Demand Forecasting of a Public Bike-Sharing System Reflecting Dynamic Spatial Data

Jiwon Lee*, Jiyoung Kim**, Kiyun Yu*

* Seoul National University, {vert802, kiyun}@snu.ac.kr
** Konkuk University, elliekim@konkuk.ac.kr

Abstract. Bike sharing is booming in Korea as a leisure traffic mode, but its original purpose was to reduce traffic congestion. This study developed a demand forecasting model for bike sharing connected to a subway station. For accurate demand forecasting, we used various kinds of data to reflect the spatial distribution of travel demand. We used machine learning method as prediction model, and random forest had the best predictive result.

Key words. Public Bike-sharing System, Social Media, Random Forest

1. Introduction

Bike-sharing systems are one of the useful solutions for traffic congestion and last-mile problem in urban transportation systems (Xu et al., 2018). Also, since the MaaS (Mobility as a Service) model emerged as a global trend, the status of bike sharing is increasing.

We aimed to develop a demand-forecasting model for bike sharing near subway stations, which are familiar to users as public transportation. Lots of temporal and spatial data are required for accurate demand forecasting, but previous studies were biased toward one side of the data. Time-focused studies made predictions mainly using weather data (Campbell et al., 2016 ; Gebhart and Noland, 2014), and space-focused studies considered only the influence of the static state (Faghih-Imani et al., 2014).

New forms of micro-level data have been found to contain rich information about place semantics and individual interactions with the physical world (Liu et al., 2015). In this paper we actively used various kinds of data to overcome the limitations of previous studies. In order to index the diversity of space in terms of time, we used social media data and used real-time smart-card data to consider linkage with public transportation.

Published in "Adjunct Proceedings of the 15th International Conference on Location Based Services (LBS 2019)", edited by Georg Gartner and Haosheng Huang, LBS 2019, 11–13 November 2019, Vienna, Austria.

This contribution underwent double-blind peer review based on the paper. https://doi.org/10.34726/lbs2019.51

2. Research Methods

2.1. Influential Factors

The factors affecting public bicycle demand can be classified into weather, public transportation, and land use. Unlike other traffic modes, weather influences the use of bicycles, so various types of data, such as temperature, humidity, wind speed, rainfall, and snowfall, were collected.

Public transportation factors were constructed using smart card data, which record the number of people who get off at subway station near a bike dock. They may use a public bike for the last mile to final destination.

Finally, we collected land-use factors reflecting the dynamic state of space by time zone. After defining a purpose for a public-bike trip, we collected data corresponding to each rental demand. The bike-trip purposes were divided into work, school, home, and recreational trips. These were selected by referring to Household Travel Surveys (HTS) items. In order to reflect the dynamic state of space, we extracted demand quantities for each trip purpose and then multiplied the distribution by the time period.

Work, school, and home trips were based on the distribution chart of departure times for each purpose provided by HTS. Table 1 shows the distribution of departure times by trip purpose. For example, demand for work trips by time zone was calculated by multiplying the number of employees in a district by the distribution value by time zone provided by HTS. In the same way, we calculated school trip was by using the number of students, and home trips were calculated using population (ages 15 to 64).

Time zone	**0-1h**	**....**	**7-8h**	**8-9h**	**....**	**12-13h**	**....**	**18-19h**	**19-20h**	**....**
Work trip	0.1%	....	32.7%	36.4%	....	0.7%	....	0.2%	0.1%	....
School trip	0.0%	....	26.9%	54.1%	....	0.5%	....	0.1%	0.0%	....
Home trip	0.2%	....	0.2%	0.3%	....	2.4%	....	21.5%	14.9%	....

Table 1. Departure time distribution by trip purpose (HTS in Korea, 2018).

However, the distribution of recreational trips, unlike other trips, which had an approximate trend, differed by POI (Point Of Interest). So we used check-in data provided by Google Place to figure out the distribution by time. Since check-in data is not accurate data reflecting the real world, standardization is required. For example, place A has more visitors than B in reality, but B may have more check-in data, because social media is user-dependent data. Another thing to consider is distance; since many bike docks can be adjacent to the same POI, we calculated the distance from dock to POI as a weight value. For a shorter distance, the more demand is calculated.

