ISSN: 2508-7894 © 2020 KODISA & KAIA.
KJAI website: http://www.kjai.or.kr
doi: http://dx.doi.org/10.24225/ kjai.2022.10.1.27

A Study on Predicting the demand for Public Shared Bikes using linear Regression*

Dong Hun HAN[1], Sang Woo JUNG[2]

Received: May 11, 2022. Revised: May 16, 2022. Accepted: June 03, 2022.

Abstract

As the need for eco-friendly transportation increases due to the deepening climate crisis, many local governments in Korea are introducing shared bicycles. Due to anxiety about public transportation after COVID-19, bicycles have firmly established themselves as the axis of daily transportation. The use of shared bicycles is spread, and the demand for bicycles is increasing by rental offices, but there are operational and management difficulties because the demand is managed under a limited budget. And unfortunately, user behavior results in a spatial imbalance of the bike inventory over time. So, in order to easily operate the maintenance of shared bicycles in Seoul, bicycles should be prepared in large quantities at a time of high demand and withdrawn at a low time. Therefore, in this study, by using machine learning, the linear regression algorithm and MS Azure ML are used to predict and analyze when demand is high. As a result of the analysis, the demand for bicycles in 2018 is on the rise compared to 2017, and the demand is lower in winter than in spring, summer, and fall. It can be judged that this linear regression-based prediction can reduce maintenance and management costs in a shared society and increase user convenience. In a further study, we will focus on shared bike routes by using GPS tracking systems. Through the data found, the route used by most people will be analyzed to derive the optimal route when installing a bicycle-only road.

keywords : Shared bikes, demand forecasts, linear regression, machine learning, AI

Major classifications : Artificial Intelligence, etc

1. Introduction

Bike-sharing systems allow people to rent a bicycle at one of many automatic rental stations scattered around the city, use it for a short journey, and return it to any station in the city (Raviv, 2013). First, as the need for eco-friendly transportation increases due to the deepening climate crisis, many local governments in Korea are introducing shared bicycles. Due to anxiety about public transportation after COVID-19, bicycles have firmly established themselves as the axis of daily transportation. Among them, the Seoul Metropolitan Government's shared bicycle service "Ttareungi" is set to introduce 40,000 units within the year 2020, and citizens can easily access rental facilities by expanding them around multi-dense areas such as subway stations and bus stops. Statistics on the number of users, which increased by 36% compared to the previous year, indicate explosive demand for shared bicycles. According to the Seoul Metropolitan Government, 377,000 people

* This paper was supported by the research grant of the KODISA scholarship foundation in 2022

[1] First Author, Student Researcher, MIIC, Eulji Univ. Korea, Email: d555v@naver.com

[2] Corresponding Author. General Manager, All for Land, Korea, Email: cki723@all4land.com

signed up as new members of "Ttareungyi" in the first half of this year, bringing the cumulative number of members to 3,109,000 as of the end of June. The total number of rentals in the first half of the year was 13,684,000, which was used by an average of 75,605 people a day. This is an increase of 30.3 percent from the same period last year.

As the use of shared bicycles spreads, the demand for bicycles is increasing by the rental office, but there are operational and management difficulties because the demand is managed under a limited budget. And unfortunately, user behavior results in a spatial imbalance of the bike inventory over time (Schuijbroek, 2017). Although it is currently trying to address fluctuations in demand by rental stations through bicycle relocation, accurate prediction of uncertain future user demand is a more fundamental solution (Lim, 2019). Furthermore, installing a station takes time and is costly, with the removal of asphalt or pavers, undergrounding of the structure and wires, hook-up to a nearby electrical source, and replacement of building materials, the public bike system has limited this expense with its “technical platform, ” which is the bike-sharing station's base and houses the wires for its bike dock and pay station (DeMaio, 2009). So, in order to facilitate operations efficiently, it is necessary to prepare in large quantities when demand is high and then retrieve them at a lower time.

Nowadays, machine learning (ML) is used in every area of computational work where algorithms are designed and performance is increased (Abdulqader, 2020). ML, a branch of Artificial Intelligence, relates the problem of learning from data samples to the general concept of inference and every learning process consists of two phases: (i) estimation of unknown dependencies in a system from a given dataset and (ii) use of estimated dependencies to predict new outputs of the system (Kourou, 2015). Therefore, in this study, the second option was chosen with simple linear regression algorithms and MS Azure ML is used to predict and analyze when demand is high.

