Journal of Korea Multimedia Society Vol. 25, No. 2, February 2022(pp. 287-296)
https://doi.org/10.9717/kmms.2022.25.2.287

A Generation and Accuracy Evaluation of Common Metadata Prediction Model Using Public Bicycle Data and Imputation Method

Jong-Chan, Kim[†], Se-Hoon, Jung[††]

ABSTRACT

Today, air pollution is becoming a severe issue worldwide and various policies are being implemented to solve environmental pollution. In major cities, public bicycles are installed and operated to reduce pollution and solve transportation problems, and operational information is collected in real time. However, research using public bicycle operation information data has not been processed. This study uses the daily weather data of Korea Meteorological Agency and real-time air pollution data of Korea Environment Corporation to predict the amount of daily rental bicycles. Cross- validation, principal component analysis and multiple regression analysis were used to determine the independent variables of the predictive model. Then, the study selected the elements that satisfy the significance level, constructed a model, predicted the amount of daily rental bicycles, and measured the accuracy.

Key words: Big Data, Imputation Method, Public Bicycle, Prediction Model, Regression.

1. INTRODUCTION

Environmental pollution has become an issue all over the world, and traffic congestion and air pollution have become serious problems in big cities in each country. In addition, air pollution issues have attracted attention in Korea due to various causes such as dust, industrial, factory, and automobile soot. Large cities in various countries are implementing various policies to prevent and improve such environmental pollution. Among these various policies, public bicycles are installed and operated in each city to solve not only environmental pollution but also traffic problems. Public bicycles are public transport systems that have a process of renting and returning bicycles at low prices by installing a self-governing lending facility in a country or local government [1-3]. They are convenient to use and are mainly used for short distances [4-7]. As a result, they are part of movement by means of transportation such as private use, bus, taxi, etc. In Korea, public bicycles first started in 2008 under the name “Nubiza” in Changwon City [8]. According to the National Transportation Research Institute’s survey on bicycle use in 2016, it was found that public bicycles operated in 12 cities nationwide by 2016 [9-10]. At present, Seoul City has expanded public bicycle rentals, so public bicycles have become available throughout the city. Consequently, the demand for public bicycles and the investment of each local government will increase. The municipalities that implement a public bicycle policy collect historical data of the public bicycles being operated and are man-

※ Corresponding Author : Se-Hoon, Jung, Address: (36729) 1375 Gyeongdong-ro, Andong-si, Gyeongsnag-buk-do, Korea, TEL : +82-54-820-6894, FAX : +82-54-820-6825, E-mail : jungsh@anu.ac.kr

Receipt date : Jan. 12, 2022, Revision date : Jan. 28, 2022
Approval date : Feb. 3, 2022

[†] Dept. of Computer Engineering, Sunchon National University(E-mail : seaghost@scnu.ac.kr)

[††] School of Creative Convergence, Andong National University

※ This work was supported by a grant from 2020 Research Fund of Andong National University

aged by each local government or private enterprise. As of April 2017, the average daily lending amount was 400 at 38 lenders. In other cities, public bicycle data is stored in real time, but it is not utilized. Therefore, research using the collected bicycle data is needed. In this study, the public bicycle data of Suncheon city, the daily weather data of the Korea Meteorological Administration, and the air information data of Korea Environment Corporation are collected.

This study proposes a bicycle rental amount prediction method considering only the factors that are significant among the collected data. Firstly, the data for generating the predictive model is collected and the collected data is subjected to cross validation and principal component analysis. Secondly, data that satisfies the significance level is adopted through multiple regression analysis. Thirdly, the study makes a prediction about the bicycle rental amount given the significant variables as the input value of the prediction model. Finally, the study confirms the accuracy of the prediction method of the proposed bicycle rental amount.

