A Comparative Analysis of Machine Learning Techniques for Bike-Sharing Demand Forecasting in Seoul

Subeen Leem[1], Jihoon Moon[1,2], and Seungmin Rho[3,*]
[1]*Department of Medical Science, Soonchunhyang University, Asan, South Korea*
[2]*Department of AI and Big Data, Soonchunhyang University, Asan, South Korea*
[3]*Department of Industrial Security, Chung-Ang University, Seoul, South Korea*
*Contact: smrho@cau.ac.kr, phone +82-2-820-5630

Abstract—**In this paper, we developed five machine learning models for bike-sharing demand forecasting in Seoul to compare their prediction performance. We first collected the Seoul Bike Sharing Demand Prediction dataset, a publicly accessible dataset, from the University of California, Irvine (UCI) Machine Learning Repository website and divided the dataset into ten months and two months for training and test sets, respectively. We employed five machine learning techniques, including multiple linear regression, decision tree, k-nearest neighbor, support vector machine (SVM), and random forest (RF), to conduct a 1-hour interval bike-sharing demand forecasting in Seoul. We verified their prediction performance regarding the root mean square error, R-squared, and mean absolute error. The SVM model derived an excellent prediction performance compared to other machine learning models. We also exhibited variable importance and a partial dependence plot to identify the key influential factors of the RF model and confirmed that the hour and temperature information are significant factors in model construction.**

I. Introduction

Bike-sharing systems have been rapidly spreading worldwide with the remarkable advances in intelligent transportation systems and information technology since the first decade of the 2000s. These systems simplify public bike rental. Mobile applications based on the global positioning system allow people to gather information about nearby bike stations to rent bikes. To date, many countries have implemented bike-sharing procedures [1]. They have become essential elements of transport systems in major cities due to improving several factors, such as health problems, traffic jams, and environmental conditions. After using a bike, people can lock it at any docking station in the city. To expand the public use of bicycles, the managers of these systems allocate trucks to collect parked bikes at various stations and gradually move the bicycles to their original stations.

Known in English as Seoul Bike, Ddareungi is a public rental bike-sharing system launched in 2015 in Seoul, South Korea [2]. Ddareungi was started to overcome such problems as high oil prices, traffic jams, and environmental pollution and create a healthy environment for Seoul citizens. These public bicycles are available for those 13 years or older and are manufactured using durable and lightweight materials. Ddareungi provides users with driving stability and convenience. Users can rent and store bicycles at any station. The bike stations are installed in areas with heavy traffic, such as subway entrances and exits, bus stops, residential complexes, government offices, schools, and banks. In addition, with the help of internet-capable devices or mobile phones, people can determine the stops they can rent and check their travel history and physical activity.

Due to these innovative technologies and convenience, the utilization of bike rentals is increasing daily in Seoul. Thus, the demand for bicycle rentals must be managed, and continuous and convenient user service must be maintained. One key to these services is to predict the number of rental bikes needed to keep the bike-sharing system working continuously with the growing number of users. This paper proposes several bike-sharing demand forecasting models based on machine learning techniques. First, we divide the Seoul bike-sharing demand dataset into training and test sets. We use multiple linear regression (MLR), decision tree (DT), k-nearest neighbor (KNN), support vector machine (SVM), and random forest (RF) to conduct a 1-hour interval bike-sharing demand forecasting in Seoul and compare their prediction performance.

The rest of this article is arranged as follows. Section 2 offers the general procedure of the proposed method. Section 3 explains the experimental outcomes to verify the validity of the proposed scheme. Finally, Section 4 represents the conclusion and future research directions.

II. Forecasting Model Construction

The University of California, Irvine (UCI) Machine Learning Repository website [3] has a publicly available dataset we collected. This dataset provides the number of public bicycles rented per hour in Seoul, Republic of Korea, and related weather data and holiday information. The dataset period is 12 months (365 days) from Dec. 2017 to Nov. 2018, and the attribute information of the dataset is presented in Table 1.

The seasons, holidays, and functioning days in the dataset are composed of character variables. We converted these variables into numeric variables: seasons (1: spring, 2: summer, 3: autumn, and 4: winter), holidays (1: workdays and 0: national holidays), and functioning days (1: functional hours, 0: nonfunctional hours). The number of rental bicycles rented per hour is a dependent variable from the dataset.

We split the dataset into a training set and testing set in a 83:17 ratio. The training set period is ten months (Dec. 2017 to

TABLE I
INFORMATION ON THE SEOUL BIKE-SHARING DEMAND PREDICTION DATASET

No.	Name	Description	Data Type
1	Date	Year-Month-Day	Timestamp
2	Rented Bike Count	Count of Bikes Rented Per Hour	Dependent
3	Hour	Hour of the Day	Independent (Timestamp)
4	Temperature	Celsius	Independent (Weather)
5	Humidity	%	Independent (Weather)
6	Wind Speed	m/s	Independent (Weather)
7	Visibility	10 m	Independent (Weather)
8	Dew Point Temperature	Celsius	Independent (Weather)
9	Solar Radiation	MJ/m2	Independent (Weather)
10	Rainfall	mm	Independent (Weather)
11	Snowfall	cm	Independent (Weather)
12	Seasons	Spring, Summer, Autumn, Winter	Independent (Timestamp)
13	Holiday	Holiday, No Holiday (Workday)	Independent (Timestamp)
14	Functioning Day	Yes (Functional Hours), No (Nonfunctional Hours)	Independent (Timestamp)

Sep. 2018). The testing set period is two months, from Oct. 2018 to Nov. 2018. The training set was used to construct the MLR, DT, KNN, SVM, and RF models on the R environment (RStudio v. 1.3.1073 with R v. 4.1.2), a programming language popular among data scientists.

