the aforementioned issues, this study will analyze the users from the perspective of user value. The contributions and motivations of this paper are as follows. (1) Based on the order data of a car-sharing company, various variables (indicators) are defined to represent the characteristics of the users. (2) A clustering algorithm is applied to divide the users into different categories, and the value and behavior characteristics of each category are analyzed. (3) The prediction model is developed to predict the category of a new user with the observed data in advance.

The rest of this paper is organized as follows: In Section II, the empirical data used in this study and variable definition are introduced. In Section III, the framework of the method is presented, including the clustering and prediction model for car-sharing users. In Section IV, case studies are carried out to validate the proposed model. Finally, conclusions are summarized in Section V.

2. DATA DESCRIPTION AND METHODOLOGY

2.1. Data Description

2.1.1. Data Collection
The data used in this study is provided by a car-sharing company in Lanzhou, China. The data includes order data and user data. The time range of order data is from September 7, 2018 to March 25, 2019, including order number, start and end time of the orders, and so on. User data contains the registration information of each user at the car-sharing company before April 3, 2019. Table I is an example of the order data used in this study, and Table II is an example of registered user data.

2.1.2. Data Preprocessing
This study focuses on two types of data, named as short-term data and long-term data, respectively. With detail analysis of travel behavior of car-sharing users based on short-term data, this study attempts to capture the long-term characteristics of a user. Since 14 weeks can reflect the user's behavior characteristics, the time ranges for long-term data is set to be 14 weeks. Due to the invalid data in the original data and the time span range limitation for each user, the following data preprocessing process is required.

2.1.2.1. Delete invalid data
There exist invalid samples in the original data due to the errors of data collection devices, signal transmission process, and so on. In this study, invalid data in the original data was deleted.

2.1.2.2. Filter users
According to the objective of this study, valid users are selected subject to the following two constraints: First, during the period from September 7, 2018 to March 25, 2019, the order number of the user should be not smaller than 1. Second, for each user, the time of the first order should be earlier than December 18, 2018, to ensure the time range of each user is longer than 98 days.
2.1.2.3. Extract order data
According to the filtered user ID, the order information of each user within 98 days after the first use of the shared car is extracted from the order data. Finally, the valid data contains 85993 order samples for 5787 users.

2.2. Variable Definition
To well reveal the characteristics of car-sharing users, several new variables will be defined in this subsection, which can be calculated based on the filtered data samples.

2.2.1. Clustering Variables
In this study, the RFM model is applied to evaluate the value of car-sharing users. However, by considering the special characteristics of car-sharing users, two additional variables are introduced to extend the RFM model. First, the car-sharing companies generally provide several types of shared cars, and users can choose appropriate cars according to their economic ability. To this end, the car type is introduced. Second, the user's familiarity with the car-sharing system may affect their choice behavior, so this paper defines user registration duration.

2.2.1.1. Total consumption amount
The total consumption amount refers to the total amount of travel consumption of a single user within $T$ days after the first use of the vehicle. The formula is as follows:

$$A_t = \sum_{d=1}^{T} A_{i,d}$$

where $A_t$ is the total amount spent by user $i$ within $T$ days; and $A_{i,d}$ represents the amount consumed by user $i$ on day $d$.

2.2.1.2. Order amount of a user
The order amount refers to the number of trips taken by a user within $T$ days after the first use of the shared car. It is calculated as follows:

$$C_t = \sum_{d=1}^{T} C_{i,d}$$

where $C_{i,d}$ represents the number of trips made by user $i$ on day $d$.

| Order ID | Rental Time | Return Time | Rental Station | Return Station | Car Type | Mileage (km) | Duration (min) | Amount (Yuan) | User ID |
|----------|-------------|-------------|----------------|----------------|----------|--------------|----------------|---------------|---------|
| 180***0000 | 2018/9/7 0:00:00 | 2018/9/7 0:35:12 | AST_3000 48hm20kc | AST_2000 59f5458 | EC200 | 15 | 36 | 17.75 | bzk1 |

| User ID | Company ID | Registration Time | Gender | Age |
|---------|------------|-------------------|--------|-----|
| 100***k6bk | com_10003rrszqmm6 | 2017/6/24 16:03:30 | male | 34 |

