ISSN: 2508-7894 © 2022 KODISA & KAIA.
KJAI website: http://www.kjai.or.kr
doi: http://dx.doi.org/10.24225/kjai. 2022.10.1.9

Prediction of the number of public bicycle rental in Seoul using Boosted Decision Tree Regression Algorithm

Hyun-Jun KIM[1], Hyun-Ki KIM[2]

Received: April 25, 2022. Revised: : April 26, 2022. Accepted: May 16, 2022.

Abstract

The demand for public bicycles operated by the Seoul Metropolitan Government is increasing every year. The size of the Seoul public bicycle project, which first started with about 5,600 units, increased to 3,7500 units as of September 2021, and the number of members is also increasing every year. However, as the size of the project grows, excessive budget spending and deficit problems are emerging for public bicycle projects, and new bicycles, rental office costs, and bicycle maintenance costs are blamed for the deficit. In this paper, the Azure Machine Learning Studio program and the Boosted Decision Tree Regression technique are used to predict the number of public bicycle rental over environmental factors and time. Predicted results it was confirmed that the demand for public bicycles was high in the season except for winter, and the demand for public bicycles was the highest at 6 p.m. In addition, in this paper compare four additional regression algorithms in addition to the Boosted Decision Tree Regression algorithm to measure algorithm performance. The results showed high accuracy in the order of the First Boosted Decision Tree Regression Algorithm (0.878802), second Decision Forest Regression (0.838232), third Poison Regression (0.62699), and fourth Linear Regression (0.618773). Based on these predictions, it is expected that more public bicycles will be placed at rental stations near public transportation to meet the growing demand for commuting hours and that more bicycles will be placed in rental stations in summer than winter and the life of bicycles can be extended in winter.

Keywords : Azure Machine Learning Studio, Boosted Decision Tree Regression, Shared bike, Machine Learning

JEL Classification Code : Basic Technology, Technical Application, Artificial Intelligence Convergence

1. Introduction

The demand for public bicycles operated by the Seoul Metropolitan Government is increasing every year. Seoul's public bicycle project, which began in 2015 with 450 rental stations and 5,600 public bicycles, grew up to 2,523 rental stations and about 37,500 public bicycles as of September 2021 and is now an inconvenient public service for Seoul citizens. However, as the size of Seoul's public bicycle business grows, there are also negative economic aspects. The Seoul Metropolitan Government spent about 34.4 billion won in 2020 on the cost of injecting new bicycle rental stations and maintaining existing bicycles. It has been observed that the budget was invested more than five times compared to the 6.5 billion won budget in 2016, and the number of bicycles operated has increased while existing bicycles and rental shops have aged and maintained, and the size of budget input is also increasing. As the demand for public bicycles in Seoul increases and budget input increases, it is recording a deficit every year, and budget efficiency should be increased for sustainable projects. To compensate for this problem, this paper predicts the number of public bicycles rented according to various environmental factors such as season, time, temperature, etc., and efficiently arranges public bicycles based on the prediction results to improve the usability of existing bicycles. It is expected that if public bicycles are efficiently deployed and usability is increased, they will be able to reduce the budget for new public bicycles and respond positively to the increasing demand for public

1 First Author, Student, Medical IT, Eulji University, Korea. Email: hyunjune8955@g.eulji.ac.kr
2 Corresponding Author, CEO, Shinnam information and Communications, Korea. Email: r48019@naver.com

bicycles. In addition, it is intended to study which regression method is more suitable for predicting the number of public bicycles by comparing the existing fact data and prediction data using various regression methods in addition to the Boosted Decision Tree Regression method.

2. Literature Review

2.1. Azure Machine Learning Studio

Azure Machine Learning Studio is one of Microsoft's cloud computing platforms that began service in 2010. Following PaaS in 2011, IaaS service was launched in 2013. Although there are many open source libraries and tools for machine learning, Azure Machine Learning Studio can solve problems such as lack of flexibility in computing resources for learning, difficulty configuring GPU-based environments for machine learning, difficulty installing and setting tools for learning, and difficulty managing experimental records and versions.

2.2. Decision Tree

Decision tree learning methods use decision trees as predictive models that link observation and target values for an item. This is one of the predictive modeling methods used in statistics, data mining, and machine learning. The decision tree learning method aims to generate a model that predicts the value of the target variable based on several input variables. Therefore, it is one of the most useful techniques in supervised classification learning. Figure 1 is a simple example of a decision tree. Decision tree learning is learning by continuing the True/False questions. For example, assuming that animals such as eagles, penguins, dolphins, and bears are distinguished, as in Figure 1, they can be divided into hawks, penguins/dolphins, and bears, and eagles and penguins can be divided by asking, "Can you fly?" In this way, a model that classifies data according to a specific criterion is called a decision tree model. The advantages of the decision tree are as follows. First, it is easy to interpret the results, so intuitive interpretation is possible. Second, it is a nonparametric model. It is free to assume that the statistical model requires. It is possible to explore linear/nonlinear relationships by considering the interaction between the third variables.

