Prediction of silent users of car-sharing based on Logistic Regression Model

Jun Bi¹, ², Zun Yuan¹*, Qiuyue Sai¹ and Dongfan Xie¹

¹School of Traffic and Transportation, Beijing Jiaotong University, Beijing, 100044, China
²Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Beijing Jiaotong University, Beijing, 100044, China
*Corresponding author’s e-mail: 17120927@bjtu.edu.cn

Abstract. Car-sharing is a new transportation mode with the rapid development of mobile Internet. It is necessary for the operation companies to keep the number of carsharing users and intelligent transportation can be realized by discovering the users before they loss and preventing the loss. In this paper, based on the massive order data of a car-sharing company in Beijing, the behavior of the users is analyzed before they become silent. In the light of Logistic Regression Model, the loss rate of users in next month is predicted based on the car-renting behavior in the former three months, and some suggestions are proposed to prevent the loss. The result shows that there is a significant difference in the behavior between the lost users and non-lost users. The accuracy of the prediction is 97.9%. The method proposed in the paper provides a theoretical basis for enterprises to deal with the early warning and recall of the lost users.

1. Introduction
Car-sharing is an emerging mode of transportation, also known as time-sharing car rental, which enables users to use mobile APP to reserve the car in real time, pick up and return the car in different locations. It not only improves the users’ travel experience but also relieves the issue of urban traffic congestion. However, with the development of car-sharing and the increase of users, it is inevitable to face the loss of old and new users. Silent users denote the users who change from active state to inactive or lost state, and in this study, they refer to the users who have orders in the first three months but have no order in the fourth month. To prevent the loss of the users and promote the profit of carsharing companies, it is necessary to make a prediction of the silent users and identify the users before they loss, so that effective measures can be taken in advance.

A great deal of researches about the carsharing users are carried out in the last decades. On the one hand, some researches focus on the travel characteristics of users and the travel selection based on the travel requirements. In [1], the travel selection of car-sharing users is studied from family economic conditions, travel costs, car demand, personal attributes and other aspects. Based on the travel data of carsharing users in Germany and Canada, the linear regression method is used to predict the reservation demand [2], and it is found that the factors of family status, education level as well as mobility have a significant impact on users’ choice of carsharing behavior by building a binomial regression model [3]. Also, in order to analyze the main factors affecting the travel of car-sharing users, a multiple linear regression model is established by using the transaction data provided by car-
sharing operators in Seoul, South Korea. It takes the travel volume as the dependent variable and the environment, population and traffic conditions as independent variables [4]. On the other hand, some researches concentrate on modeling and predicting the users’ travel selection in space by analyzing the individual user’s travel behavior and environmental influence. In [5], a dynamic model of daily path selection based on Dogit model is established to analyze the influence of individual path preference on the evolution trajectory of daily path traffic. To improve the accuracy of location prediction of the single user, a Markov location prediction method based on mobile behavior similarity and user clustering is proposed [6]. The above researches analyze the relevant factors affecting users' travel and establish the corresponding model to make a prediction. However, it remains a problem on how to locate and identify the users and estimate the probability of the user loss rate from the perspective of time. And that is our work in this paper.

In this paper, based on the data of a carsharing operation company in Beijing, the car-renting behavior of the silent users is analyzed and a Logistic Regression Model is built to predict the loss of users. It contributes much to the decision making and operation optimizing for the carsharing company. The results can serve as a reference for the operation company to adopt different operation strategies for different target users, and thus the benefit of the enterprise can be maximized. The research in this paper plays a vital role in promoting the development of car-sharing.

This paper is organized as follows. The background, the significance and the content of the research are introduced in section 1. The data used in the research are described in section 2. In section 3, the behavior patterns of silent users are analyzed based on the data. In the following, the model of predicting the loss of users is introduced and built in section 4. Finally, some conclusions close this paper.

2. Data source
The data studied in this paper are come from a carsharing operating company in Beijing. Up to October 2018, this company owns more than 1,000 new-energy vehicles and 171 outlets. From January 2017 to October 2018, 128,000 pieces of order data and 20G track data are collected. The data include the order number, user number, car return time, car rental mileage, car rental duration, fees, discount amount and GPS data of the order. For the company is in the initial operation stage, it always issues a large number of coupons for various amounts in the form of price reduction and discount. Since the original data contain a large number of data unrelated to the research content, it is necessary to filter the original data. Consequently, six data including time, user number, car rental time, mileage, fees and discount amount are selected for the following analysis and modeling in this paper.

3. Analysis of the behavior patterns of silent users
The silent users is defined as the users who have orders in the first three months and have no order in the fourth month. In this paper, the carsharing company owns 35 thousands users, in which silent users accounts for 39% and non-silent users accounts for 61%. The number of silent users are huge and it can not be ignored. So it becomes a key issue for the enterprise to study the silent users’ behavior and predict whether the user would be silent in advance.

Firstly, the order distribution of the silent users and non-silent users in the first three months is presented in Table 1. It shows that the number of silent users who place orders less than five times accounts for 71.49% while that of the non-silent user accounts for only 24.45%. In other intervals of the order number, the proportion of non-silent users is larger than that of silent users. It indicates that silent users have a much less car renting orders than non-silent users.

| Order number | Non-silent user | Silent user |
|--------------|----------------|-------------|
| 1-5          | 24.45%         | 71.49%      |
| 6-10         | 15.57%         | 12.90%      |

Table 1. Distribution of user orders in the first three months.
Then, the cumulative survival function of silent users is studied, which is shown in Figure 1. The survival probability of users drops from 1 to 0.6 in the first 30 days. And when the user's survival time reaches about 250 days, the survival probability drops to 0.3 and remains stable.