2.2.1.3. Car type
According to the data used in this study, there are 4 types of cars, as showed in Table III. One can see that the charging costs of them are quite different. Here, the preferred car type for a user is assumed to be the most frequent type among the user’s orders, which is represented as follows:

$$T_i = \text{mode}(T_{i,d}), \quad c = 1,2,3,4$$

where $T_{i,d}$ represents the type of car used by user $i$ within $T$ days. $c \in \{1,2,3,4\}$ is the car type, as in Table III. $\text{mode}(\cdot)$ is a function of the mode.
2.2.1.4. Registration duration
In this study, registration duration refers to the time difference between the user’s registration and the end of $T$ days after the first use of shared cars, which is as follows:

$$T_i^{\text{regis}} = D_i^{\text{end}} - D_i^{\text{regis}}$$  \hspace{1cm} (4)

where $D_i^{\text{end}}$ represents the end of $T$ days after the first used of shared car by user $i$. $D_i^{\text{regis}}$ is the registration time of user $i$.

2.2.1.5. Last gap time
Last gap time represents the time difference between the time of last order by user $i$ and the end of $T$ days, which is,

$$T_i^{\text{gap}} = D_i^{\text{end}} - D_i^{\text{last}}$$  \hspace{1cm} (5)

where $D_i^{\text{last}}$ represents the last time user $i$ used the car in $T$ days.

The above five indicators reflect users’ activity, loyalty and contribution to the company. Therefore, these indicators will be used as independent variables for the clustering of car-sharing users.

2.2.2. Behavior Characteristics of Car-sharing Users
The variable of car use behavior can reflect the behavior characteristics of different users and analyze the difference of their car use behavior. This subsection defines the users’ behavior variables from the following five aspects.

2.2.2.1. Station heterogeneity ratio
Station heterogeneity ratio is defined as the proportion of the number of orders with different rental and return stations of user $i$ during $T$ days, which could represent the behavior preference of users regarding to the car-sharing stations. It is defined as follows:

$$R_i^{\text{station}} = \frac{c_i^{\text{diff}}}{c_i}$$  \hspace{1cm} (6)

where $c_i^{\text{diff}}$ represents the number of orders with different rental and return stations for user $i$ during the $T$ days; and $c_i$ denotes all the orders of user $i$ during $T$ days.

2.2.2.2. Order date
Let $R_i^{\text{weekday}}$, $R_i^{\text{weekend}}$ and $R_i^{\text{holiday}}$ respectively represent the proportions of orders for weekday, weekend and holiday in the total travel orders, which are,

$$R_i^{\text{weekday}} = \frac{c_i^{\text{weekday}}}{c_i}$$  \hspace{1cm} (7)

$$R_i^{\text{weekend}} = \frac{c_i^{\text{weekend}}}{c_i}$$  \hspace{1cm} (8)

$$R_i^{\text{holiday}} = \frac{c_i^{\text{holiday}}}{c_i}$$  \hspace{1cm} (9)

where $c_i^{\text{weekday}}$, $c_i^{\text{weekend}}$, $c_i^{\text{holiday}}$ respectively represent the orders numbers for weekdays, weekends and holidays within the $T$ days.

2.2.2.3. Travel distance
Let $D_i^{\text{short}}$, $D_i^{\text{long}}$ and $D_i^{\text{average}}$ represent the shortest, longest, and average distances in orders of user $i$, respectively.

$$D_i^{\text{short}} = \min(D_i)$$  \hspace{1cm} (10)
\[ D_{i}^{long} = \max(D_i) \]  
\[ D_i^{average} = \frac{\sum D_i}{\sum \sum D_i} \]

where \( D_i \) represents the set of travel distances of user \( i \).

### 2.2.2.4. Order duration

According to the travel time distributions, order duration of users is divided into short travel (0-2 hours), medium short travel (2-12 hours), medium long travel (12-24 hours) and long travel (>24 hours), and the corresponding proportions of orders of user \( i \) are represented by \( R_i^s \), \( R_i^{ms} \), \( R_i^{ml} \) and \( R_i^l \).