2.3 Machine Learning

Machine learning or machine learning is a field of artificial intelligence in computer science that has evolved from the study of pattern recognition and computer learning theory. Machine learning can be said to be a technology that studies and builds systems that learn based on empirical data, perform predictions, and improve their own performance. Rather than performing strictly defined static program instructions, the algorithms of machine learning take the approach of building specific models to derive predictions or decisions based on input data. Machine learning algorithms can be largely divided into supervised learning, unsupervised learning, and reinforced learning according to the form of inputting information and data into the learning system. Map learning is the process of learning a function that maps an input to a corresponding known output. In other words, it can be said to be a process of inferring a function from training data. Unsupervised learning refers to a method of learning by building a model only with input without output. In general, most of the data mining techniques correspond to this. Finally, reinforcement learning refers to a method in which a learner selects an action to affect the environment with an action and obtains a compensation value with feedback on it and uses it as a guide to a learning algorithm. Supervised learning, unsupervised learning, and reinforcement learning are being studied in various fields of artificial intelligence.

2.4. Prediction of the number of public bicycle rental using deep learning

The Seoul Metropolitan Government's public bicycle business is growing every year. In particular, public bicycles have a positive effect in many areas, such as preventing environmental pollution and improving citizens' health, and for this reason, many analyses and studies are being conducted on the prediction and characteristics of the number of public bicycles. Ajou University Transportation Research Institute developed a deep learning model to predict the rental volume of public bicycles in 2020, built an ARIMA model and an LSTM-based deep learning model, and compared and evaluated prediction errors using MSE and MAE evaluation indicators. As a result, the MSE of the deep learning model decreased by 66% and the MAE was 52%, resulting in a small error in the deep learning model, and through this model, it was determined that the number of loans could be reduced through the application of the deep learning model in the field of predicting public bicycle rental. In addition, research on predicting demand for public bicycles using artificial intelligence is being conducted.

3. Dataset Description

In order to predict the number of public bicycle rentals

in Seoul, this study downloaded and used the Seoul Bike Data.csv file with information on the environment and the number of rental public bicycles in Seoul from 2017 to 2020. Table 1 is the attribute item of the dataset.

Table1: Data set column list

Column	Content
Data	year – month - day
Rented Bike count	Number of bicycles rented per hour
Hour	Hours of the day
Temperature	Temperature (C)
Humidity	Humidity (%)
Windspeed	Wind speed(m/s)
Visivility	Visibility
Dew point temperature	Dew point temperature
Solar radiation	Amount of solar radiation
Rainfall	Precipitation (mm)
Snowfall	Suitable amount (cm)
Seasons	Seasons(Spring, Summer, Fall, Winter)
Holiday	Holiday(Holiday / Not Holiday)
Functional Data	Operation date

The dataset has 8,760 columns and 14 attributes. It includes environmental information and rental information related to public bicycles in Seoul from 2017 to March 1, 2020.

4. Experimental process

4.1. Dataset Reprocessing

To predict the number of public bicycle rentals in Seoul, the SeoulBikeData.csv file was imported. After that, in order to increase the effect of learning the data, missing values must be removed. To remove the missing value, the row of missing value was removed, and to remove the missing value, the 'Clean Missing Data' option was added to remove the missing value, and data with missing value was excluded by selecting remove the entry. To confirm that there is no problem with Dataset by eliminating missing values and to predict the number of bicycle rentals by an environmental factor, six public bicycles were predicted by selecting Hour, Rented Bike Count, Data, Temperature, Humidity, and Seasons properties. Finally, it was divided into training data and test data through data Split at a ratio of 8:2. In this research was conducted using a data set that completed preprocessing.

4.2 Model Selection and Train

After completing the data split through the process up to 3.2, the learning model must be selected with the training data and taught. In this paper, since the goal was to predict the number of rental bicycles with several environmental factors, the Boosted Decision Tree Regression method was selected and the learning was conducted. The Boosted Decision Tree Regression model is chosen because this study predicts the Rented Bike Count based on several environmental factor columns, and the Boosted Decision Tree Regression model is useful in predicting the value of the target variable based on the data of several input variables. For this reason, it was determined that the most effective algorithm in this paper is Boosted Decision Tree Regression. Data is predicted by comparing selected columns such as Temperature, Seasons, Hour, and Humidity with each other.

