Computer Communications 153 (2020) 353–366

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Using data mining techniques for bike sharing demand prediction in metropolitan city

Sathishkumar V E, Jangwoo Park, Yongyun Cho *

Department of Information and Communication Engineering, Sunchon National University, Suncheon Si, South Korea

ARTICLE INFO

Keywords:
Data mining
Predictive analytics
Public bikes
Regression
Bike sharing demand

ABSTRACT

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes. A Data mining technique is employed for overcoming the hurdles for the prediction of hourly rental bike demand. This paper discusses the models for hourly rental bike demand prediction. Data used include weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information. The paper also explores an filtering of features approach to eliminate the parameters which are not predictive and ranks the features based on its prediction performance. Five Statistical regression models were trained with their best hyperparameters using repeated cross-validation and the performance is evaluated using a testing set: (a) Linear Regression (b) Gradient Boosting Machine (c) Support Vector Machine (Radial Basis Function Kernel) (d) Boosted Trees, and (e) Extreme Gradient Boosting Trees. When all the predictors are employed, the best model Gradient Boosting Machine can give the best and highest R^2 value of 0.96 in the training set and 0.92 in the test set. Furthermore, several analyzes are carried out in Gradient Boosting Machine with different combinations of predictors to identify the most significant predictors and the relationships between them.

1. Introduction

Currently, the bike-sharing scheme is well-received throughout the world. It is a shared bike service to individuals, which is free of charge and for a short term basis at a minimal rate. Most bike-sharing systems permit people to borrow and return a bike from a bike station to another station that belongs to the same network. Bike-sharing gains a vast range of attention in recent years as part of initiatives to boost the use of cycle, improve the first mile/last mile link to other modes of transportation, and to minimize the negative effect of transport activities on the environment. Bike-sharing has significant impacts on establishing a larger cycling community, increasing the use of transportation, minimizing greenhouse gas emissions, enhancing public health and also traffic troubles.

Bike-sharing program progress is slow initially but after 1960 effective tracking tactics for bikes with improved technology are developed [1]. During this decade, the development gave rise to the rapid spread of bike-sharing systems across several continents. South Korea is turning into a land of two wheels, as cycling facilities provide transit flexibility vehicle emission reductions, health advantages, low congestion fuel efficiency, and financial benefit for individuals and also the paths are extended in cities across the nation as well as in rural areas.

The bike-sharing system is developed with accessible bikes for all residents for all cities across South Korea. The benefit of bike-sharing over renting is that riders can take a bike out of any system station and return it to any other station, enhance mobility, and benefit a greater number of users. Anyone can enjoy the benefits of the bike-sharing facility by being a member of a bike-sharing program and the user has access to a city-wide bike fleet for private utilization, either at minimal cost or free of charge. Many bike-sharing systems are automated based on cell phones or smart cards of the user. The first Korean city to introduce a bike-sharing system is Changwon Gyeongsangnam-do (Province of South Gyeongsang), that launched the Nubija (Nearly Useful Bike, Fun Attraction) in 2008, with 230 autonomous bikes running over 4600 bikes by 2012. Also, Seoul, Busan, and Daejeon, as well as in Suncheon, Jeollanam-do (South Jeolla Province) have the bike-sharing services. Currently, in 17 districts of Seoul, bike-sharing systems are offering 3200 bicycles. District governments own and command us, with the expectation of bike shares from Yeouido and Sangam World Cup Park [2]. Fig. 1 shows the Seoul Bike Ddareungi spot.

* Corresponding author.
E-mail address: yycho@scnu.ac.kr (Y. Cho).

https://doi.org/10.1016/j.comcom.2020.02.007
Received 3 December 2019; Received in revised form 6 January 2020; Accepted 3 February 2020
Available online 6 February 2020

Fig. 1. Seoul Bike Ddareungi spot.

So the constant raise of users necessitates the prediction of the number of rental bikes that were needed to make the bike sharing system to consistently work. Therefore, this research aims to use machine learning and data mining based algorithms to predict required number of rental bikes required at each hour. In this method, data mining is used as it has the reliability to solve complicated issues. Across various cities, a growing body of research has investigated weather and climate impacts on cycling, usually across combination with several other factors that can affect cycling. The results differ in the degree to which climate influences use. Pucher et al. [3] shows that U.S. cities with relatively high levels of cycling have mild winters and often little rain compared to the extreme heat and moisture that disrupts cycling. Furthermore, in Pucher and Buehler's analysis [4] estimating the percentage of cycling trips to work in U.S. and Canadian cities, rainfall and temperature are statistically important variables associated with lower cycling rates. This shows the influence of weather data in cycling patterns and selected weather parameters were used in this research.