2. Related Research

2.1. Machine Learning

Machine learning is an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment (El Naqa & Murphy, 2015). It consists of designing efficient and accurate prediction algorithms (Mohri M, 2018), and is closely related to the fields of pattern recognition, computational statistics, and artificial intelligence (Paluszek, 2016). The term "Machine Learning" was first used by Arthur Samuel, an IBM researcher in the field of artificial intelligence, in his paper “Studies in Machine Learning Using the Game of Checkers” (Kang & Choi, 2021). It has origins in the artificial intelligence movement of the 1950s and emphasizes practical objectives and applications, particularly prediction and optimization. Computers “learn” through machine learning by improving their performance at tasks through “experience” (Goodman & Lessler, 2019).

The purpose of machine learning is to learn from data and many studies have been done on how to make machines learn by themselves without being explicitly programmed (Mahesh, 2018). Machine learning can be separated into two types of study. The first is supervised learning. Supervised learning accounts for a lot of research activity in machine learning and many supervised learning techniques have found applications in the processing of multimedia content (Cunningham, 2008). This type of learning is analogous to humans learning from past experiences to gain new knowledge in order to improve our ability to perform real-world tasks, but since computers do not have “experiences”, machine learning learns from data, which is collected in the past and represents past experiences in some real-world applications (Liu, 2011). Unsupervised learning, however, the model does not provide correct results during the training. To be specific, unsupervised learning denotes how a network can study to signify some input designs in a method that reproduces the numerical arrangement of the total gathering of input designs or patterns (Dike, 2018).

2.2. Linear regression

Many learning algorithms that are widely used in practice rely on linear predictors, first and foremost because of the ability to learn them efficiently in many cases, and in addition, linear predictors are intuitive, are easy to interpret, and fit the data reasonably well in many natural learning problems (Shwartz, 2021). Regression analysis is a statistical approach to studying and modeling the relationship between variables, describing the status of data, estimating parameters, and fitting the model to predict and control (Mohammad, 2015). Linear regression is one of the simplest and most common machine-learning algorithms that the mathematical approach is used to perform predictive analysis (Maulud, 2020). Linear regression analysis is used to predict the value of a variable based on the value of another variable. For example, predict tomorrow's stock market price given current market conditions and other possible side information, Predict the age of a viewer watching a given video on YouTube, Predict the location in the 3d space of a robot arm end effector, and give control signals sent to its various motors can be predicted (Murphy, 2012).

Regression analysis is a statistical technique for investigating and modeling the relationship between variables (Montgomery, 2012). So, in statistics, linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models (Wikipedia, 2021).

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions (IBM, 2020). The concept of linear regression was first proposed by Sir Francis Galton in 1894. The linear regression analysis uses the mathematical equation, i.e., y = mx + c, that describes the line of best fit for the relationship between y (dependent variable) and x (independent variable), and the regression coefficient, i.e., r2 implies the degree of variability of y due to x (Kumari, 2018).