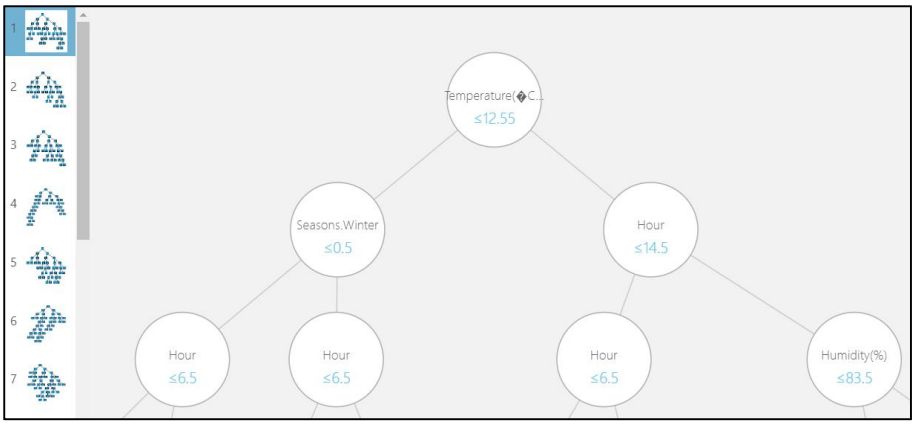

Figure 1: Check the trained prediction model

Figure 1 is a tree representation of a prediction model that connects the columns and the target value columns (Rented Bike Count). In the case of the learning method, learning is performed by comparing attributes from the root node. In other words, the number of shared bicycle rental is predicted by comparing the environment with attributes such as temperature, season, and time. Looking at Figure 1, it is possible to check the tree in which each attribute is compared with the temperature.

4.3 Score Model

If a learning model is created through the Train Model, the test data and the predicted value can be checked through the Score Model. As for the predicted result value, the result value for predicting the number of bicycle rentals can be checked with Scored Labels, and the correlation distribution can be confirmed by comparing Rented Bike Count and Scored Labels. At this time, if the correlation distribution is well connected in a straight line, it can be seen as a well-predicted result. Rented Bike Count refers to the actual number of public bicycle rentals, and Scored Labels is a value predicted through the training model. The correlation distribution of these two values in a straight line means that the actual value and the predicted value are almost similar, and the prediction can be seen as well.

In addition, to identify the number of rentals by time zone, first, compare Hour and Scored Labels, second compare Temperature and Scored Labels, and finally, compare the predicted results with the existing results. The output values are shown in Figure 2,3.

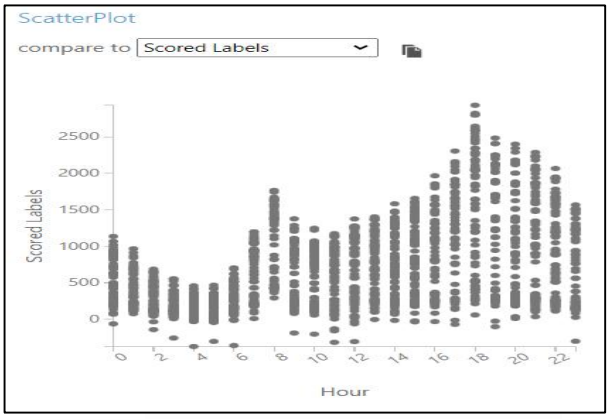

Figure 2: Comparison of Hour and Scored Labels

In addition, Hour and Scored Labels were compared to identify the number of loans by time zone. Figure 2 is a comparison of the predicted number of bicycle rental by time zone. According to the prediction results, it was predicted that there would be the highest demand for public bicycles during rush hour at 18:00, and overall, it was expected that there would be a high demand for public bicycles during 8 a.m. and afternoon hours. On the other hand, demand for public bicycles was relatively low during the time except for work hours and work hours.

