Capturing the characteristics of carsharing users based on a data-driven method

Qiuyue Sai1, Dongfan Xie1,a, Jun Bi1, and Fujun Ding2

1MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing 100044, China
2Gansuyixiangxing New Energy Developments Ltd, Lanzhou 730000, China

Abstract. In order to capture the characteristics of carsharing users, we applied a data-driven method to classify the users into typical clusters according to the contributions to the carsharing companies. The data used in this study are collected in the order information from the carsharing company in Beijing, China. The RFM model is employed to determine the influencing factors for clustering analysis. Accordingly, the most recent consumption, consumption frequency and spending are adopted to represent the user value. The users can thus be classified into three groups, named as potential users, high value users and the loss of users. It is expected that the results can improve the user management of the carsharing company.

1.Introduction
In recent years, population explosion and resource shortages have contributed to the increasing development of sharing economy [1]. Carsharing is the representative of sharing economy in the transportation field, which has significant effects to reduce traffic congestion and environmental pollution. Carsharing companies have enjoyed increasing development due to the public concern on resource shortages and encouragement from government sectors. As the number of carsharing users increased, identifying the value of different users becomes a critical issue for carsharing company. When determining the car rental schemes, the user value has significant impacts on the benefit of carsharing companies. Therefore, attention should be payed to the classification of carsharing users.

Existing literatures have been conducted to develop effective approaches to improve the diffusion of carsharing [2-3]. The impacts of carsharing on traffic environment were discussed in the studies. Considering the user travel characteristics in particular areas, the potential demand and market of carsharing were assessed [4-5]. the selection intention for carsharing was analyzed using questionnaires [6-7]. the multiple discrete-continuous extreme value models were proposed by considering the travel time, travel distance and travel costs [8]. However, these studies do not discuss the differences of car rental behavior among different carsharing users. As a matter of fact, different users have different contribution to the carsharing companies due to their different car rental behavior, which should be considered to improve the operation efficiency of carsharing companies.

In this study, we use a great quantity of field data to explore the car rental behavior of carsharing users. Based on the influencing factors provided by the data, the carsharing users are classified into several groups by using a data-driven method. The results may be used by the carsharing company operators to identify the users with different values.
2. Data collection and processing
The data used in this study are collected from the order information of Beijing Travel Auto Service Co., Ltd. which is a carsharing company in Beijing, China. The company has 1,000 vehicles and 170 parking spaces to provide carsharing services. As the users send their car rental requests to the company by using their mobile phones, the data of order information are generated and stored in the database. The data online record the car rental information of users, such as travel distance, travel duration and travel costs.

The dataset used in this study includes 81,455 data points. To clean up the raw data, we remove the duplicated and incorrect data points. After data processing, 70,391 data points are retained. The experimental data records the car rental behavior of 11,473 car sharing users and 958 vehicles during the period from December 17, 2016 to October 16, 2017.

3. Methodology
The clustering method is employed to classify the carsharing users based on the field data. The users with similar car rental behavior are clustered into the same user group. In general, the greater differences among the different groups would contribute to the better classification results for the carsharing users. To perform the user clustering process, it needs to select the variables that can reflect the preference of user orders. Then, clustering algorithm can be applied to classify the users.

For carsharing companies, the valuable users creation for the business is crucial. The RFM model is an classical model for measuring user value and user profitability [9]. There are three important elements in RFM, the most recent consumption, consumption frequency, and spending. The most recent consumption represents the time span that a user experiences between the current car rental request and the previous one. It is obvious that a shorter time span presents the user has better response to the carsharing services. The consumption frequency can reflect the usage times of the carsharing services during a certain period. The users who have more usage times would have higher satisfaction for the carsharing services. The spending is the critical influencing factor for the operation of carsharing companies.