The composition of this paper is as follows. In Section 2, we introduce the existing research related to the prediction model. Section 3 describes the data specification and prediction model design used in the proposed prediction model. Section 4 describes the implementation and accuracy of the proposed prediction model. Finally, Section 5 presents conclusions and future research directions.

2. RELATED RESEARCH

Unstructured data refers to a data form which is not being structured. It is quite difficult to make an attempt on informatizing a certain relationship identified in a large amount of data shared on the internet effectively. Also, sometimes an answer can be obtained from the personal insight into the data rather than from the data itself, or acquire a new insight through analysis [11]. Big data is being used widely in the services sector or for social analytics and its demand for the construction of infrastructure for social analytics employing the online wording or related wording is constantly increasing as well [12]. At the same time, the various types of infrastructures for social analytics are being recommended along with the approach to end-users based on the public trend and its changes, as well as the changes in the wording in them. It is also necessary to consider what kind of data should be used. Although it is possible to proceed with research or marketing based on the self-collected data but such a way may become quite a burden when learning about big data for the first time. As dataset itself has considerable value, it is not common to provide it openly for personal profit. However, a number of websites are offering a series of quality datasets currently. Meanwhile, in the Republic of Korea, the Big Data Campus is providing necessary infrastructure but as its demerit is that it is available for offline use only and those who wish to use the infrastructure will have to actually visit Sangamdong. Despite such a negative aspect, it is quite helpful for learners of big data.

2.1 Multi Regression Analysis

Regression analysis is a statistical method to explain causal relations in nature or in society with explanatory variables to influence and response variables to be influenced. A regression model expresses response variables with the function of explanatory variables, and an estimated regression model is used to predict the values of response variables with those of explanatory variables. Binomial types expressed in Boolean values are used for response variables in regression analysis. When there are three values of response variables or more, multinomial and continuous types are used. Regression analysis, in general, is on the premise of linear relations between independent and dependent variables. There are interactive effects in such linear relations just like the increasing values of independent variables will lead to the certain increase or decrease of dependent variables between weight and height, for instance. Eq. (1) shows a linear functional formula to present relations between correlated independent and dependent variables[13]. Multiple regression analysis has the same basic concept as simple linear regression analysis, but it uses two independent variables or more. Predictive abilities can be increased by using many different independent variables. This model was used to match linear relations between Y group of quantitative dependent variables and X group of independent variables.

$$Y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+...++\beta_nx_n+\epsilon \quad (1)$$

2.2 Principal Component Analysis

PCA is a technique of unsupervised learning to reduce information loss of multi-dimensional input vectors through analysis and to return them to lower-dimensional vectors. It is one of the multivariate data processing techniques presented in a couple of principal component values. When there is a vector of n dimension, eigenvector is obtained through average vector and variance covariance matrix from the application of Eq. (2) and (3)[14]. Then eigenvector is arranged according to the size of the corresponding proper value to add a new matrix. The new matrix is applied as a transformation matrix to convert vector x into vector y as seen in Eq. (4)[14]. Then new variables in row y have non-correlation and are arranged in the order of monotone decreasing variance to reduce the dimensions with the big principal components of high variance value.