The paper is structured as follows. Section 2 provides a comprehensive analysis of the literature review. Section 3 deals with methods of research. Section 4 provides data set description, exploratory analysis, data feature filtering and importance. Section 5 discusses various evaluation metrics used for evaluating the models, Section 6 provides model development process, Section 7 deals with results and discussion and Section 8 concludes the paper.

2. Related works

In the bike-sharing demand prediction, multiple pieces of research are carried out and some of the important works are discussed in this section. In Washington D.C. Capital Bike Sharer network, multiple linear regression and random forest algorithms are used to predict demand for rental bike [5]. A short-term prediction for the use of docking stations in Suzhou, a case is implemented in China [6]. In docking stations with one-month historical data, LSTM and GRU are used for the prediction of bike availability. As a benchmark, Random Forest algorithm is used.

There are numerous researches focussed on global bike demand prediction. Only few researches focussed on predicting bike demand in a station level [7]. A similar pre-processed dataset is used to train Random Forest algorithm which is a ensemble based model, Gradient Boosting Machine and Artificial Neural Network. Whereas, the two training methods: prediction models for check-in and check-out and other model trained with processed bike count changes directly, are applied to the three models to improve predictive accuracy. The outcomes indicate that these bike numbers change data training methods are more efficient. A method of traffic prediction per station with sub-hour granularity, a Spatio-temporal based mobility model which uses historical bike-sharing data is created [8]. The model quality is measured using a completer one-year dataset from public bike-sharing system which is the world's largest database with 2800 stations and above. It also consists more than 103 million reports of check-in/out. Results indicate an 85 percentile relative error of 0.6 for both check-in and check-out estimates.

Deep Learning techniques are also suggested for bike sharing demand prediction by many researches. Graph Convolutional Neural Network with Data-driven Graph Filter model which is successful in learning hidden pairwise heterogeneous correlations between stations to predict station-level hourly demand in a wide-scale bike-sharing network, is proposed by Lin et al. [9]. Deep learning approaches provide better results in all researches conducted in recent days but in order to execute deep learning methods more time and computation cost is required.

Historical bike data is used to predict bike sharing demand prediction using a spatio-temporal method is discussed by Zhang et al. [10]. Spatial information is used to generate more useful information. A traffic prediction model with sub-hour roughness is built on a per station basis. Data analysis and visualization system are developed to demonstrate the concept and prediction of mobility better and provide results in real-time prediction. An Efficient Gaussian Process Regressor which is based on similarity is used to predict the number of rental bikes which are required to be rented at each docking station, cluster developed and throughout the city. In addition to significantly enhance training and effectiveness of online prediction, the regressor recognizes external impact factors addressing data imbalance issues and captures the non-linearity of spatio-temporal data more effectively. Third the formulation of the General Least Square (GLS) aimed at collectively enhancing the predictions gathered by mutual strengthening. The GLS makes the final rent prediction very realistic. Transition-based Inference approach is made to infer from the causality between return and lease the city-wide bike return basis of the predicted rent demands. Based on historical data, weather data, and time data; a real-time model is developed to predict bike rent and return in diverse areas of the city during the future period [11]. A data-based network is built using a community-based detection method on the network, and two communities with the highest demand for shared bikes are identified. Data from stations in both communities are being used as a data set, and a two-layer deep LSTM model is used to predict bike rent and return, using long short term memory gating mechanism, and the ability to process recurrent neural network sequence information.

"SmartBIKER" is proposed as an efficient model for bike-sharing systems based on major city events in the historical data [12]. SmartBIKER model bike demands trends during major events use a trend prediction model to assess low or high demand bike stations and develop a relocation strategy that significantly reduces relocation costs while optimizing usage stations. The experimental assessment reveals that this approach is practical, effective and exceeds state-of-the-art relocation and predictive schemes. To predict the bike count, a Bimodal Gaussian Inhomogeneous Poisson (BGIP) is used [13]. The BGIP consists of three measures. Firstly, Poisson inhomogeneous approach is then used to explain how people approach a docking station to borrow or return the rental bikes. Secondly, Gaussian function which is bimodel is utilized to describe the intensity function of Poisson system that is inhomogeneous. A method for measuring the influence of external factors on the utilization of rental bikes is used to consistently expose the trend change in the usage of bikes by the state. Third, by estimating the average use of bikes based on sequences of rental bike check-out and rental bike check-in, the number of rental bikes is predicted. Findings indicate that this algorithm surpassed the baseline algorithms to solve the issue of bike prediction: to correctly predict the number of rental bikes and to assess whether at least one rental bike is available at a rental bike docking station.