\[ R_i^s = \frac{\sum_{d=1}^{T} \sum_{d=1}^{T} C_i^s}{\sum_{d=1}^{T} \sum_{d=1}^{T} C_i} \]  
\[ R_i^{ms} = \frac{\sum_{d=1}^{T} \sum_{d=1}^{T} C_i^{ms}}{\sum_{d=1}^{T} \sum_{d=1}^{T} C_i} \]  
\[ R_i^{ml} = \frac{\sum_{d=1}^{T} \sum_{d=1}^{T} C_i^{ml}}{\sum_{d=1}^{T} \sum_{d=1}^{T} C_i} \]  
\[ R_i^l = \frac{\sum_{d=1}^{T} \sum_{d=1}^{T} C_i^l}{\sum_{d=1}^{T} \sum_{d=1}^{T} C_i} \]

where \( C_i^s \), \( C_i^{ms} \), \( C_i^{ml} \) and \( C_i^l \) represent the order numbers of user \( i \) for the four durations within the \( T \) days.

### TABLE III. CHARGING STANDARDS FOR DIFFERENT TYPES OF CARS

| Car type | Charging standards                  | Classification |
|----------|------------------------------------|----------------|
| Zhidou   | 0.15 Yuan/min                       | 1              |
| E200     | 0.15 Yuan/min + 0.5 Yuan/km        | 2              |
| EC200    | 0.15 Yuan/min + 0.7 Yuan/km        | 3              |
| E5       | 0.15 Yuan/min + 1.0 Yuan/km        | 4              |

### 2.2.2.5. Order time

A day is divided into 24 segments on average. The order time refers to the time segment with the most orders of user \( i \).

\[ T_i^{start} = \text{mode}(S_i) \]

where \( S_i \) represents the set of order begin times for user \( i \) within the \( T \) days.

### 3. PREDICTION MODELING OF CAR-SHARING USERS

The modeling framework of this paper is as follows. First, several new variables are defined to represent the travel behavior of car-sharing users. Second, together with the inherent features of car-sharing system, the RFM model is extended for user value evaluation. Third, the users can be divided into several categories in terms of their values, and the characteristics of each user category can be clearly analyzed. Finally, a prediction model is developed to predict the long-term value category of the users based on short-term data.

#### 3.1. Clustering Algorithm for Classification of Car-sharing Users

For a car-sharing company, the value contributions may be very distinct for the users. Therefore, it requires the company to classify the users based on their values.

Because K-means algorithm is easy to understand and implement, it is applied to classify the users with the five input variables. The Calinski-Harabaz and SSE (sum of squares due to error) measurement
indicators [23, 24] are used to determine the perfect clustering numbers. Together with the clustering performance with various \( k \), the K-means clustering algorithm can be briefly described as follows.

**Step 1:** Randomly generate \( k \) (\( k=2 \)) initial centers within the data domain; set the convergence criterion.

**Step 2:** For each data sample, calculate the distance between the sample and the \( k \) cluster centers. Assign each sample into the nearest cluster.

**Step 3:** Recalculate the center of each cluster and set them as the new clustering centers.

**Step 4:** Repeat steps 2 and 3 until convergence criterion has been satisfied.

**Step 5:** If \( k<K \) (\( K \) is a sufficiently large number), let \( k=k+1 \), and repeat steps 2-4.

**Step 6:** Choose \( k \) with the best clustering performance according to the Calinski-Harabaz and SSE measurements. Output the clustering results.

### 3.2. Prediction Modeling of Car-Sharing Users Based on Random Forest

Due to the lack of understanding of new users (including the travel behavior, user value, and so on), the company is unable to conduct optimized user management strategies. To this end, this subsection attempts to propose a prediction model to predict the long-term value of a new user based on the short-term data.

#### 3.2.1. Algorithm Introduction

The prediction model is developed based on the random forest approach due to its advantages as follows [25]: (1) The random forest approach has sufficiently high accuracy in classification. (2) The importance of each feature vector can be measured by performance indicators such as the Gini index. (3) The random forest approach does not suffer from overfitting.

The basic principle of the random forest algorithm is to establish a lot of decision trees to form a "forest" of a decision tree. Multiple trees are used to make decisions to obtain more accurate and stable predictions [26].

#### 3.2.2. Prediction Model Application

In this study, the random forest algorithm is applied on the modeling for user value prediction.

**Step 1:** Select the features of car-sharing users.

Based on the car-sharing data, the input vector \( I \) consists of the user value features \( V \), personal attributes of users \( A \), and travel behavior characteristics of users \( C \). It is formulated as,

\[
I = [V, A, C]
\]  

The output feature \( O \) is the value type of the user. On the basis, the training data can be extracted. Apply the random forest algorithm to train the prediction model, calculate and output the importance of each input feature.