$$YY^Tu_k=\lambda_ku_k \quad (2)$$

$$Y^TYv_k=\lambda_kv_k \quad (3)$$

$$v_k=Y^Tu_k/\lambda_k^{\frac{1}{2}} \quad (4)$$

2.3 Data Imputation

The raw data used for data analysis or prediction model can be collected in various ways. The missing values are often found in raw data obtained from observations or experiments, and those data with such missing value are commonly discarded; thus resulting in information loss and generation of an analysis model that is biased toward a particular data. Especially, research on the statistical models to be used for analyzing the data containing the non-negligible missing values is still being carried out actively by many statisticians [15-18]. Studies on the methods of imputation, one of the statistical methodologies that can analyze the data with missing value or non-response, are also being conducted widely. As such, this paper aims to apply such imputation technique to the transmission power analysis & prediction model after modifying (complementing) it. It is essential to understand the data missing mechanism to determine an appropriate imputation technique. The data missing mechanism refers to the correlation between the missing value and data variables. The data missing mechanism can be distinguished as Missing Completely At Random (MCAR), Missing At Random (MAR), or Non-Ignorable (NI) according to the dependency of data missing on these variables. MCAR means that data missing is unrelated to any variables regardless of a missing value being included in a variable. In this case, the missing data is distributed randomly. Missing at Random refers to a condition wherein data missing is related only to the variables with missing value. Non-negligible refers to a condition wherein data missing is related only to the variables with a missing value.

3. DESIGN OF PREDICTION MODELS USING DATA

Fig. 1 shows the overall system flow diagram for the design of the prediction model proposed in this study and the accuracy measurement. The proposed system consists of five stages: Big Data collection, preprocessing, data analysis, prediction model generation, and model verification. At first,

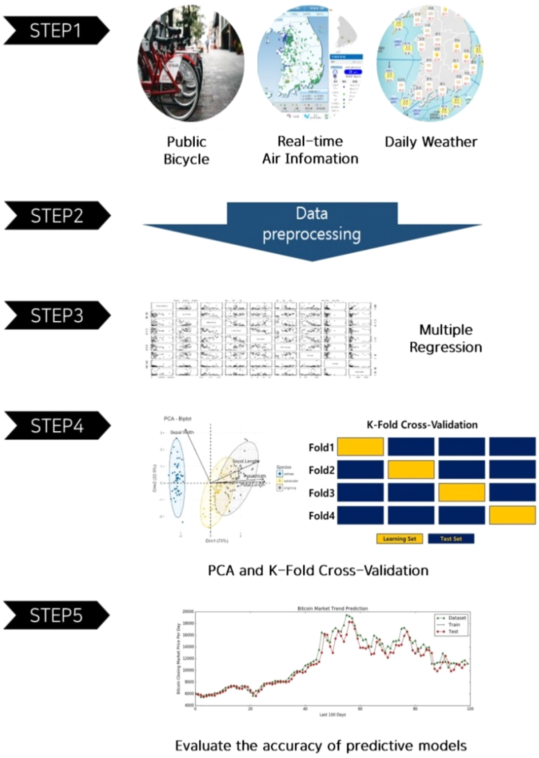

Fig. 1. Overall Flow Chart of Predictive Model Generation and Verification.

the study analyzes the data affecting the bicycle rental amount, the Korea Environment Corporation's real-time atmospheric information, which is the data affecting the amount of rental, and the daily weather data of the weather agency are collected. Secondly, unstable data such as missing or outliers are included in the data collected before analysis and this data reduces the stability and accuracy of the model. Therefore, preprocessing was performed to remove the data. Thirdly, the regression analysis confirms the relationship between the daily rental amount, the real-time waiting information, and the daily weather of the Korea Meteorological Administration to generate a prediction model that best predicts the bicycle rental amount. Fourthly, eliminating the independent variables through the previous regression analysis, a daily bicycle rental amount prediction model is created based on cross validation, multiple regression analysis, and principal component analysis. Finally, the study compares the accuracy and error rate of the predictive model with the final selected factors and the predictive model with the unnecessary factors.

3.1 Big Data Collection

The data used in this study collected domestic weather data, atmospheric data, and Suncheon public bicycle data from December 20, 2015 to December 14, 2017. The weather and air data are based on data provided by the Korea Meteorological Administration and the Korea Environment Corporation. Details of the collected data are as follows: Firstly, weather data of KMA provides information on average temperature, maximum temperature, minimum temperature, cloudiness, precipitation, average wind speed, and maximum instantaneous wind speed. In this study, the data of the meteorological office is used in units of one day to generate a forecasting model between the weather and daily bicycle lending amount. Secondly, the Korea Environmental Protection Agency's atmospheric information data is composed of information on air pollution measured by region and provides information on a daily average and hourly basis. In this study, daily average data was used to generate a bicycle lending forecasting model. Thirdly, the public bicycle operation data provided from Suncheon city was used. Operational data consisted of 11 variables including bike unique ID, rental time, return time, and terminal name. A total of 102,329 types of data was used.