Influence of time related information, weather/climate data, docking station venue details are considered for Bike usage prediction [14].

Adjacent stations are clustered with similar rental bike usage patterns and lagged parameter is introduced to simulate the influence of environment/weather on usage number. In the end, ARMA error is used as a multi-factor regression model to predict for a while check-out or check-in of the number of rental bikes in each cluster. Dataset of New York Bike Sharing System is used as the evaluation dataset and four baseline models were considered to assess the models prediction level. The markovian approach is used to predict the spreading of rental bikes for a rental bike sharing system which is functioning in urban area, taking into account the climate and historical event influences [15]. This approach is used in the city bike network collection in New York city. The primary insights are appropriate to urban execution and are more briefly explained as follows: (a) the results of the model address important traits of the city, i.e. for New York City, in relation with climate feature, only the temperature variable impacts the utilization of rental bike, whereas the effect of historical events considered is relatively minimal; (b) distribution of the hourly rental bike is predicted a day in advance, and that is of particulate concern of the city manager (C) the customer is aware of the possibility of finding an accessible bike or free parking space 1 day in advance at the station. Further analysis of the city comparison is given in the terms of traffic detail, vehicle usage and density of the population.

The present paper draws on the knowledge gained from early empirical studies and builds a data mining method to predict rental bike-sharing demand in Seoul City bike sharing service (Seoul Bike). Prior studies have made considerable efforts to develop statistical models for predicting demand for shared rental bikes. However, their use is restricted by many of the issues mentioned above, particularly for operational applications. Instead, this study seeks to exploit recent developments in data mining techniques to tackle the problem of prediction. This research uses data mining algorithms to predict and identify the best performing algorithm for hourly rental bike demand in the bike-sharing scheme. The best performing algorithm is then trained with various combinations of features to assess the effect of several key features and to understand their relationships.

3. Methodology

3.1. Linear regression

Linear regression (LM) in the most simplest and nursing method, that is equated with the relationship between the Y attribute of the scalar output and one or even more X attributes of the input quantity. The case of an independent attribute is known as simple linear regression, and the method is called as multiple linear regressions when more than one independent attributes are considered. Data is designed using linear predictor functions in linear regression, and from data, the unknown model parameters are estimated. Usually, Linear regression refers to a system where the conditional mean of Y is an affine function of X, given the value of X.

The model is assumed as in Eq. (1).

$$Y = \beta_0 + \beta_1 X + s \tag{1}$$

Here β_0 and β_1 are two unknown constants representing the intercept and slope, also known as parameters or coefficients, and s is the term of error.

3.2. Gradient Boosting Machine

Boosting algorithms usually makes use of weaker parameters to make them strong. Gradient Boosting Machine (GBM) entails classification machine learning methods as well as regression issues where prediction model estimation is conducted concerning poor prediction models, typically decision trees. Like other forms of boosting, it develops the model step by step. It also generalizes them by allowing an arbitrary differentiable loss function to be optimized. Like other methods of boosting, gradient boosting iteratively infuses weak "learners" into a single powerful learner. In the case of least-squares regression scenario, by minimizing the mean square error $\frac{1}{n}\Sigma\left[\hat{y}_i - y_i\right]^2$, it is simpler to describe where the goal is to "teach" a model to predict form $\hat{y} = F(x)$ values.