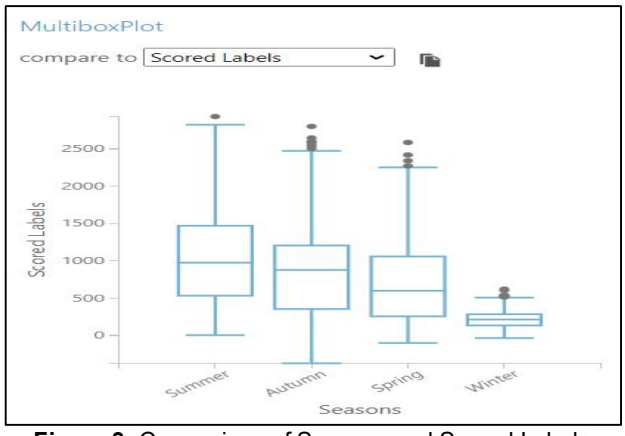

Figure 3: Comparison of Seasons and Scored Labels

Figure 3 predicts the number of public bicycle rental by season. More than 3,000 units were expected on average in summer, and 2,500 to 3,000 units were expected in spring and autumn. Finally, 500 to 1,000 units were expected in winter. In summary, the demand for rental of public bicycles is relatively low on cold days from minus 15 degrees to 5 degrees Celsius, and the demand for rental of public bicycles is high on warm and hot days from 20 degrees Celsius to 30 degrees Celsius. Demand for public bicycles decreases in the cold winter, but an increase can be expected in spring, summer, and autumn.

4.4 Evaluate Model

Finally, it is necessary to confirm the accuracy and results of the prediction model, and the accuracy was confirmed by performing the evaluation model process as the next step of the score model. For the evaluation model, we evaluate how accurately the Boosted Decision Tree Regression algorithm predicts the demand for public bicycle rental numbers. At this time, the average absolute error, root mean square error, relative absolute error, relative square error, and determination effect can be confirmed through the model evaluation process. The average absolute error refers to an average of an absolute value for the difference between the processed or transmitted image and the original image. The mean square error refers to the difference between the measured value or the predicted value and the value that is thought to be true. That is, the average of the squares for the errors is called the mean square error. Relative absolute error is the

ratio of how much the measured value of an object differs from the actual value. Finally, the coefficient of determination refers to the ratio of the variance of the dependent variable described as an independent variable. In short, it is a numerical representation of how accurately the object can be described with a statistical model. The results obtained through model evaluation are as follows. The mean absolute error is 150.691208, the root mean square error 226.933151, the relative absolute error 0.286483, the relative square error 0.121198, and finally the coefficient of determination 0.878802. The predictive accuracy of the model should be determined. The determination coefficient represents the accuracy of the model, and the result is 0.878802, so the model made to predict the number of public bicycle rental in this paper shows an accuracy of 87.8%

.

4.5 Train Model Accuracy Comparison

In this paper, the number of rentals for public bicycles was predicted according to the columns through the Boosted Decision Tree Regression model. The purpose of this study is to compare the accuracy between models through various regression algorithms other than Boosted Decision Tree Regression. I compare the performance of Boosted Decision Tree Regression with other algorithms. The process is as follows. First of all, after importing the dataset, data preprocessing is carried out through 'Clean Missing Data' and 'select Columns in Dataset'. After completing preprocessing, it is classified into learning data and test data through data split. Then, I Trained the Boosted Decision Tree Regression Algorithm and other regression algorithms to compare their performance. Finally, the performance of the trained algorithm was compared through the 'Evaluate Model'. As a result, the algorithms that show high accuracy in predicting public bicycles are listed in order as follows. Figure4 shows the accuracy of the algorithm's demand prediction for the number of public bicycles.

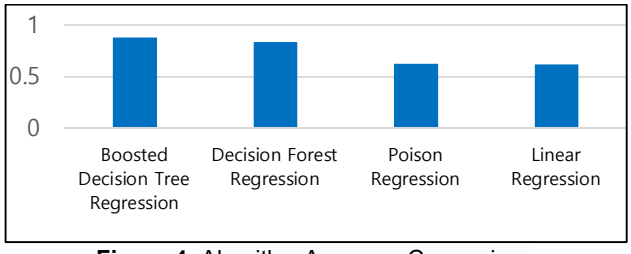

Figure 4: Algorithm Accuracy Comparison

First Boosted Decision Tree Regression Algorithm (0.878802), second Decision Forest Regression (0.838232), third Poison Regression (0.62699), and fourth Linear Regression (0.618773). It can be seen that the Boosted Decision Tree Regression algorithm is about 87%, which shows higher accuracy than other algorithms. The reason why the Boosted Decision Tree Regression algorithm shows the highest accuracy is as follows. Boosted Decision Tree Regression learns by fitting the residual of the trees that preceded it. Thus, boosting in a decision tree ensemble tends to improve accuracy with some small risk of less coverage. In conclusion, the Boosted Decision Tree Regression algorithm shows the best performance in predicting the number of bicycle rental using various environmental variables.