3.2 Big Data Preprocessing

Before the data analysis, a preprocessing process was performed to remove the anomaly data that may affect the accuracy and stability of the analysis model. Operational data includes not only the outliers of user's cause such as confirmation that the bicycle failed after rental by the user and immediate return due to remorse, but this has no relation with the amount of rental from actual bicycle data because the data can reduce the accuracy rate in the prediction model generation and is classified as abnormal data. The above data was selected when the user's rental time and return time were less than one minute, or the driving status event was not a rental event such as breakdown or movement. Unlike existing studies, the missing data imputation method proposed in this study adopted an algorithm wherein the MLE and K-NN algorithms are combined. Compared to the existing K-NN algorithms, which determine the imputation value based on small-size or partial data, this algorithm can be used for large-size data or for all the data. Table 1 gives a description of each variable.

Fig. 2 shown the flow cahrt for missing value data imputation in data preprocessing part. A data correction technique included in a preprocessing process first replaces missing values from the collection of raw data with the result values of maximum likelihood estimation for the entire data. The goal is to enable the utilization of observed data in instances including missing values in the appli-

Table 1. Data Set.

Value	Name	Explanation
x_1	Average temperature	Average Celsius temperature for the day
x_2	Highest temperature	Maximum Celsius temperature for the day
x_3	Lowest temperature	Minimum Celsius temperature for the day
x_4	Cloudiness	Average daily cloud amount, 0.0 to 10.0
x_5	Precipitation	Total daily precipitation
x_6	Average wind speed	Average daily wind speed
x_7	Maximum wind speed	Maximum daily wind speed
x_8	PM10	A density of 10 μm or less in diameter, fine dust
x_9	PM2.5	A density of 2.5 μm or less in diameter, ultrafine dust
x_{10}	SO_2	Sulfur dioxide
x_{11}	NO_2	Nitrogen dioxide
x_{12}	O_3	Ozone
x_{13}	CO	Carbon monoxide

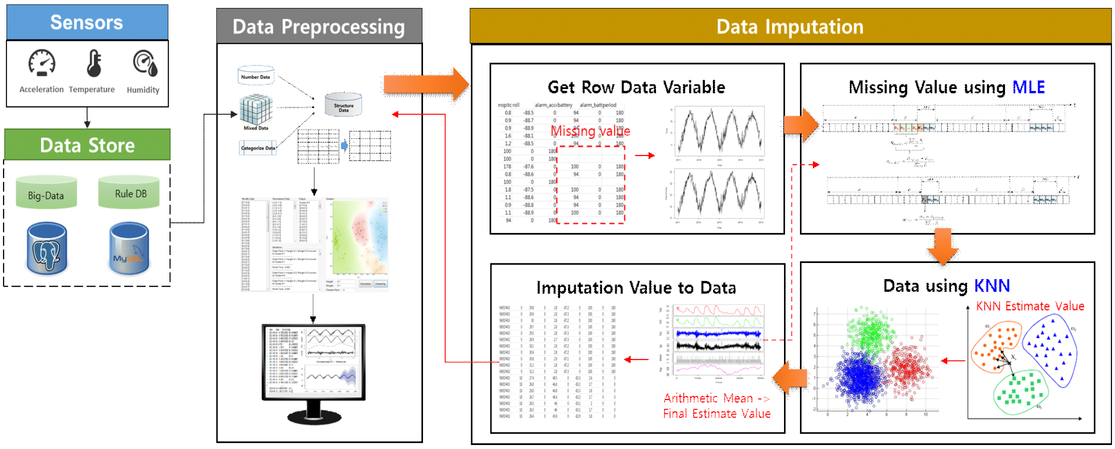

Fig. 2. Flow Chart of Data Imputation using MLE and KNN.