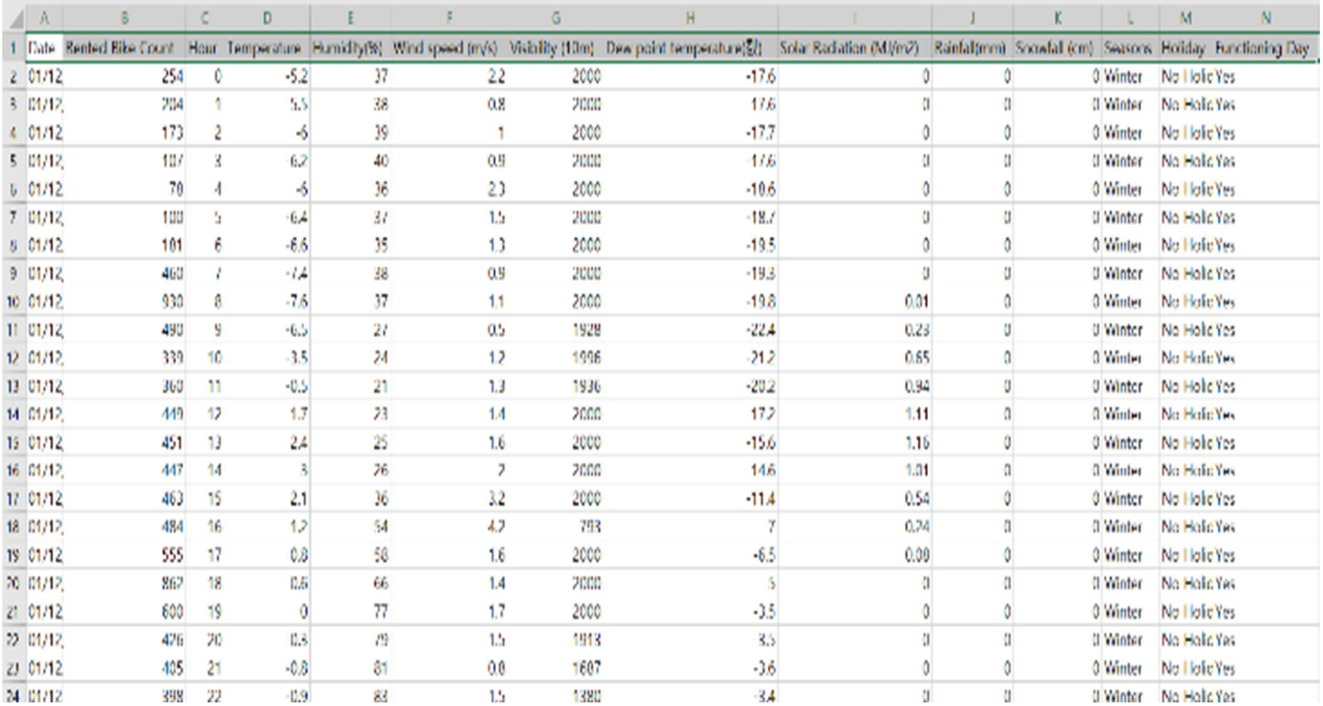
2.3. Azure Machine Learning

Azure Machine Learning is a cloud service that accelerates and simplifies the lifecycle of machine learning projects. One of the central themes of Azure Machine Learning is the ability to quickly create machine-learning experiments, evaluate them for accuracy, and then "fail fast" to shorten the cycles to produce a usable prediction model (Barnes, 2015). And by making it easier for developers to use the predictive models in end-to-end solutions, Azure Machine Learning enables actionable insights to be gleaned and operationalized easily (Barga, 2015). So, Machine learning experts, data scientists, and engineers can use this service to perform routine workflows, namely learning and deploying models and managing ML Ops. You can create models in Azure Machine Learning or use models built on open-source platforms such as Pytorch, TensorFlow, or scikit-learn (Microsoft, 2021).

3. Experiment

3.1. Data setup

As for the data, on-demand public bicycles in Seoul from 2017 to 2018 were used.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Date	Rented Bike Count	Hour	Temperature	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(℃)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
2	01/12	254	0	-5.2	37	2.2	2000	-17.6	0	0	0	Winter	No Holic	Yes
3	01/12	204	1	-5.5	38	0.8	2000	-17.6	0	0	0	Winter	No Holic	Yes
4	01/12	173	2	-6	39	1	2000	-17.7	0	0	0	Winter	No Holic	Yes
5	01/12	107	3	-6.2	40	0.9	2000	-17.6	0	0	0	Winter	No Holic	Yes
6	01/12	78	4	-6	36	2.3	2000	-18.6	0	0	0	Winter	No Holic	Yes
7	01/12	100	5	-6.4	37	1.5	2000	-18.7	0	0	0	Winter	No Holic	Yes
8	01/12	181	6	-6.6	35	1.3	2000	-19.5	0	0	0	Winter	No Holic	Yes
9	01/12	460	7	-7.4	38	0.9	2000	-19.3	0	0	0	Winter	No Holic	Yes
10	01/12	930	8	-7.6	37	1.1	2000	-19.8	0.01	0	0	Winter	No Holic	Yes
11	01/12	490	9	-6.5	27	0.5	1928	-22.4	0.23	0	0	Winter	No Holic	Yes
12	01/12	339	10	-3.5	24	1.2	1996	-21.2	0.65	0	0	Winter	No Holic	Yes
13	01/12	360	11	-0.5	21	1.3	1936	-20.2	0.94	0	0	Winter	No Holic	Yes
14	01/12	449	12	1.7	23	1.4	2000	-17.2	1.11	0	0	Winter	No Holic	Yes
15	01/12	451	13	2.4	25	1.6	2000	-15.6	1.16	0	0	Winter	No Holic	Yes
16	01/12	447	14	3	26	2	2000	-14.6	1.01	0	0	Winter	No Holic	Yes
17	01/12	463	15	2.1	36	3.2	2000	-11.4	0.54	0	0	Winter	No Holic	Yes
18	01/12	484	16	1.2	54	4.2	793	-7	0.24	0	0	Winter	No Holic	Yes
19	01/12	555	17	0.8	58	1.6	2000	-6.5	0.08	0	0	Winter	No Holic	Yes
20	01/12	862	18	0.6	66	1.4	2000	-5	0	0	0	Winter	No Holic	Yes
21	01/12	600	19	0	77	1.7	2000	-3.5	0	0	0	Winter	No Holic	Yes
22	01/12	426	20	-0.3	79	1.5	1913	-3.5	0	0	0	Winter	No Holic	Yes
23	01/12	405	21	-0.8	81	0.8	1687	-3.6	0	0	0	Winter	No Holic	Yes
24	01/12	398	22	-0.9	83	1.5	1380	-3.4	0	0	0	Winter	No Holic	Yes