2.2. Prediction Model based on Machine Learning

There are various methods for machine learning to predict demand. In this study, bicycle demand was predicted by using random forest and gradient boosting, which has proved to be effective in various data sets of classification and regression problems.

Both algorithms use a decision tree as a basic element for constructing a model, and it is an ensemble technique that creates a powerful model by grouping several decision trees. Decision trees have an advantage of being fast, and easy to explain, because visualization is possible. Random Forest uses aggregation as a model-combining method. Gradient boosting compensates for the error in the binary tree by using a boosting method that gradually increases the model to be used. The overall analytical process is illustrated in Figure 1.

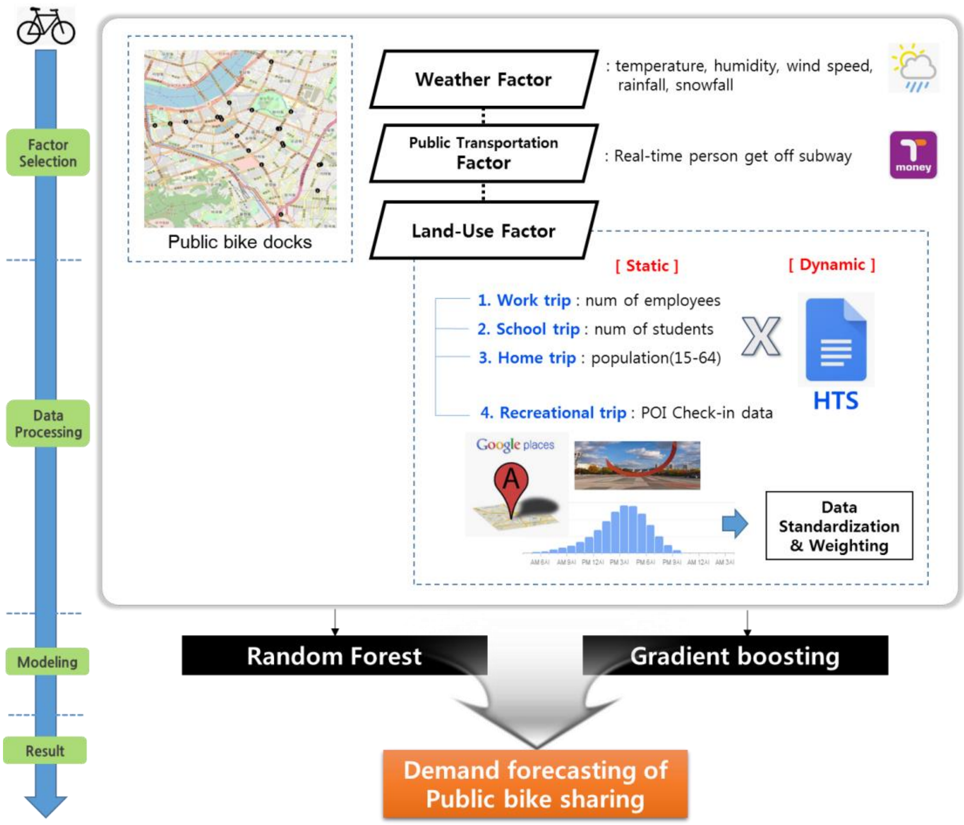

Figure 1. Study flow chart.

2.3. Accuracy assessment

In order to evaluate the performance of the proposed prediction model, we selected Root Mean Squared Logarithmic Error(RMSLE) as the evaluation parameter shown in Equation(1).

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2} \qquad (1)$$

where n is the number of hours in the test set ; p_i is predicted count ; a_i is the actual count ; And $\log(x)$ is the natural logarithm.

This method gives a penalty to underestimated items rather than overvalued items, and the closer the value is to 0, the higher the precision. Also, in order to measure the generalization performance, we did k-fold cross validation to learn multiple models by repeatedly dividing the data.