5. Conclusions

This paper aims to predict the number of bicycle rentals according to factors such as time and environment, for a stable supply in line with the rapidly increasing demand for public bicycles in Seoul. The reason for this prediction is that public bicycles provided by the Seoul Metropolitan Government record deficits every year, but as demand continues to increase recently, the purpose is to efficiently deploy bicycles through data analysis and use the budget for public bicycles.

In this study, the UCI Machine Learning Repository predicted the number of bicycle rentals according to various environmental factors based on the 2020 Seoul Public Bicycle Dataset and used SeoulBikeData.csv Dataset for the project to process missing values, data split, and model training using Boosted Decision Tree Regression. The demand for public bicycles was expected to be the highest at 18:00 during the rush hour, and overall, the demand for public bicycles was expected to be high during the 8 a.m. and afternoon hours. On the other hand, the demand for public bicycles was relatively low during the hours excluding office hours and office hours. The prediction of the number of seasonal public bicycle rental cases is as follows. More than 3,000 units were expected on average in summer, and 2,500 to 3,000 units were expected in spring and autumn. Finally, 500 to 1,000 units were expected in winter. In summary, the demand for renting public bicycles is relatively low on cold days from minus 15 degrees to 5 degrees Celsius, and the demand for renting public bicycles is high on warm and hot days from 20 degrees Celsius to 30 degrees Celsius. In summary, it was found that it would be the most in the afternoon rush hour, such as 8 a.m., and in addition, the expected temperature, season, and bicycle rental numbers were compared in the order of summer-fall-spring-winter. There is no big difference in demand in summer, fall, and spring, but it has been confirmed that demand decreases sharply in winter.

Table2: Maximum minimum number of rentals

	Time	Seasons	Temperature	Rental range
Max rental	6p.m	Summer	24~28 degrees	More than 3000 units
Minimum rental	4a.m	Winter	-15 degrees	Under 100 units

Table 2 shows information about the maximum rental and the minimum rental. The time when the number of loans was the highest was 6 p.m. in summer, with more than 3,000 loans. On the other hand, the lowest number of loans was 4 a.m. in winter, recording less than 100 loans. Based on these predictions, it is expected that more public bicycles will be placed at rental stations near public transportation to meet the growing demand for commuting hours and that more bicycles will be placed in rental stations in summer than winter and the life of bicycles can be extended in winter. As a result of performing the evaluation to determine the accuracy of the predicted value obtained in this way, a significant coefficient of determination of 0.878802 could be obtained. Accuracy comparison results according to the algorithm showed high accuracy in the order of first boost decision tree regression algorithm (0.878802), second decision forest regression (0.838232), third poison regression (0.62699), and fourth linear regression (0.618773). It can be said that the Boosted Decision Tree Regression algorithm is the most appropriate to predict the number of public bicycle based on the accuracy results.

References

An, S. H., Yeo, S. H., & Kang, M. S. (2021). A Study on a car Insurance purchase Prediction Using Two-Class Logistic Regression and Two-Class Boosted Decision Tree. *Korean Journal of Artificial Intelligence*, *9*(1), 9-14.

Cho, K. M., Lee, S. S., & Nam, D. H. (2020). Forecasting of Rental Demand for public bicycles Using a Deep Learning Model. *Korea Institute of Intelligent Transport System, 19*(3), 28-37.

Kang, M. S., Kang, H. J., Yoo, K. B., Ihm, C. H., & Choi, E. S. (2018). *Getting started with Machine Learning using Azure Machine Learning studio*. Seoul, Korea: Hanti media.

Kang, M. S., & Choi, E. S. (2021). *Machine Learning: Concepts, Tools and Data Visualization.* Singapore, Singapore: World Scientific Publishing Company.

Kim, H. G., & Kim, S. I. (2019). A Study on the Direction of Public Bicycle Development in Korea. Journal of Digital Convergence. *Journal of Digital Convergence*, *16*(8), 263-267.

Kim, K. O. (2018). *A Study on the Characteristics of Using Shared Bicycles for the Operation of Shared Bicycle System Considering Bicycle Imbalances*. Seoul, Korea: Seoul Research Paper Contest, Seoul researcher, Seoul, Korea.

Kim, K. S., & Jeong, Y. H. (2021). A Study on Crime Prediction to Reduce Crime Rate Based on Artificial Intelligence. *Korean Journal of Artificial Intelligence*, *9*(1), 15-20.

Mun, J. H., & Jung, S. W. (2021). A customer credit Prediction Researched to Improve Credit Stability based on Artificial Intelligence. *Korean Journal of Artificial Intelligence*, *9*(1), 21-27.

Sports Korea News (2021). One out of three Seoul citizens rode "Ttareungi". *Sports Korea News*, 2 August, Section 1. Seoul, Korea.