Figure 1: Seoul bike demand data

The list of data is as follows.

The data set was pre-processed and confirmed that data exists only when 'Functioning Day' is 'Yes', and that the missing value is 0.

Table 1: Example of Data Set

Date
Rented Bike Count
Hour
Temperature
Humidity (%)
Wind speed (m/s)
Visibility (10m)
Dew point temperature
Solar Radiation (MJ/m2)
Rainfall (mm)
Snowfall (cm)
Seasons
Holiday
Functioning Day

3.2. Data Analysis

The order of data analysis is in the order of Select Columns in Data Set – Split Data – Linear Regression – Train Model – Score Model - Evaluate Model.

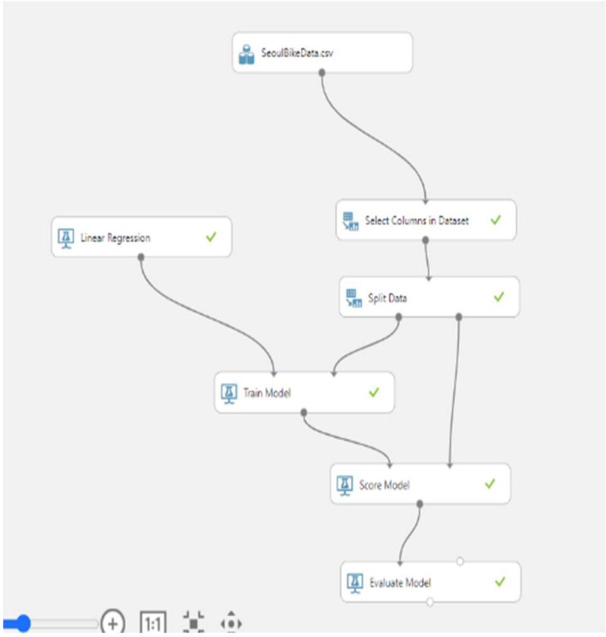


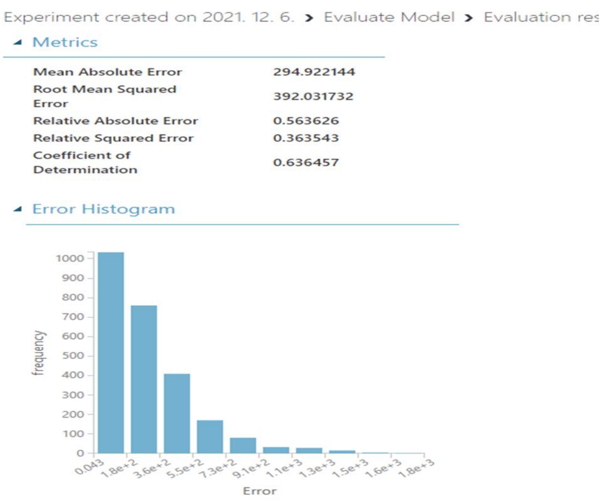
Figure 2: Azure ML flow

According to the Figure 2, After importing the Dataset, exclude 'Functioning Day', which is unnecessary data, from Select columns in dataset to Launch column selector.

Next, Split Data is divided into 7:3 ratios to separate training data and test data. For random sampling, Random seed enters 3123. In this analysis, Linear Regression is substituted as an appropriate analysis prediction method. For the Train Model, substitute the Rented Bike Count, which is the value to be predicted by Single columns. To determine the suitability of the model, run is performed after substituting the Evaluate Model.

4. Result

To check the analysis result, Run Score Model - Scored dataset - Visualize to verify.