**Step 2:** Specify the final features of the prediction model.

To simplify the prediction model, and further to improve the practicability and efficiency of the prediction model, we select the most essential feature with sufficiently high importance. Sum of the contributions of the front 14 features is above 98%. Accordingly, only the front 14 features, denoted as \( I_n = [V_n, A_n, C_n] \), are selected as the inputs of the proposed prediction model. Details of the features are listed in Table IV.

**Step 3:** Set the observation duration.

This study takes the order data for 14 weeks. It employs the data of front \( l_k \) \( = k \times \Delta t \) weeks to train the prediction model. Here, \( k (k \in \{1, 2, \ldots, 14\}) \) denotes the number of weeks; and \( \Delta t \) is the number of days in a week. Then, the long-term value type of users can be predicted.

**Step 4:** Training and testing of the random forest model.

The pre-processed data includes 5787 users. 80% of the samples are used for training, and the remaining 20% samples are used for testing. Taking \( k = 1 \) for example, the training process are as follows.
| Feature type               | Input feature   | Value type       |
|---------------------------|-----------------|------------------|
| Value features            |                 |                  |
| Last gap time             | Continuous variable |
| Order amount of a user    | Continuous variable |
| Total consumption amount | Continuous variable |
| Car type                  | Discrete variable |
| Travel behaviour          |                 |                  |
| Shortest distance         | Continuous variable |
| Proportion of weekend travel | Continuous variable |
| Proportion of holiday travel | Continuous variable |
| Proportion of weekday travel | Continuous variable |
| Proportion of short travel | Continuous variable |
| Longest distance          | Continuous variable |
| Station heterogeneity ratio | Continuous variable |
| Average distance          | Continuous variable |
| User attributes           |                 |                  |
| Age                       | Continuous variable |

When $k = 1$, the input feature vector is $I_{n1} = \{V_{n1}, A_{n1}, C_{n1}\}$, and the output is the value type according to the data of 14 weeks. The prepared pre-processed data samples are used to train the random forest model. The prediction accuracy is employed as the performance index of training and testing.

Step 5: Determine the shortest observation duration.

In general, the longer the observation duration is, the higher the prediction accuracy is. Nevertheless, the managers of the car-sharing companies usually want to know the value types of users as soon as possible, so as to adjust the user management strategy in advance. To this end, a threshold prediction accuracy $A^*$ is set in this study. Let $k = k + 1$, and repeat step 4; Until $k = k^*$, then $A(k^*) \geq A^*$ is satisfied. $k^*$ is set to be the observation duration.

Step 6: Predict the value types of new users.

After determining $k^*$, the value type of a new user within 14 weeks can be predicted according to the user attributes and the travel behavior within $k^*$ weeks. In addition, users' travel behaviors may change with time, and users' value types are not immutable. Therefore, the company can use this model to predict the value type of users in any periods of time.

4. CASE STUDY

4.1. User Clustering and Analysis

4.1.1. Clustering Results

Five variables are selected as clustering indicators, and the K-means clustering algorithm is used to classify all users into 2-8 categories. When cluster number is 3, the Calinski-Harabaz value is the largest and the curvature of the first trip of SSE is the largest. Therefore, the perfect cluster number is 3. According to the clustering results in Table V, the three types of users are named as high-value users, low-value users, and potential users, respectively. Detail characteristics of each type of users are discussed as follows.

- High-value users belong to the type with the largest consumption amount and order amount of a user. The last gap time is small. All the results indicate that this type of users is the most active and has the highest value. Nevertheless, the fees of their preferred car types are relatively low, indicating that this type of users is economic sensitivity. There are 585 high-value users, accounting for 10.11% of
the total number. However, they have the largest orders number, which is 37672, accounting for 43.81% of the total order number.

- **Low-value users** have the largest last gap time, and the consumption amount and the order amount are the lowest. In addition, this type of users has long registration duration. Accordingly, this type of users seldom uses the shared cars during the short term, and they may change to be the lose users in the future. There are 2654 low-value users, accounting for 45.86% of the total users. They have a total of 9164 orders, accounting for 10.66% of the total order numbers.