In addition, the car-renting behavior of the silent users is analyzed shown in Table 2. The total cost and discount cost of silent users are significantly higher than that of non-silent users, and the proportion of discount amount of silent users is higher than that of non-silent users. It can be seen that silent users tend to travel long distances and are more sensitive to coupons.

4. Model of predicting silent users

Due to the much difference of the behavior between silent users and non-silent users, it provides a reference for the subsequent prediction of silent users. In this section, Logistic Regression Model is used to make a prediction of whether the users would be silent in the next month. By predicting in advance, effective measures can be taken in time to prevent the user loss.
4.1. Modeling

Logistic Regression is a kind of the classification prediction algorithm. It forecasts the probability of future results based on the performance of historical data, which has the advantages of simple principle and fast computation speed. In this paper, whether the user would be silent (in other words, whether the user would rent the car) is set as the dependent variable, and the car rental behavior of the user in the last three months, such as the order number and the average amount of each order, are set as independent variables. By building a model and analyzing the users’ car rental behavior, the car-renting orders or the silent users in next month can be predicted.

Logistic Regression Model is a probabilistic nonlinear regression model, which is also a multivariate analysis method to study the relationship between classification result (y) and some influencing factors (x). Sigmoid function is adopted as the predicting function and it associates the attribute variables in terms of probability which has the value between 0 and 1. The function is expressed in (1) and the corresponding functional curve is presented in Figure 2.

\[
g(x) = \frac{1}{1 + e^{-x}}
\]  

(1)

where \(x\) is the input of the Sigmoid function, \(g(x)\) is the probability. It can be seen in Figure 2 that the value of \(g(x)\) is 0.5 when \(x\) is zero. The output is closer to 1 when the input become larger and inversely the output is closer to 0 when the input become smaller. This is exactly the predictive function that we want for the binary classification algorithms.

By combining with the linear regression function which is defined as (2), the prediction function of Logistic Regression Model is derived as follows (3).

\[
h_{\theta}(x) = \theta^T x
\]  

(2)

\[
h_{\theta}(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}
\]  

(3)

where \(\theta\) is the parameter vector, \(x\) is the independent variable. To obtain the parameter vector of Logistic Regression Model, the cost function of logistic regression is defined as (4).

\[
L(\theta) = - \frac{1}{N} \sum_{i=1}^{N} \left[ y_i \log\left( \frac{1}{1 + e^{-\theta^T x_i}} \right) + (1 - y_i) \log\left( \frac{1}{1 + e^{-\theta^T x_i}} \right) \right]
\]  

(4)

It is supposed to get the lowest value of the cost function and calculate the corresponding \(\theta\) and \(g(x)\). The input of \(x\) is classified by the value of output which is changed into 0 or 1 based on a defined threshold value. In this way, a process of the prediction is finished.

In this paper, Logistic Regression Model is used to predict whether users are silent in the fourth month according to their behaviors in the first three months. The modeling data are all based on users themselves. The output ‘y’ represents whether the user would be silent, which denotes the user is silent.
when ‘y’ equals to one and not silent when ‘y’ equals to zero. The input ‘x’ is the user’s behavior data that include eight indexes, which are the number of orders, average cost per order, average discount amount per order, average using time per order, average mileage per order, survival days, the longest distance and the shortest distance. To fit the model, the behavior data of users in four months is taken as a basis. The data in the first three months are the input of the prediction model, and the data in the fourth month are served as a result data.

4.2. Results
To test the model and make the model fit the practical data well, the users are separated into the training set and the testing set. The accuracy of the model, which denotes the ratio of the right-predicting users and the total users, is tested using the testing set. The accuracy of the model in this paper is 97.9%. The confusion matrix is presented in Table 3.

| Predicting results | 0     | 1     |
|--------------------|-------|-------|
| Actual results     | 3531  | 159   |
|                    | 32    | 5130  |

5. Conclusion
Taking the car-sharing users as the research object, the behavioral differences between silent users and non-silent users are analyzed in this paper. A Logistic Regression Model is built to predict whether users would be silent in the next month, and the prediction accuracy is 97.9%. It provides a reference for enterprises to reasonably prevent the loss of users and better manage users. According to the preference of using coupons by silent users, carsharing companies can develop a reasonable coupon distribution strategy to retain the silent users and improve user stickiness. In this paper, only the prediction of silent users is researched. The solution of preventing the users become silent is not discussed. This will be our future work.

Acknowledgments
This paper is sponsored by National Natural Science Foundation of China (71961137008).

References
[1] Q.J. Dai. (2012) Analysis of car-sharing policy based on consumer choice behavior. Journal of changchun university of science and technology (social science edition), 25(11): 106-107.
[2] Schmöller S, Weikl S, Müller J, et al. (2015) Empirical analysis of free-floating carsharing usage: The Munich and Berlin case. Transportation Research Part C, 56: 34-51.
[3] Coll M, Vandersmissen M, Thériault M. (2014) Modeling spatio-temporal diffusion of carsharing membership in Québec City. Journal of Transport Geography, 38: 22-37.
[4] Kang, J., Hwang, K., Park, S. (2016) Finding factors that influence carsharing usage: Case study in seoul. Sustainability, 8(8): 709.
[5] L.J. Tian, X.L. Jiang, T.L. Liu, Y.X. Zhao. (2016) Research on daily travel behavior considering route preference based on dogit model. Transportation system engineering and information, 16(6): 228-235.
[6] S.K. Lin, S.Z. Li, J.Z. Qiao, D. Yang. (2016) Markov location prediction based on user mobile behavior similarity clustering. Journal of Northeastern University (natural science edition), 37(3): 323-326.