Experiment created on 2021. 12. 6. > Evaluate Model > Evaluation res

Metrics

Mean Absolute Error	294.922144
Root Mean Squared Error	392.031732
Relative Absolute Error	0.563626
Relative Squared Error	0.363543
Coefficient of Determination	0.636457

Error Histogram

Figure 3: Score model

Figure 3 shows the visualization of the Evaluate Model to verify the reliability of the Score Model. The coefficient of determination is a measure of the degree to which the estimated linear model is suitable for a given data. It denoted R2, is the quotient of the explained variation (sum of squares due to regression) to the total variation (total sum of squares total SS (TSS)) in a model of simple or multiple linear regression (Di Bucchianico, 2008). If the value is greater than or equal to 0.6, this result is judged to be reliable, and the measured value of the result is 0.636457.

Title: Seaborn visualization pseudo-code

```
- Shared bicycle data visualization
subplots rows=2, cols=2
# 4 layers of windows
set_size_inches (20,8)
data = train,
x="year", “month”, “day, “Season”
y="Rented Bike Count"
```

Figure 4: Data visualization pseudo-code

The Score Model above shows that if this date, time, temperature, and season are the same, it is necessary to prepare the number of bicycles at the level of Scored Label, but it is difficult to determine the correlation in a single graph. Therefore, it is necessary to visualize data to know the correlation with seasons and temperatures. Although it is also supported by MS Azure ML, visualization using

Python and Seaborn data visualization packages was carried out to make it easier and easier to understand at a glance.

The visualization of the above data shows that the demand for bicycles in 2018 is increasing compared to 2017, and the demand for bicycles is lower in winter than in spring, summer, and fall. It can be seen that demand is low in December, January, and February and that the usage is low on average on 1, 2, 3, and 12 days as of the date by subdividing.

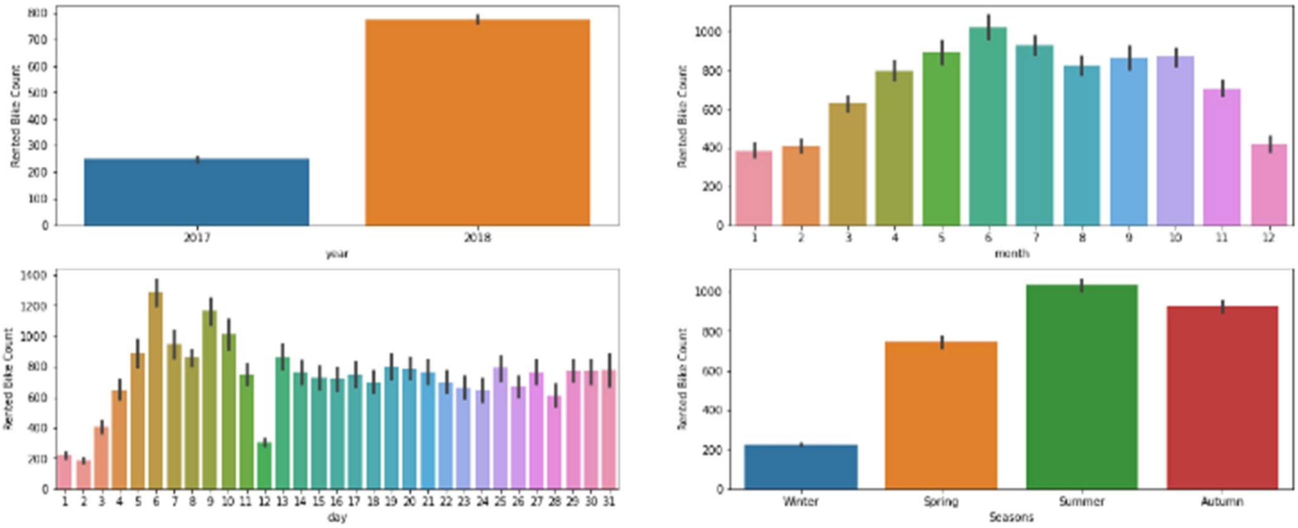


Figure 5: Data visualization

5. Conclusions

Analyzing the demand trend of Seoul's public bicycle "Ttareungi" through linear regression, it is recommended to increase the number of bicycles available in summer and fall through trends according to the year and changes in demand according to the season and to collect and manage them in winter and early spring.

Analysis and visualization of such data can reduce management costs within maintenance and limited resources in the sharing society, which is becoming a recent trend, as well as increase user convenience. In addition, the increase in shared vehicles, shows the possibility of further increasing satisfaction by increasing or expanding the number of stops or rental stations used by many people through GPS linkage in the future. In a further study, we will focus on shared bike routes by using GPS tracking systems. Through the GPS data found, the route used by most people will be analyzed to derive the optimal route when installing a bicycle-only road.