At each stage m, $1 \le m \le M$, 1 range of gradient boosting, the existence of an imperfect model can be assumed (at the beginning, a very weak model can be used that predicts only the mean y in the training set). By developing a new model that adds an estimator h to provide a better model, $F_{m+1}(x) = F_m(x) + h(x)$ the gradient boosting algorithm $F_{(m+1)}(x) = F_m(x) + h(x)$ improves: to find h, the gradient boosting solution starts with the analysis that an ideal h means as given in Eq. (2) or Eq. (3).

$$F_{m+1}(x) = F_m(x) + h(x) = y \tag{2}$$

and, in the same way,

$$h(x) = y - F_m(x) \tag{3}$$

Therefore, the gradient boost fits h with the residual. Like other boosting variants, each attempt to correct the errors of their predecessor. The discovery that residuals are the negative gradients of the squared error loss function for given model results in a generalization of this definition to loss functions apart from squared error and to sorting and ranking problems. Therefore, gradient boosting is a gradient descent algorithm, and generalizing it implies "plugging in" another loss and gradient. GBM showed outstanding results in various domains such as hospital admissions prediction [16] and energy consumption prediction [17].

3.3. Support Vector Machine

Support Vector Machine (SVM) is developed in 1996 to seek the finest hyper-plane classification of data [18]. The data are transformed into a high-dimensional feature space where data can be linearly segregated through kernel functions to differentiate two groups that are not linearly separable. Similarly, SVM makes a non-linear mapping of kernel function data then proceeds to linear regression in this new high-dimensional feature space. The regression function SVM was detailed in Eq. (4):

$$f(x) = \sum_{i=1}^{m} \left(\alpha_i^* - \alpha_i\right) k\left(x_i, x\right) + b \tag{4}$$

The value in the function space $f(x_i)$ and $f(x_j)$ is equal to the inner product of two vectors, x_i and x_j, i.e. $K(x_i, x_j) = f(x_i)\, f(x_j)$. All required calculations can be performed directly in the input space, without using the kernels to calculate the map $f(x)$. Some of the functions of the kernel are: (a) Linear Kernel $K(x_i, x_j) = x_i \bullet x_j$, (b) Polynomial Kernel $K(x_i, x_j) = (x_i \bullet x_j + 1)$, (c) Radial basis function kernel is given by Eq. (5),

$$k\left(x_i, \gamma_i\right) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \gamma > 0, \tag{5}$$

where the kernel parameters are x_i, x_j and γ. These kernel parameters are the tuning parameters and best values for these parameters are selected for improving the prediction accuracy. A radial-based function has numerical advantages and is used in this study [19].

3.4. Boosted Trees

Boosted Trees (BT), usually regarded as blackboost, gradient boosting to optimize functions of arbitrary loss where trees of regression are used as base learners. This algorithm employs regression trees as base learners after the classical gradient boosting. The key difference between GBM and BT is that arbitrary, optimized loss functions, can be specified through the blackboost family argument, whereas GBM employs hard-coded loss functions. Besides somewhat more flexible are the base learners (conditional inference trees). The regression fit is a predicting machine with a black box and therefore hard to interpret.

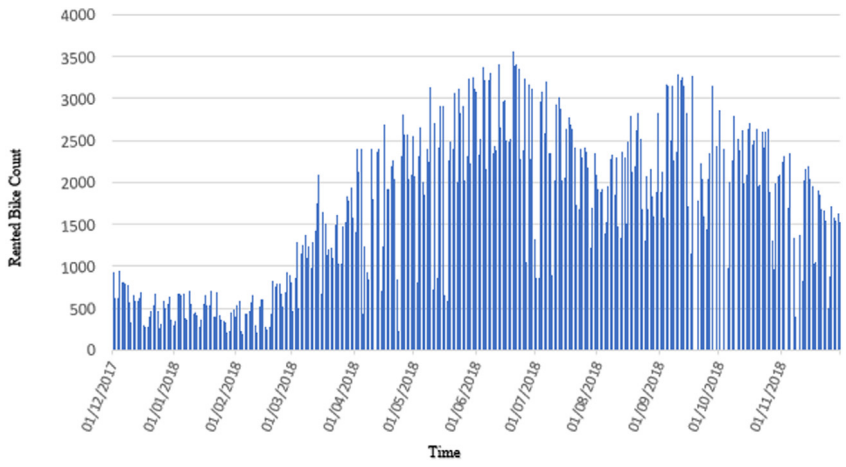


Fig. 2. Rented bike count measurement for the whole period.