cation of a K-Nearest Neighbor algorithm. The maximum likelihood estimation result value of an instance among the instances of maximum likelihood estimation is changed into a missing value to apply a K-Nearest Neighbor algorithm to the entire data except for the instances changed into missing values to calculate the estimated value of a missing value. The third approach is to determine the estimated value to replace a missing value by applying a K-Nearest Neighbor algorithm to an instance with a missing value. The preprocessing process proceeds with corrected missing values.

3.3 Selection of factors influencing rental amount

Principal component analysis and multiple regression analysis were conducted after cross validation to analyze the effect of collected data on rental amount. Cross-validation can solve the problem that model accuracy is reduced by underfitting when model is learned with less data. In this study, K-Fold validation was used to divide the data into K folds. One-Fold was used as the verification data and the rest was classified as the training data. In general[19], when K-Fold-Validation is used, K is used as a method of adding or subtracting from 10, and K is set to 8 to 16 in this study. Fig. 3 is a schematic diagram that separates training data and verification data when k is 4 during K-Fold verification. Principal component analysis is an analytical method that uses the variance covariance relationship between the data to find the principal component represented by the linear combination and to reduce the dimension to the important K main components and to utilize it in model development. The principal component analysis is performed before the regression analysis because of the characteristics of the regression analysis. The multi collinearity of the regression coefficient becomes unstable due to the strong correlation between the independent variables.

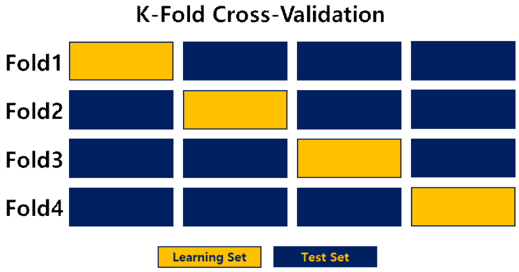

Fig. 3. The K type value is 4 the state of k-fold cross-validation.

Therefore, the effect of independent variables on dependent variables cannot be accurately explained. Principal component analysis was performed to improve the accuracy of the prediction model. Multiple regression analysis is applied to a regression model in which there are two or more independent variables and each independent variable has a linear relationship with the dependent variable. Table 2 shows the principal component analysis results. The main component of the independent variables was the cumulative proportion of 85% or more.

Table 3 show multiple regression results. To apply the principal component analysis to the regression analysis, the matrix data is subjected to principal component analysis on the existing data. Multiple regression analysis was used to exclude independent variables that did not meet the significance level of 0.05 to find significant independent variables for dependent variables. As a result,

Table 2. Principal component analysis result

Classification	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC1	2.0065	0.4474	0.4474
PC2	1.4367	0.2294	0.6767
PC3	1.2308	0.1683	0.8450
PC4	0.9404	0.0982	0.9433
PC5	0.4781	0.0254	0.9687
PC6	0.4432	0.0218	0.9905
PC7	0.2329	0.0060	0.9966
PC8	0.1646	0.0030	0.9996
PC9	0.0617	0.0004	1.0000

Table 3. Multiple regression analysis.

Classification	Standard Coefficient	Standard Error	t-Value	p-Value
Average Temperature	-13.490	2.0545	-6.566	1.18e-10
Highest Temperature	24.529	2.0273	12.099	2e-16
Precipitation	-3.228	0.3780	-8.525	2e-16
PM10	-1.319	0.3078	4.286	2.14e-05
PM2.5	-2.857	0.5368	-5.322	1.48e-07

p-values of average temperature, maximum temperature, precipitation, fine dust, and ultrafine dust data were measured to be less than 0.05, which was confirmed to be a significant variable to the amount of bicycle lending.