- **Potential users** are regarded to those who have high possibility to become high-value users in the future. The consumption amount and order amount of a user of potential users are between other two types of users. They have short registration duration and similar last gap time with the high-value users. There are 2548 potential users, accounting for 44.03% of the total users; and order number is 39157, accounting for 45.53% of the total orders.

### 4.1.2. Heterogeneous Behavior of Various Users

This section introduces the heterogeneous behavior of different types of users from three aspects: user personal attributes, value attributes, and travel behavior characteristics.

According to the clustering result, one can see that the personal attributes are significantly distinct among the three types of users. 78% of the high-value users are young people with ages not higher than 34; while the proportion is 60% for potential users. In addition, we find that most of the users are male, which is higher than 80% for all three types of users.

From the distribution of average consumption amount per order for different type of users, the distributions of consumption are similar for all types of users. The average consumptions of most users are lower than 25 Yuan. Nevertheless, proportions of low and high consumptions are obviously different.

The characteristics of travel behavior are analyzed in terms of order date, order time, travel order duration and station heterogeneity.

- **Order date**: The date is classified into three categories, i.e., weekday, weekend and holiday. The proportion of high-value users travelling on weekdays is the highest, while that of low-value users is the lowest. In addition, low-value users have the highest proportion of holiday orders.

- **Order time**: For high-value users and potential users, there exist obvious morning and evening peaks. In detail, the morning and evening peaks for car-rental are 7:00-8:00 and 17:00-18:00, respectively; while those for car-return are 8:00-9:00 and 18:00-19:00, respectively. In particular, for the low-value users, there are no car-rental and car-return peaks.

- **Order duration**: For all the three types of users, the durations for most orders are short, which are shorter than 2 hours. Differently, the proportions of short-duration orders for both high-value users and potential users are particularly high, both of which are over 80%. However, the proportion of long-duration orders (longer than 2 hours) is about 34%, which is larger than that of the high-value users and potential users.

- **Station heterogeneity**: Based on the pre-processed data, the proportion of station-heterogeneity order for high-value users is highest, which is 83.0%; while it is lowest for low-value users, only 68.9% of the orders are station-heterogeneity.

In conclusion, according to the analysis of travel behavior of the users, it can be seen that high-value users prefer to use shared cars in commute travel; while low-value users usually choose shared cars in some special cases, and they rarely use shared cars as a commute travel mode. In particular, the potential users have obvious similarity with high-value user in various aspects, such as order date, order time, and so on.

### 4.2. User Value Prediction and Analysis

Based on the pre-processed data, the proposed prediction is trained and tested with different observation durations. Parts of the results are shown in Table VI.
To intuitively understand the impact of observation duration on prediction accuracy, the variations of prediction accuracy with observation duration are plotted in terms of potential users, low-value users, high-values, and all users, as shown in Fig. 1. One can see that as observation duration increases from 1 week to 14 weeks, all curves increase gradually, indicating that the prediction accuracy can be promoted with longer observation duration.

![Variation of prediction accuracy with observation duration](image)

**Figure 1.** Variation of prediction accuracy with observation duration

**TABLE V.** INDICATOR RANGE OF DIFFERENT TYPES OF USERS

| Type                | Total consumption amount for a user (Yuan) | Order amount of a user | Last gap time (day) | Registration duration (day) | Car type | Number of users | Order number |
|---------------------|-------------------------------------------|------------------------|---------------------|-----------------------------|----------|----------------|--------------|
| High-value users    | 366.915-9955.52                           | 9-262                  | 0-74                | 98-822                      | 1, 3     | 585            | 37672        |
| Low-value users     | 5-1558.76                                 | 1-36                   | 15-98               | 98-893                      | 3, 1     | 2654           | 9164         |
| Potential users     | 8.05-1630.23                              | 2-61                   | 0-85                | 98-908                      | 1, 3     | 2548           | 39157        |