3.5. Extreme Gradient Boosting Tree

In 2016 Chen and Guestrin developed Extreme Gradient Boosting Tree (XGBTree), a scalable tree-boosting machine learning platform [20]. XG Boost is the most popular method because it is used by 17 solutions out of 29 winning solutions in the Kaggle 2015 machine learning competition. Gradient boosting is the basic model of XG Boost, combining iteratively weak basic learning models into a stronger learner [21]. The residuals are used to correct the earlier predictor that the specified loss function can be optimized at each iteration of the gradient and defined by Eq. (6).

$$J(\Theta) = L(\Theta) + \Omega(\Theta) \tag{6}$$

The parameters trained from the data are given are referred to as L is the function of training loss, such as square loss or loss of logistics, which evaluates how well the model fits on training data. This is the regularization term that measures the complexity of the template, such as the L1 norm or the L2 norm. Simpler versions are more likely to perform against overfitting. A set F of k tree votes or averages the performance of the model as the base model is a decision tree and can be defined by Eq. (7):

$$\hat{y} = \sum_{i=1}^{k} f_k(x_i), \; f_k \in F \tag{7}$$

The objective function can be concrete in Eq. (8) at the time iteration as shown below,

$$J^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{t} \Omega(f_k) \tag{8}$$

where n is the number of predictions. Here the $\hat{y}_i^{(t)}$ can be given as in Eq. (9):

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{9}$$

In [20] for a decision tree, the regularization term $\Omega(f_k)$ is defined as in Eq. (10):

$$\Omega(f_k) = YT + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{10}$$

where Y is the complexity of each leaf. T is the number of leaves in the decision tree. W is a parameter that is the vector of scores on leaves to level the penalty. Then, Taylor's second-order expansion is taken to the loss function in XG Boost instead of the first order in general gradient boosting. The objective function can be derived as in Eq. (11) if the loss function is a Mean Squared Error (MSE):

$$J^{(t)} \approx \sum_{i=1}^{n} \left[g_i w_{q(x_t)} + \frac{1}{2}\left(h_i w_{q(x_t)}^2 \right) \right] + YT + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{11}$$

The constants have dropped. The q(entry) is a function where the corresponding leaf data point is allocated. And h_i is the derivative of the first and second loss function of MSE. In [22], the loss function is determined by the sum of the loss value for each data sample. Because each data sample corresponds to just one leaf node, the loss function can also be represented by the sum of the loss value for each leaf node. So Eq. (12),

$$J^{(t)} \approx \sum_{J=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2}\left(\sum_{j \in I_j} h_i + \lambda \right) w_j^2 \right] + YT \tag{12}$$

According to [23], G_j and H_j can be defined as in Eq. (13):

$$G_j = \sum_{i \in I_j} g_i, \quad H_j = \sum_{i \in I_j} h_i \tag{13}$$

In recent days, the performance of XGBTree is higher than other prediction algorithms, so XGBTree is considered to compare its performance with other algorithms used in this study.

4. Dataset description

4.1. Data preparation

This work comprehends the relation between rental bike used in each hour and the different predictors such as weather information and time information. Additionally it examines the efficiency of various regression models: (a) LM, (b) GBM, (c) SVM, (d) BT and (e) XGBTree, for predicting the demand for public rental bikes and ranking the influence of predictors or parameters in the prediction.

Data for one year (2017 December to November 2018) is downloaded from the Seoul Public Data Park website of South Korea, where the hourly public rental history of Seoul bikes is available [24]. Dataset time-span is 365 days (12 months). The number of rental bikes rented at every hour is determined from the data. Fig. 2 shows the total number of rented bikes for the entire period. The figure shows that the rental bike count is highly variable at each hour.

Fig. 3 shows the boxplot and histogram of the data. The histogram plot shows the frequency of rented bike count at each hour. Boxplot displays the position of the median value in the black line. As can be seen, there is a long tail in the data distribution. In the box plot, the median is represented inside the blue rectangle by a thick black line and has a value of 505. The value of the lower whisker is 0 and the upper whisker is 2400. There are also more dispersed data above the

Table 1
Data variables and description.