4. MODEL ANALYSIS RESULT

Table 4 shows the development environment for the design and verification of the proposed prediction model. Currently, R and Python are used as the data analysis language.

R is a language optimized for statistical calculations. It has the advantage of providing various data types and packages needed for statistical calculation, but it has a disadvantage of limited use range. Python offers a variety of statistical calculation packages, but it can be used more universally than R. In the proposed study, R is used as a data analysis language and Python is used as a prediction model verification and evaluation language.

In this study, we compare the performance of various prediction models generated through multiple regression analysis. The performance evaluation was carried out based on nine lending factors. The error was calculated for the accuracy of the model. The error is root mean square error (RMSE). The error can be measured by comparing the difference between the actual lease amount and the predicted lease amount. RMSE can be expressed by Eq. (5). y_i and $y_i^{'}$ in Eq. (5) is the number of actual measured leases and the number of predicted leases, respectively. It is the square root of the root-mean-square and is based on the standard deviation, so that large errors can be minimized.

Table 4. Development Environment.

Classification	Details
CPU	Intel Core i7-6800K 3.4GHz
RAM	24GB
Analysis Language	R, Python
Verification Language	Python 3.8
IDE	Pycharm 2020, R-studio 1.3

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^{'})^2} \tag{5}$$

Table 5 shows the accuracy and error of predictive models generated by regression analysis after principal component analysis and data imputation. To find the optimal prediction model, simple regression analysis with dependent variable and independent variable 1:1, and multiple regression analysis with dependent variable and independent variable in 1:N format were conducted. As a result, the accuracy increases as the degree of independent variables increases. However, the model with the highest accuracy was confirmed to have accuracy of 80.89% for the model consisting of precipitation, average temperature and maximum temperature.

Table 6 shows the accuracy and error rate of the predictive models generated through regression analysis after cross validation. The cross-validation was conducted in K-Fold method. Regression analysis showed that the significance level of each independent variable was 0.05 or less. Using

Table 5. Data Imputation and regression analysis.

Classification	Degree	Analysis method	Accuracy
Simple regression	1	Rental amount <- Average temperature	31.29%
		Rental amount <- Highest temperature	62.01%
		Rental amount <- Average temperature	57.69%
		Rental amount <- PM10	3.13%
		Rental amount <- PM2.5	4.10%
Multiple regression	2	Rental amount <- Precipitation + Highest temperature	70.42%
		Rental amount <- Precipitation + Average temperature	58.92%
		Rental amount <- Precipitation + PM10	19.54%
		Rental amount <- Precipitation + PM2.5	42.11%
		Rental amount <- Highest temperature + Average temperature	55.77%
		Rental amount <- Highest temperature + PM10	62.16%
		Rental amount <- Highest temperature + PM2.5	51.46%
	3	Rental amount <- Precipitation + Average temperature + Highest temperature	80.89%
		Rental amount <- Precipitation + Highest temperature + PM10	68.75%
		Rental amount <- Precipitation + Average temperature + 2.5	66.92%
	4	Rental amount <- Precipitation + Average temperature + Highest temperature + PM10	76.74%
		Rental amount <- Precipitation + Average temperature + Highest temperature + PM2.5	72.51%
	8	Rental amount <- All variables	72.05%

Table 6. Cross validation and regression analysis.

k-value	Accuracy	RMSE
8	77.16%	8.20%
9	71.97%	9.20%
10	77.67%	7.90%
11	85.51%	5.50%
12	78.45%	6.80%
13	68.49%	10.60%
14	69.26%	10.90%

the significant variables, we constructed the model and measured the accuracy and error rate of the model according to the number of intersection verification. The number of cross validation crosses was set from 8 to 14 at least. The verification result showed the highest performance when k was 11.