In this study, the K-means algorithm is used to classify the users [10]. According to many experiments, we believe that the number of clusters is 3, which can best distinguish customers. So, in K-means algorithm, the number of clusters is three in this study. Each user data object includes the latest consumption time, the consumption frequency, and the consumption amount. The difference between the objects can be calculated by the distance between the two objects. The distance formula used here is the Euclidean distance,

\[ d(i, j) = \sqrt{(x_{ip} - x_{jp})^2 + \cdots + (x_{op} - x_{op})^2} \] (1)

where \( x_{ip} \) and \( x_{jp} \) are the factors of user \( i \) and user \( j \).

Meet the conditions
- \( d(i, j) \geq 0 \) represents a value that is not negative between objects.
- \( d(i, j) = 0 \) represents the distance between the object itself is zero.
- \( d(i, j) = d(j, i) \) represents the distance between objects is a symmetric function.
- \( d(i, j) \leq d(i, n) + d(h, j) \) represents if the distance between the two objects is greater than the third edge, the distance between the two objects is expressed by one edge, and \( h \) is the third object.

The objects are then assigned to the nearest cluster. After all the objects have been allocated, the centers of \( k \) clusters are recalculated. The result is output until there is no change to the cluster center from the previous \( k \) cluster centers. The clustering criterion function used in this study is the square error sum criterion, and the sum of squared errors is used to sample distributions in which all kinds of samples are intensive and the number of samples is small.
Where, $m_j$ is the mean of the sample in type and the center of the c set, which can be used to represent c types.

4. Results and discussion

According to the results of the clustering algorithm in the previous section, we draw a pie chart showing the proportions of all kinds of users. As shown in Figure 1, the largest cluster contains 7,698 users, accounting for 67.1% of the total users. The smallest cluster contains 987 users, accounting for 8.6% of the total users. The third category accounts for 24.3%.

![Pie chart showing user proportions](image)

Figure 1. Proportions of users in three user groups

As shown in Figure 2, the three types of users are plotted against all the parameters of the analysis of the overall user's relationship. In the graph, each column represents a class of users. Columns 1, 2, and 3 correspond to the first class of users, the second class of users, and the third class of users, respectively. Each row of the graph represents a variable, where the first row indicates the frequency, the second row indicates the amount of consumption, and the third row shows the time difference between the last consumption and the current date. The white square in the figure shows the distribution of the overall object parameters. The middle line of the square box represents the median and the positions of the boxes correspond to the upper and lower quartiles of the data respectively. The blue dot in the middle indicates the median distribution under a certain parameter of a certain type of user, and the blue horizontal bar indicates the upper and lower quartile range of the parameter of this type of user.

![Feature distribution box chart](image)

Figure 2. Feature distribution box of the three user groups

It can be concluded from Figure 2 that the frequency distributions of first and third user classes are approximately equal to the overall frequency distribution. The frequency distribution of the second user class is greater than the overall frequency distribution, indicating a high number of trips. The first and third types of users’ total average consumption are equal to the total average amount of consumption. The total average consumption of the second type of users is greater than the total
amount of consumption. So the second type of users will spend more on car rental. The time difference between the last travel of the first type and the current time is slightly smaller than the overall last travel time compared with the current time difference. The last travel time difference of the second type of users is smaller than the overall travel time. It is obvious that the second type of users’ travel time difference is less than the first type users’. The third type of users last travel time difference is greater than the overall travel time difference. We can see that the second and third type of users have recently rented the corporate car, and the third type of users have used the car for a long time.

Table 1. Three-parameter summary table of three user groups.

|                  | Frequency | Money  | R      |
|------------------|-----------|--------|--------|
|                  | number    | Average| median | variance | sum     |
| potential users  | 7700      | 4.38   | 3.00   | 17.053   | 33716   |
|                  | 7700      | 384.70 | 235.00 | 411.373  | 2962166 |
|                  | 7700      | 123.31 | 125.00 | 25.462   | 949475  |
| high value users | 989       | 29.17  | 25.00  | 250.403  | 28853   |
|                  | 989       | 2910.74| 2334.00| 3693534.087| 2878726|
|                  | 989       | 95.99  | 86.00  | 427.151  | 94935   |
| Losing users     | 2787      | 2.81   | 2.00   | 12.459   | 7821    |
|                  | 2787      | 256.05 | 127.00 | 187147.787| 713604  |
|                  | 2787      | 220.05 | 210.00 | 1593.216 | 613275  |

The first type of users are called as potential users. Who accounted for 67.1% of the total number of people, per capita use 4 times, per capita income 384.7 and total revenue is 2962166 account for 45.2% of the total revenue.