We built the DT and KNN and SVM models using the *rpart* and *caret* packages, respectively. The optimum hyperparameter values of the KNN and SVM models were tuned using the *tuneLength* function. For the RF model construction, we used the *ranger* [4], a creative software package for implementing RF, and set two RF hyperparameter values, namely, *nTree* and *mTry* to 128 and 4, respectively [5].

III. EXPERIMENTAL RESULTS

The prediction performance of the regression models is considered using a variety of criteria; namely, regarding the root mean square error (RMSE), R-squared (R^2), and mean absolute error (MAE). Equations (1) to (3) are used to determine the RMSE, R^2, and MAE, respectively.

$$\text{RMSE} = \sqrt{(\sum(A_t - F_t)^2 / n)}, \quad (1)$$

$$R^2 = 1 - (\sum(A_t - F_t)^2 / \sum(A_t - \overline{A})^2 \quad (2)$$

$$\text{MAE} = (\sum|A_t - F_t / n), \quad (3)$$

where, n denotes the number of prediction time points and t denotes the prediction time point. Moreover, A_t and F_t signify the actual and forecasted values at time t, respectively, whereas $\overline{A}$ denotes the average of all actual values.

Table 2 lists the performance comparison of the machine learning models on the testing set. We confirmed that the SVM model outperformed the other machine learning models in terms of RMSE and MAE, whereas the RF model exhibited the second excellent performance. In addition, the RF model derived a better prediction performance than the other machine learning models in terms of R^2, whereas the SVM model exhibited the second excellent performance.

In Fig. 1, we can verify that if the temperature is below 17 degrees Celsius, the bike-sharing demand is low, and if the

TABLE II
PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	RMSE	R^2	MAE
MLR	478.23	0.529	341.10
DT	485.29	0.428	339.61
KNN	459.04	0.526	327.83
SVM	408.86	0.615	278.64
RF	415.97	0.750	297.93