Table 7 shows the results of the accuracy and error rate of the predictive models generated through regression analysis after correlations between variables were analyzed by cross validation and principal component analysis. The cross validation was carried out using K-Fold method. The multiple regression analysis showed that the significance level of each independent variable was less than 0.05, and the model was constructed to measure the accuracy and error rate of the model according to k of the cross validation. The cross validation k was set from 8 to 12 at least. When k is 12, the highest accuracy and the lowest error rate were confirmed.

5. CONCLUSION AND DISCUSSION

In this study, bicycle rental amount per day was predicted by using cross validation, principal component analysis and multiple regression analysis, imputation method. Through cross validation, the occurrence of underfitting was eliminated, the cor-

Table 7. Results of the predictive model with cross validation, PCA, and multiple regression analysis.

k	Analysis method	Accuracy	RMSE
8	Rental amount <- Precipitation + Average temperature + Highest temperature	77.16%	7.10%
	Rental amount <- Precipitation + Average temperature + Highest temperature + PM10	75.84%	8.50%
	Rental amount <- Precipitation + Average temperature + Highest temperature + PM2.5	78.19%	7.90%
9	Rental amount <- Precipitation + Average temperature + Highest temperature	78.29%	7.60%
	Rental amount <- Precipitation + Average temperature + Highest temperature + PM10	75.80%	8.70%
	Rental amount <- Precipitation + Average temperature + Highest temperature + PM2.5	73.89%	9.60%
10	Rental amount <- Precipitation + Average temperature + Highest temperature	71.66%	10.7%
	Rental amount <- Precipitation + Average temperature + Highest temperature + PM10	77.30%	7.40%
	Rental amount <- Precipitation + Average temperature + Highest temperature + PM2.5	69.42%	9.80%
11	Rental amount <- Precipitation + Average temperature + Highest temperature	82.57%	5.40%
	Rental amount <- Precipitation + Average temperature + Highest temperature + PM10	75.55%	8.30%
	Rental amount <- Precipitation + Average temperature + Highest temperature + PM2.5	73.98%	7.20%
12	Rental amount <- Precipitation + Average temperature + Highest temperature	86.04%	3.60%
	Rental amount <- Precipitation + Average temperature + Highest temperature + PM10	75.97%	8.20%
	Rental amount <- Precipitation + Average temperature + Highest temperature + PM2.5	72.75%	9.50%

relation between each variable was analyzed by principal component analysis, and significant variables were selected by regression analysis. The accuracy of the prediction method considering only the selected significant variables was improved by 5% and the errors were reduced compared to all variables. To evaluate the performance of the prediction model, various prediction models were generated and compared. As a result, the accuracy of the prediction model through all three analyzes was excellent at 86.04%. Because the forecast for the daily average lending amount was carried out, the data variance was larger than the hourly forecast and the effect on the precipitation dependent variable was found to be abnormally higher than the effect of other independent variables on the dependent variable. In the future, we will conduct research on forecasting the amount of rental per hour by utilizing public bicycle data of various cities as well as public bicycle data of two years. Based on the data correction algorithm used in the study, a semantic data analysis system will be built to resolve the imbalance of additional large data and figure out semantic correlations among data.

REFERENCE

[1] O. Eoin and D.B. Shmoys, "Data Analysis and Optimization for Citi Bike Sharing," *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 687-694, 2015.

[2] C. Chekuri, N. Korula, and M. Pal, "Improved Algorithms for Orienteering and Related Problems," *ACM Transactions on Algorithms*, Vol. 8, No. 3 pp. 1-27, 2012.

[3] Z. Yongping and Z. Mi, "Environmental Benefits of Bike Sharing: A Big Data-Based Analysis," *Applied Energy*, Vol. 220, pp. 296-301, 2018.

[4] G.N. Oliveiraa, J.L. Sotomayor, R.P. Torchelsen C.T. Silva, and J.L-D. Comba, "Visual Analysis of Bike-Sharing Systems," *Computers & Graphics*, Vol. 60, pp. 119-129, 2016.