Parameters/Features	Abbreviation	Type	Measurement
Date	Date	year-month-day	–
Rented Bike count	Count	Continuous	0, 1, 2, …, 3556
Hour	Hour	Continuous	0, 1, 2, …, 23
Temperature	Temp	Continuous	°C
Humidity	Hum	Continuous	%
Windspeed	Wind	Continuous	m/s
Visibility	Visb	Continuous	10 m
Dew point temperature	Dew	Continuous	°C
Solar radiation	Solar	Continuous	MJ/m2
Rainfall	Rain	Continuous	Mm
Snowfall	Snow	Continuous	cm
Seasons	Seasons	Categorical	Autumn, Spring, Summer, Winter
Holiday	Holiday	Categorical	Holiday, Workday
Functional Day	Fday	Categorical	NoFunc, Func
Week status	Wstatus	Categorical	Weekday (Wday), Weekend (Wend)
Day of the week	Dweek	Categorical	Sunday, Monday, …, Saturday

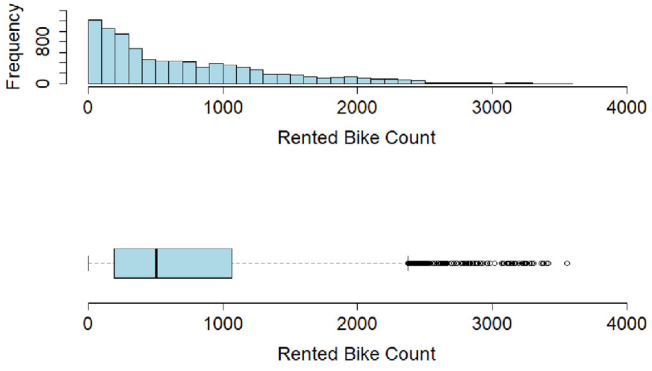


Fig. 3. Rented bike count distribution. Top: Histogram, Bottom: Boxplot.

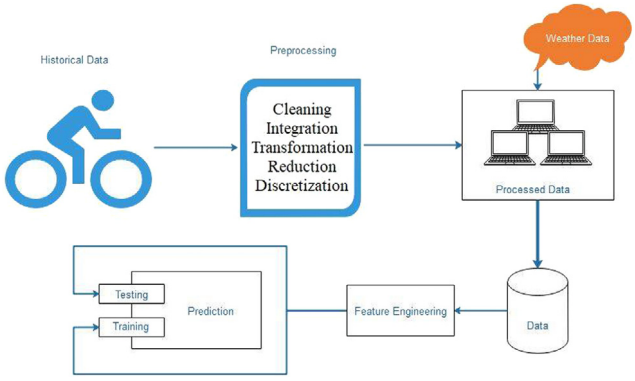


Fig. 4. System flow.

median value. And, there are many outliers displayed above the upper whisker that are marked with circles.

As all modes of transportation depends primarily on the weather conditions, the corresponding climate details such as temperature, humidity, wind speed, visibility, the temperature of the dew point, rainfall and snowfall at each hour is added The data generated now comprises of the total number of rental bikes rented at each hour with date/time variable and weather details. The next move is to build certain additional features from the date/time parameter to enhance the performance of the machine learning algorithms. This method of using domain knowledge to create additional features from available data is known as feature engineering. Week status (Weekend or weekday) and weekdays (Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday) are derived from the date/time factor. Data on holidays (official working days and holidays) is gathered and added [25]. There are four seasons in South Korea (Autumn/Spring, Summer and Winter). Added this data as well. Eventually, the functional and non-functional days, the days when the rental bike system does not operate are included. Fig. 4 shows the entire system's process flow. Table 1 lists all variables, features or parameters with their corresponding abbreviation, type (continuous or categorical) and measurement.

4.2. Exploratory data analysis

The final consolidated rental bike data is partitioned into two namely, training set for building the regression and testing set for assessing the model performance by using the data partition function generated by CARET package. Usually larger part of data is need to teach the models and so the 75% of the final data is utilized for model training and the remaining 25% of the data is used for testing purpose. The dimensions of training and testing set is shown in Table 2.

Table 2
Training and testing dataset.

Dataset	Number of observations
Training set	6571 and 16 variables
Testing set	2189 and 16 variables

Exploratory Data Analysis (EDA) is the most common data analysis methodology to analyze the data visually using different parameters and could be used to summarize the data. Fig. 5 presents pair plots displaying the relations of the rented bike count in the training set with all variables. This figure is developed using the Psych package [26]. The figure shows the bivariate scatter plots displayed in the boxes below the diagonal, the histogram graphs along the boxes in the diagonal and Spearman correlation values displayed above it. Spearman correlation is computed as a measure of monotonous relationships between two features or variables. A correlation value of 1 is considered as a total positive correlation, −1 is considered total negative correlation, and if 0 no correlation exists between the variable. The linear regression line fits are shown in red for each pair.