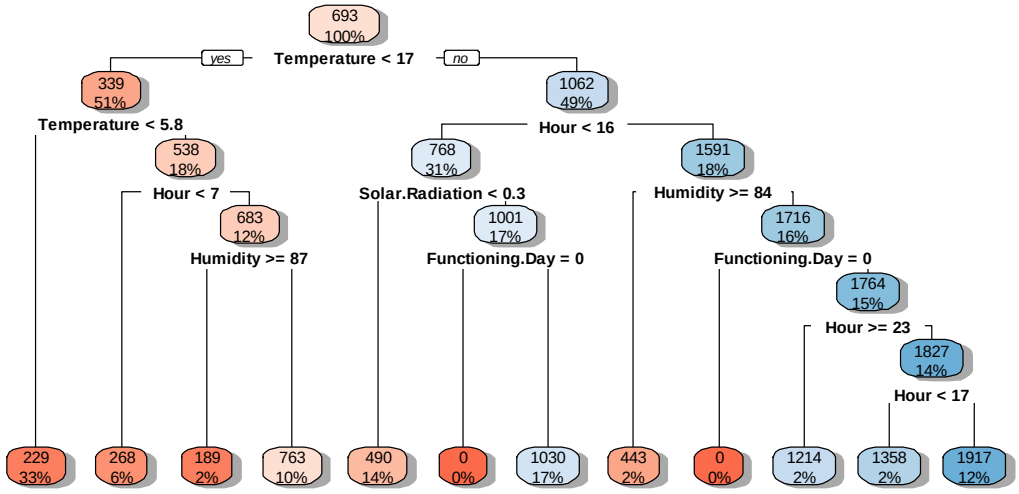

Fig. 1 Decision tree plot for bike-sharing demand forecasting

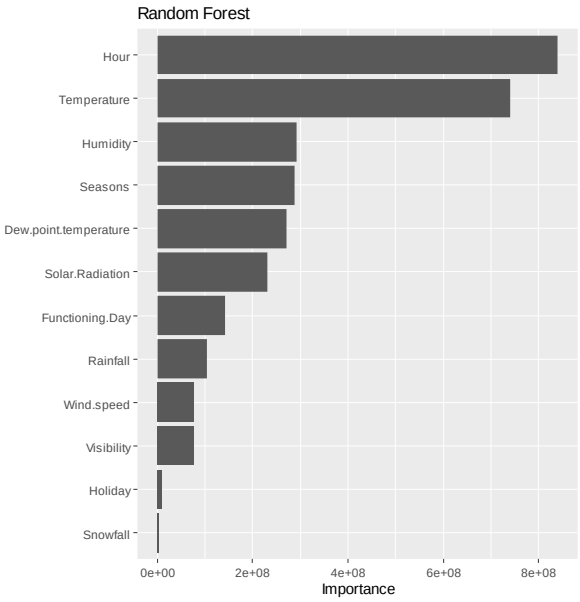

Fig. 2 Variable importance for the random forest model

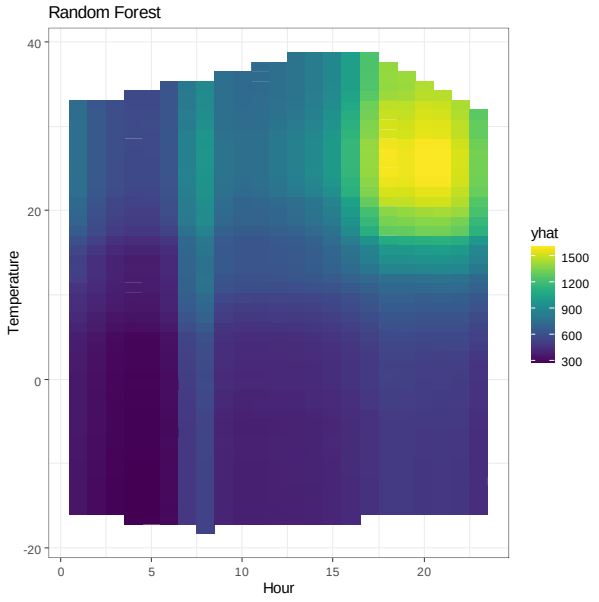

Fig. 3 Partial dependence plot between the temperature and hour variables

temperature is above 17 degrees Celsius, the bike-sharing demand is high on the first branch. In addition, we can verify that the lower bike-sharing demand with temperatures below 6.8 degrees Celsius or before 7 a.m. on the left branch. On the right branch, there is a lot of bike-sharing demand from 6 p.m. to 11 p.m.

In the training set, we examined variable importance to establish the most influential factors, as shown in Fig. 2. We also considered a partial dependence plot (PDP) to represent the change in projected values based on the difference in their characteristics, as depicted in Fig. 3. In Fig. 2, *Hour* was the most influential variable, whereas *Temperature* was the most influential weather condition and the second most influential variable for model construction.

We employed multipredictor PDPs to express the change in prediction values when the significant factors changed. In Fig. 3, we used *Hour* and *Temperature* as the most important timestamp and weather elements to display the PDPs. The prediction value is indicated by "*yhat*" in this figure. We confirmed that the bike-sharing demand in Seoul is high during commuting hours and higher temperature results in higher bike-sharing demand.

IV. Conclusions

This paper developed five machine learning models for bike-sharing demand prediction in Seoul, South Korea. We collected the Seoul bike-sharing demand dataset from the UCI Machine Learning Repository website. We converted character variables, such as seasons, holidays, and functioning days, into numeric variables for model construction. We split the datasets into training (10 months) and testing (2 months) sets and adopted machine learning techniques, such as MLR, DT, KNN, SVM, and RF.

We confirmed that the SVM model outperformed the other machine learning models in terms of RMSE and MAE. The RF model also derived a better prediction performance than the other machine learning models in terms of R2. Variable importance and PDP were analyzed to identify the hidden links between the external components. Hour and Temperature were ranked as the most significant variables in predicting rental bicycle demand at each hour. We identified a substantial number of bicycle rentals during commuting time.

The proposed method forecasted bicycle rental demand for the entire Seoul area to aid in developing urban policies that provide various options for better health, a cleaner environment, and climate action. Future research will project rental bicycle demand by region while accounting for seasonal fluctuations. Forecasts based on regional rental bicycle demand will be significantly more helpful in constantly delivering public rental bicycles, which we believe will play an essential role in the population's physical, social, and mental well-being.

Acknowledgment

This paper was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0008703, The Competency Development Program for Industry Specialist).

References

[1] J. C. García-Palomares, J. Gutiérrez, and M. Latorre, "Optimizing the location of stations in bike-sharing programs: A GIS approach," *Applied Geography*, vol. 35, pp. 235–246, 2012.

[2] E. J. Kim, J. Kim, and H. Kim, "Does Environmental Walkability Matter? The Role of Walkable Environment in Active Commuting," *International Journal of Environmental Research and Public Health*, vol. 17, no. 4, p. 1261, 2020.

[3] V. E. Sathishkumar, J. Park, and Y. Cho, "Using data mining techniques for bike sharing demand prediction in metropolitan city," *Computer Communications*, vol 153, pp. 353–366, 2020.

[4] M. N. Wright and A. Ziegler, "ranger: A fast implementation of random forests for high dimensional data in C++ and R," *Journal of Statistical Software*, vol. 77, pp. 1–17, 2017.

[5] J. Moon, S. Park, S. Rho, and E. Hwang, "Robust building energy consumption forecasting using an online learning approach with R ranger," *Journal of Building Engineering*, vol. 47, p. 103851, 2022.