The second type of users are named as high value users, accounted for 8.6% of the total number of people, per capita use of 29 times, per capita income 2910.74 and total revenue is 2878726 account for 43.9% of total revenue.

The third type accounts for 24.3% of the total number of users. Per capita’ use frequency is 3 times. The last travel time from the current date is 220 days per capita. The third type of users create revenue 220 yuan per capita, total revenue 713,604, accounting for 10.9% of total revenue. Accordingly, the third type is named as the loss of users.

5. Conclusions

The purpose of this study is to analyse consumers’ characteristics according to different consumer groups by clustering optimization method. The result of this paper is helpful for the enterprise to analyze the changes in the market and make the marketing plan.

This paper divides users into three categories. Potential users accounts for 67% of the total, accounting for the largest proportion. We believe that the large proportion of such users is due to lack of marketing and a large number of potential users worth further development. Such users are expected to develop into high value Users. The companies should take measures to try to retain.

High-value users is accounting for 8.6% of the total, accounting for a small proportion of 44.6% of the total revenue generated. This type of users is expected to be more active and have higher loyalty so as to create greater value for the enterprise.
The third type of users accounting for 30% belong to the loss of users. The value created by this kind of users accounts for 10.9% of total revenue. The loss of the third type of users is because that the leasing services that did not meet user expectations or the user just try to use it before the promotional stop. For the third type of people should be issued questionnaires and find out the reasons for the loss.

With the continuous development of shared economy, carsharing will be more widely used in the future. Carsharing companies will also have greater space for commercial development. If we can effectively use data mining methods, we can provide decision-makers with accurate users’ classification, loyalty, profitability and other useful information. We can guide companies to develop optimal corporate marketing strategy which can increase profits and accelerate the development of enterprises. In this paper, the users of shared car are divided into three categories through cluster analysis. Among them, high value users create half of the total profits, the company should focus on this part of users. 67% of potential users indicates that a large number of users of this enterprise can continue to develop. This type of customers is expected to become a high-value user and create greater profits for the business. Therefore, based on the results of this paper, we can make customer-centred decisions for carsharing companies, so that shared car companies can be better developed.

Acknowledgments
This work was supported by the National Key R&D Program of China (2018YFC0706005, 2018YFC0706000).

References
[1] Hamari, J., Sjöklint, M., Ukkonen, A., Journal of the Association for Information Science & Technology, 67(9), 2047-2059.(2016)
[2] Fellows, N. T., Pitfield, D. E., Transportation Research Part D Transport & Environment, 5(1), 1-10. (2000)
[3] Correia, G. H. D. A., Antunes, A. P., Transportation Research Part E: Logistics & Transportation Review, 48(1), 233-247. (2012)
[4] Zheng, J., Scott, M., Rodriguez, M., Sierzchula, W., Platz, D., Guo, J. Y., Adams, T. M., Transportation Research Record, 2110, 18–26. (2009)
[5] Celsor, C., Millard-Ball, A., Transportation Research Record, 19.(2006)
[6] Firnkorn, J., Müller, M., Ecological Economics, 70(8), 1519-1528. (2011)
[7] Wappelhorst, S., Sauer, M., Hinkeldein, D., Bocherding, A., Glaß, T., Transportation Research Procedia, 4(1), 374-386.(2014)
[8] Jian, S., Rashidi, T. H., Dixit, V., Transportation Research Part A: Policy & Practice, 103, 362-376.(2017)
[9] Wei, J. T., Lin, S. Y., Wu, H. H., African Journal of Business Management, 4(19), 4199-4206. (2010)
[10] Tao, X. M., Xu, J., Yang, L. B., Liu, Y., Journal of Electronics & Information Technology, 2010(1), 92-97.(2010)