**TABLE VI.** PARTIAL RESULTS BASED ON THE PREDICTION MODEL WITH VARIOUS OBSERVATION PERIODS

| Actual | Prediction |
|--------|------------|
|        | 1 week     |
|        | 7 weeks    |
|        | 14 weeks   |
| **User type** | **Protentional users** | **Low-value users** | **High-value users** | **Accuracy** |
| Protentional users | 293 | 206 | 15 | 57.00% |
| Low-value users | 143 | 385 | 6 | 72.10% |
| High-value users | 53 | 20 | 37 | 33.64% |
| Accuracy | 42.23% | 52.76% | 5.01% | 61.74% |
| **User type** | **Protentional users** | **Low-value users** | **High-value users** | **Accuracy** |
| Protentional users | 393 | 110 | 11 | 76.46% |
| Low-value users | 76 | 458 | 0 | 85.77% |
| High-value users | 24 | 1 | 85 | 77.27% |
| Accuracy | 42.57% | 49.14% | 8.29% | 80.83% |
| **User type** | **Protentional users** | **Low-value users** | **High-value users** | **Accuracy** |
| Protentional users | 493 | 19 | 2 | 95.91% |
| Users            | Low-value users | High-value users | Accuracy        |
|------------------|-----------------|------------------|-----------------|
|                  | 26              | 3                | 95.13%          |
|                  | 508             | 107              | 97.27%          |
|                  | 0               |                  | 95.68%          |

In this study, the threshold of prediction accuracy is set to 80%, which not only ensured the reliability of the prediction model, but also prevented excessive costs to achieve high accuracy. As can be seen from Table VI and Fig. 1, regarding to prediction accuracy of all users, the shortest observation duration is 7 weeks. In this case, the prediction accuracies of low-value users, potential users and high-value users are 85.77%, 77.27% and 76.46%, respectively.

In brief, the case study indicates that the proposed approach can be applied to analyze the travel behavior of car-sharing user, and can well predict the value type of the users.

5. CONCLUSIONS
To support car-sharing company in user management, this paper attempts to mine and analyze the travel behavior and values of users based on the order data and user attribute data provided by a car-sharing company. On the basis, several variables are presented and defined to depict the characteristics of users. Furthermore, the RFM model is extended to cluster the users in terms of their value. The clustering results indicate that the car-sharing users can be divided into three categories: high-value users, low-value users and potential users.

Based on the classification of users, the travel behavior characteristics of various types of users are analyzed. Subsequently, a prediction model is developed based on random forest to predict the long-term value type of a new user based on short-term data. Based on the pre-processed data, the prediction model is trained and tested. By setting long-term duration 14 weeks, the results indicate that the prediction accuracy can reach over 80% with observation duration of 7 weeks.

In this study, the characteristics of car-sharing users are investigated based on empirical data, and a model is developed to predict the value types the users. Nevertheless, there are still many issues that are not clearly addressed, such as the distribution of orders in time and space, the variation of value type for a user with duration. Future studies will focus on these topics, and apply some other approach for prediction modeling to promote the prediction accuracy.

ACKNOWLEDGMENTS
This work was partially supported by the National Natural Science Foundation of China (71961137008, 71621001, 71671014, 91846202, 71771021).

REFERENCES
[1] Jung, J., Koo, Y.: 'Analyzing the effects of car sharing services on the reduction of greenhouse gas (GHG) emissions', Sustain., 2018, 10, (2), pp. 1–17.
[2] Mounce, R., Nelson, J.D.: 'On the potential for one-way electric vehicle car-sharing in future mobility systems', Transp. Res. Part A Policy Pract., 2017, 120, pp. 17–30.
[3] Illgen, S., Höck, M.: 'Literature review of the vehicle relocation problem in one-way car sharing networks', Transp. Res. Part B Methodol., 2019, 120, pp. 193–204.
[4] Schlüter, J., Weyer, J.: 'Car sharing as a means to raise acceptance of electric vehicles: An empirical study on regime change in automobility', Transp. Res. Part F Traffic Psychol. Behav., 2019, 60, pp. 185–201.
[5] Zhou, F., Zheng, Z., Whitehead, J., et al.: 'Projected prevalence of car-sharing in four Asian-Pacific countries in 2030: What the experts think', Transp. Res. Part C Emerg. Technol., 2017, 84, pp. 158–177.
[6] Kim, J., Rasouli, S., Timmermans, H.: 'Satisfaction and uncertainty in car-sharing decisions: An integration of hybrid choice and random regret-based models', Transp. Res. Part A Policy Pract., 2017, 95, pp. 13–33.
[7] Perboli, G., Caroleo, B., Musso, S.: 'Car-Sharing: Current and Potential Members Behavior Analysis after the Introduction of the Service', Proc. - Int. Comput. Softw. Appl. Conf., 2017, 2, pp. 771–776.