Fig. 5 displays the positive correlation between the count and Temp (0.56). It ensures that the demand for the rented bike rises when the temperature increases. The higher the temperature, the more rental bikes are used. The second significant correlation is between count and hour (0.39). This indicates a positive correlation as the day gets busier by the hour of the day. Positive correlations are also notable between count and wind, visb, dew and solar. There is a negative correlation of (−0.22), (−0.28) and (−0.22) between count and humid, count and rain, count and snow respectively, which signifies that humidity, rainfall and snow are the factors influencing rental bike usage in a negative manner. In other words the bike Usage decreases when the humidity or rain or snow value increases. These correlation values

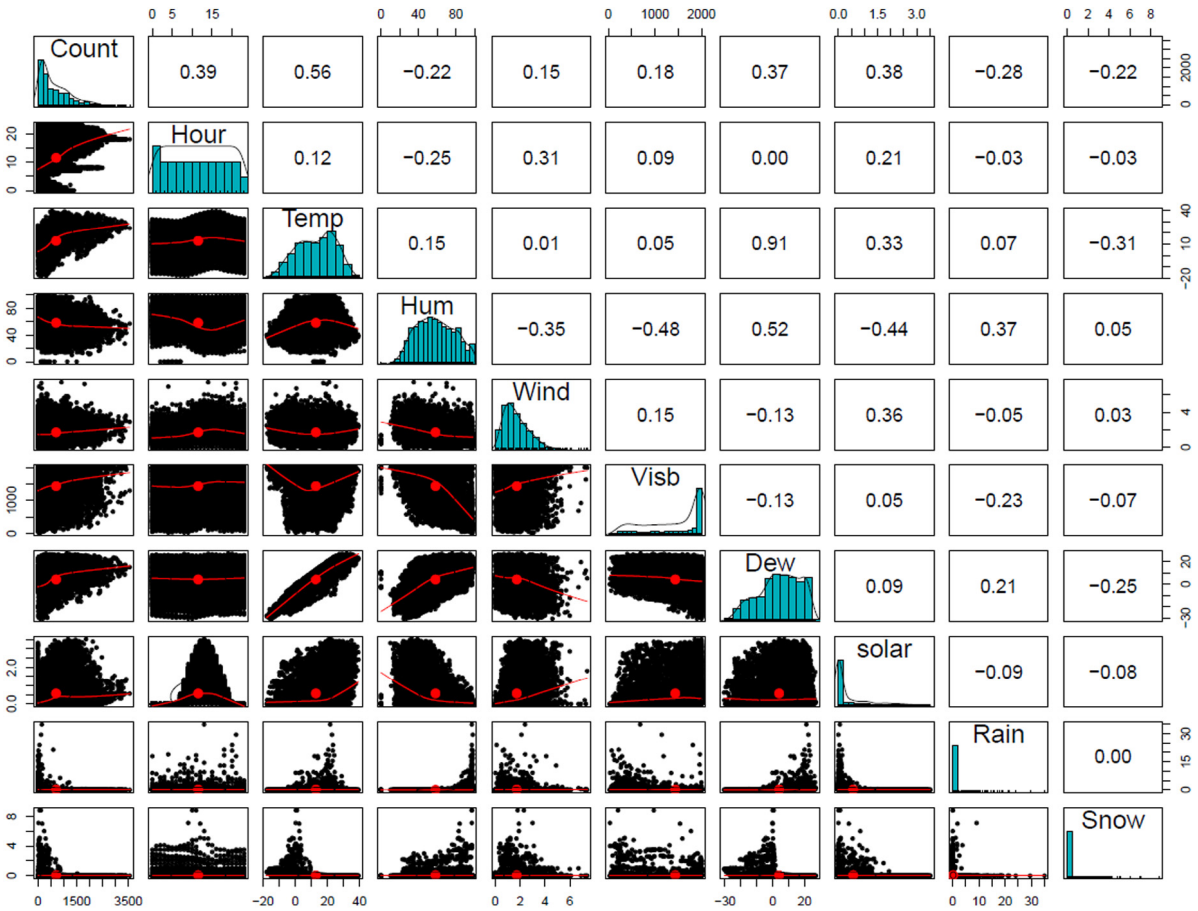


Fig. 5. Pairs plot. Relationship between bike count with Hour, Temp, Hum, Wind, Visb, Dew, Solar, Rain, Snow.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

imply that the weather variables are related to the rental bike count at each hour. The other correlation values are common correlations between weather variables.