3. Experiment and Result

The experimental area was limited to Songpa-gu, Seoul, and the rental history of 20 public bike docks near a subway station was analyzed. Songpa-gu has a variety of densely populated residential and business areas, close to the bike-riding Han River and parks, so is a good place to analyze bike demand.

From September to December 2018, 130,783 items of rental-history data were collected for four months. Figure 2 provides the detailed of data. The data for days 1-24 of a month were used as train data and data for days 25-30(31) were used as test data. Data were collected on an hour-by-hour basis and classified as weekday or weekend.

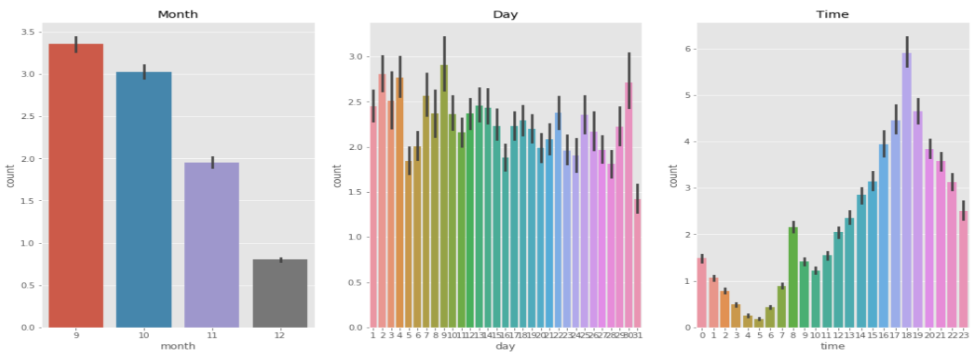

Figure 2. Rental history of public bike in research area

Experimental results show that the demand forecast using Random Forest is 0.1440 for RMSLE and 0.3258 for prediction through Gradient Boosting.

Random Forest's prediction performance is better. In addition, the model that reflects the dynamic state of public transportation and space works better than the model that uses only weather information. Table 2 shows the RMSLE results for each model.

Models	Random Forest	Gradient Boosting
(Weather + Public Transport + Land use) factors	0.1440	0.3258
(Weather + Public Transport) factors	0.2806	0.4053
(Weather) factors	0.4544	0.4588

Table 2. RMSLE result of different models

4. Conclusion

This study developed a demand forecasting model for bike sharing near subway stations using machine learning. Spatial and temporal data were introduced into the forecasting demand. In particular, various kinds of data such as smart cards and social media data were used to reflect the spatial distribution of travel demand. By integrating those data with general information, the method can be predicted more accurately bike-sharing demand.

Spatial and temporal data has information that static data does not have, which allows decision-makers to gain new insights. Especially user-created content on social media provides details for topics such as customer satisfaction, and travel behavior (Welch et al., 2019). In a future study, we plan to develop customized demand-prediction models by analyzing user opinions composed of unstructured data from social networks.

Acknowledgement

This study was supported by the research funding of the project on the development of big data management, analysis, and service platform technology for the national land spatial information research project of the Ministry of Land, Infrastructure, and Transport (19NSIP-B081011-06).

References

Campbell A, Cherry C, Ryerson M, Yang X (2016) Factors influencing the choice of shared bicycles and shared electric bikes in Beijing. Transportation Research Part C 67: 399–414.

Faghih-Imani A, Eluru N, El-Geneidy A, Rabbat M, Haq U (2014) How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. Journal of Transport Geography 41: 306-314

Gebhart K, Noland R (2014) The Impact of weather conditions on Bikeshare trips in Washington, D. C. Transportation 41(6): 1205–1225

Liu Y, Liu X, Gao S, Gong L, Kang C, Zhi Y, Chi G, Shi L (2015) Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. Annals of the Association of American Geographers 105: 512-530

Welch T, Widita A (2019) Big data in public transportation: a review of sources and methods. Transport Reviews. 39(4): 1-24

Xu C, Ji J, Liu P (2018) The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. Transportation Research Part C 95:47-60