[8] Kim, J., Rasouli, S., Timmermans, H.J.P.: 'The effects of activity-travel context and individual attitudes on car-sharing decisions under travel time uncertainty: A hybrid choice modeling approach', Transp. Res. Part D Transp. Environ., 2017, 56, pp. 189–202.

[9] Fleury, S., Tom, A., Jamet, E., et al.: 'What drives corporate carsharing acceptance? A French case study', Transp. Res. Part F Traffic Psychol. Behav., 2017, 45, pp. 218–227.

[10] Yoon, T., Cherry, C.R., Jones, L.R.: 'One-way and round-trip carsharing: A stated preference experiment in Beijing', Transp. Res. Part D Transp. Environ., 2017, 53, pp. 102–114.

[11] Jain, T., Johnson, M., Rose, G.: 'Exploring the process of travel behaviour change and mobility trajectories associated with car share adoption', Travel Behav. Soc., 2020, 18, pp. 117–131.

[12] Becker, H., Ciairi, F., Axhausen,K.W.: 'Modeling free-floating car-sharing use in Switzerland: A spatial regression and conditional logit approach', Transp. Res. Part C Emerg. Technol., 2017, 81, pp. 286–299.

[13] Rotaris, L., Danielis, R., Maltese, I.: 'Carsharing use by college students: The case of Milan and Rome', Transp. Res. Part A Policy Pract., 2019, 120, pp. 239–251.

[14] Schmöller, S., Weikl, S., Müller, J., et al.: 'Empirical analysis of free-floating carsharing usage: The munich and berlin case', Transp. Res. Part C Emerg. Technol., 2015, 56, pp. 34–51.

[15] Mueller, J., Schmoeller, S., Giesel, F.: 'Identifying Users and Use of (Electric-) Free-Floating Carsharing in Berlin and Munich', IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC, 2015, pp. 2568–2573.

[16] Ye, J., Wang, D., Zhang, H., et al.: 'What Kind of People Use Carsharing for Commuting? Case Study in Shanghai', Transp. Res. Rec., 2019, 2673, (5), pp. 770–778.

[17] Qian, C., Li, W., Ding, M., et al.: 'Mining Carsharing Use Patterns from Rental Data: A Case Study of Chefenxiang in Hangzhou, China', Transp. Res. Procedia, 2017, 25, pp. 2583–2602.

[18] Hui, Y., Wang, W., Ding, M., et al.: 'Behavior Patterns of Long-term Car-sharing Users in China', Transp. Res. Procedia, 2017, 25, pp. 4662–4678.

[19] Hui, Y., Liu, C., Ding, M.: 'Carsharing trip characteristic analysis: Do users choose carsharing rather than taxi to economize?', Transp. Res. Rec., 2018, 2672, (42), pp. 115–126.

[20] Chen, X., Cheng, J., Ye, J., et al.: 'Analysis of carsharing users and demand spatio-temporal characteristics', Journal of Tongji University (Natural Science), 2018, 46, pp. 796–803.

[21] Sai, Q., Xie, D., Bi, J., et al.: 'Capturing the characteristics of carsharing users based on a data-driven method', J. Phys. Conf. Ser., 2019, 1168, (3).

[22] Sai, Q., Bi, J., Xie, D., et al.: 'Identifying and predicting the expenditure level characteristics of car-sharing users based on the empirical data', Sustain., 2019, 11, (23).

[23] An, J., Yan, Z., Zhai, J.: 'K-means Clustering Algorithm Based on Distance Threshold and Weighted Sample', Microelectronics & Computer, 2015, 32, pp. 135–138.

[24] Guo, J., Hou, S.: 'Study on the index of determining the optimal clustering number of K-means algorithm', Software Guide, 2017, 16, (11), pp. 5–8.

[25] Feng, T., Timmermans, H.J.P.: 'Comparison of advanced imputation algorithms for detection of travel mode and activity episode using GPS data', Transp. Plan. Technol., 2016, 39, pp. 1–15.

[26] Ziegler, A., König, I.R.: 'Mining data with random forests: Current options for real-world applications', Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 2014, 4, (1), pp. 55–63.