[5] F. Yang, F. Ding, X. Qu, and B. Ran, "Estimating Urban Shared-Bike Trips with Location-Based Social Networking Data," *Sustainability*, Vol. 11, No. 11, pp. 1-14, 2019.

[6] V. Mohammad, "Analysis of Potential Evapotranspiration Using Limited Weather Data,"

Applied Water Science, Vol. 7, No. 1, pp. 187-197, 2017.

[7] W. Ling, Q. Shi, and M. Abdel-Aty, "Predicting Crashes on Expressway Ramps with Real-Time Traffic and Weather Data," *Transportation research record*, Vol. 2514, No. 1, pp. 32-38, 2015.

[8] Changwon City(2018), https://www.nubija.com/main/main.do(accessed July 1, 2018).

[9] National Institute for Transportation and Commities(2018), https://nitc.trec.pdx.edu/research/project/1041/National_Electric_Bike_Owner_Survey_(accessed July 1, 2018).

[10] Peopleforbikes(2018), https://peopleforbikes.org (accessed July 1, 2018).

[11] L. Alexandros and H. V. Jagadish, "Challeng-Es and Opportunities with Big Data," *Proceedings of the VLDB Endowment*, Vol. 5, No. 12, pp. 2032-2033, 2012.

[12] V. Marx, "Biology: The Big Challenges of Big Data," *Nature*, Vol. 255, pp. 255-260, 2013.

[13] I.T. Jolliffe, "Principal Components in Regression Analysis," *Springer Series in Statistics*, New York, NY, pp. 129-155, 1986.

[14] S.H. Jung, C.S. Shin, C.Y. Yun, J.W. Park, M.H. Park, Y.H. Kim, S.B. Lee, and Ch.B. Sim, "Analysis Process based on Modify K-means for Efficiency Improvement of Electric Power Data Pattern Detection," *Journal of Korea Multimedia Society*, Vol. 20, No. 12, pp. 1960-1969, 2017.

[15] S. Qinbao and M. Shepperd, "A New Imputa-Tion Method for Small Software Project Data Sets," *Journal of Systems and Software*, Vol. 80, No. 1, pp. 51-62, 2007.

[16] P. William, "Pseudo Maximum Likelihood Estimation: The Asymptotic Distribution," *The Annals of Statistics*, Vol. 14, No. 1, pp. 355-357, 1986.

[17] T. Venta and J. Sumner, "Maximum Likelihood Estimates of Rearrangement Distance: Implementing a Representation-Theoretic Approach," *Bulletin of Mathematical Biology*, Vol. 81, No. 2, pp. 535-567, 2019.

[18] J.C. Kim, C.B. Sim, and S.H. Jung, "A Study on Automatic Missing Value Imputation Replacement Method for Data Processing in Digital Data," *Journal of Korea Multimedia Society*, Vol. 24, No. 2, pp. 245-254, 2021.

[19] P. Burman, "A Comparative Study of Ordinary Cross-validation, V-Fold Cross validation and the Repeated Learning-Testing Methods," *Biometrika*, Vol. 76, pp. 503-514, 1989.

Kim, Jong-Chan

He received his BS, MS, and PhD degrees in computer engineering from Sunchon National University, Suncheon City, Rep. of Korea, in 2000, 2002, and 2007, respectively. He joined in department of computer engineering from Sunchon National University in September, 2021. His current research interests include image processing, computer vision, content, fuzzy control, object detection, machine learning, computer graphics.

Jung, Se Hoon

He received his BS, MS, and PhD degrees in multimedia engineering from Sunchon National University, Suncheon City, Rep. of Korea, in 2010, 2012, and 2017, respectively. He worked for the Youngsan University, Rep. of Korea, from September 2018 to February 2020. He joined in school of creative convergence from Andong National University in March, 2020. His current research interests include bigdata processing, data-mining, reinforcement learning, blockchain.