References

Abdulqader D. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). Machine Learning Supervised Algorithms of Gene Selection: A Review. *Technology Reports of Kansai University, 62*(3), 23-27.

Barga, R., Fontama, V., & Tok, W. H. (2015). Introducing microsoft azure machine learning. *In Predictive Analytics with Microsoft Azure Machine Learning*. pp 21-43. Berkeley, CA: Apress.

Barnes, J. (2015). *Azure machine learning. Microsoft Azure Essentials*(1st ed), Washington, USA: Microsoft.

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American journal of epidemiology, 188*(12), 2222-2239.

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. *In Machine learning techniques for multimedia*. 21-49. Heidelberg, Berlin: Springer.

Dehghan, M. H., Hamidi, F., & Salajegheh, M. (2015, September). Study of linear regression based on least squares and fuzzy least absolutes deviations and its application in geography. *In 2015 4th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)* (pp. 1-6). IEEE.

Dehghan, M. H., Hamidi, F., & Salajegheh, M. (2015, September). Study of linear regression based on least squares and fuzzy least absolutes deviations and its application in geography. *In 2015 4th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)* (pp. 1-6). IEEE.

DeMaio, P. (2009). Bike-sharing: History, Impacts, Models of Provision, and Future. *Journal of Public Transportation, 12*(4), 41-56.
Doi: http://doi.org/10.5038/2375-0901.12.4.3

Di Bucchianico, A. (2008). Coefficient of Determination. *The Concise Encyclopedia of Statistics*. NY, USA: Springer, Doi: https://doi.org/10.1007/978-0-387-32833-162

Dike, H. U., Zhou, Y., Deveerasetty, K. K., & Wu, Q. (2018, October). Unsupervised learning based on artificial neural network: A review. *IEEE International Conference on Cyborg and Bionic Systems (CBS)* (pp. 322-327). IEEE.

Douglas C. Montgomery, Elizabeth A. Peck, G. & Geoffrey Vinning (2012). *Introduction to Linear Regression Analysis (sixth edition).* New York, USA: Wiley.

El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. *In machine learning in radiation oncology.* 3-11. Springer, Cham.

IBM(2020). *Linear regression*. Retrieved from https://www.ibm.com/topics/linear-regression

J. Schuijbroek, R.C. Hampshire, W.-J. van Hoeve (2017). Inventory rebalancing and vehicle routing in bike sharing systems. *European Journal of Operational Research, 257*(3), 992-1004.
Doi: https://doi.org/10.1016/j.ejor.2016.08.029

Kang, M., & Choi, E. (2021). *MACHINE LEARNING: Concepts, Tools and Data Visualization*. London, En: World Scientific.

Khushbu K., Suniti Y. (2018). Linear regression analysis study. *CURRICULUM IN CARDIOLOGY – STATISTICS. 4*(1), 33-38. Doi: 10.4103/jpcs.jpcs_8_18

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal, 13*, 8-17.

K. P. Murphy (2012). *Machine learning: a probabilistic perspective.* Cambridge U.S.A.: MIT press.
Doi: https://doi.org/10.1016/j.csbj.2014.11.005.

Lim H. J., Jung K. H. (2019). Development of Demand Forecasting Model for Seoul Shared Bicycle. *The Journal of the Korea Contents Association, 19*(1), 132–140.
Doi: https://doi.org/10.5392/JKCA.2019.19.01.132

Liu, B. (2011). *Supervised learning. In Web data mining*. 63-132. Heidelberg, Berlin: Springer.

Mahesh, B. (2018). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR). 9*, 381-386.

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends, 1*(4), 140-147.

Microsoft (2021). *Azure Machine Learning.* Retrieved from https://azure.microsoft.com/ko-kr/services/machine-learning/

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning.* Cambridge U.S.A.: MIT press.

Paluszek, M., & Thomas, S. (2016). *MATLAB.* Boston, USA: machine learning Apress.

Shwartz, S. S. (2021). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge U.S.A.: Cambridge university press.

Raviv, T., Tzur, M., Forma, I. A. (2013). Static repositioning in a bike-sharing system: models and solution approaches. *EURO Journal on Transportation and Logistics. 2*(3), 187-229. Doi: https://doi.org/10.1007/s13676-012-0017-6.

Wikipedia (2021). *Linear regression.* Retrieved from May 30, 2022. https://en.wikipedia.org/wiki/Linear_regression