The count of rental bike users by each hour is plotted and visualized through seasons, weekdays and month to obtain more consistent insights about the data and also to classify the time trends. From Fig. 6 it can be seen. That, the count pattern of rental bike has a strong time element. From Fig. 6(A), it can be seen that the average count is skewed by month. And Fig. 6(B) shows that the average count is high at each hour in the summer and low in the winter. The count is quite similar during autumn and spring. The count is identical from hours 12 to 15 in the autumn, spring and summer season. Fig. 6(C) displays an average number of users every hour of the day throughout the week. It is shown that the count distribution follows identical trends over the weekdays and different patterns over the weekends. Fig. 6(D) shows the average number of users per hour of the day monthly. The pattern suits all months, but the average number of users per hour varies over different months. The count abruptly increases in all figures in hours 8 and 18, except during the weekends. This is because the hours from 8 AM and 6 PM is regarded as the peak hours, during which the usage of the rental bike is high in Seoul. These figures in Fig. 6 show a strong association between the users of the rental bike and the factors like weather and day/time.

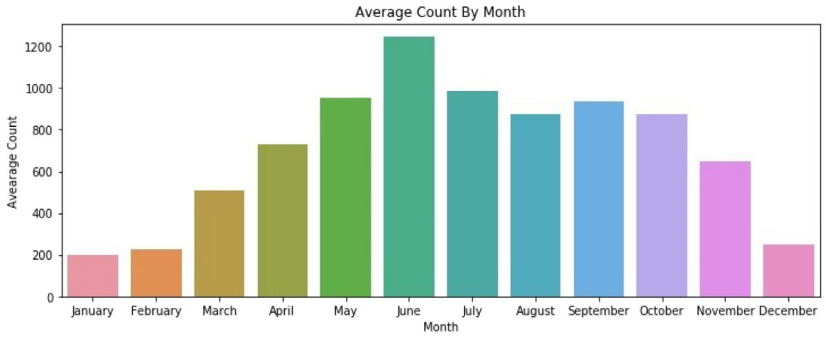
4.3. Data filtering and importance

There are several parameters/features in the data set. It is beneficial to identify the key features that boost the rental bike count prediction at each hour. Boruta package [27] is utilized to pick all relevant attributes. Boruta algorithm is a random forest based on wrapper algorithm used to select important features in the dataset with respect to the dependent variable. For several prediction algorithms, the Boruta algorithm is used for variable filtering [17,28–30].

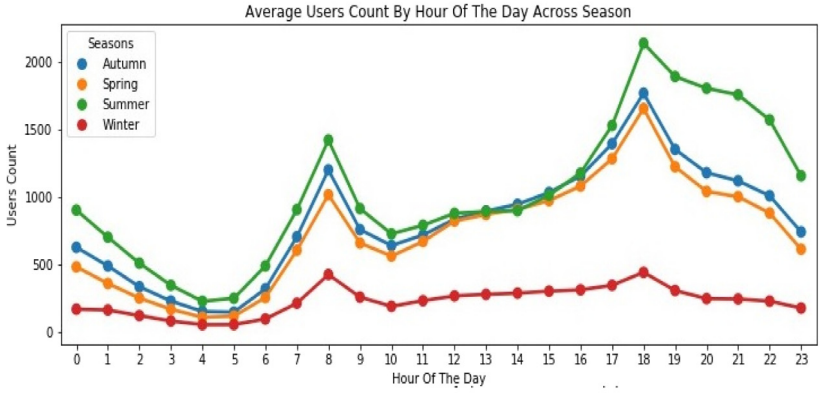
In order to clearly study the influence of categorical variables with the rental bike count, the categorical variables are converted to dummy variables. R package dummies [31] are used to construct dummy attributes. The characteristics of the Seasons, Weekend, Weekend, Working Day and Week Day are turned into dummy variables. After conversion of dummy variables, finally the number of independent variables is 26. Two random variables rv1 and rv2 are introduced to test the Boruta algorithm. Selection of the Boruta variable or feature selection helps to lower the complexity and interpretability of the model. Boruta package estimates the attribute's value with the shadow attributes generated by the original shuffling.

As shown in Fig. 7, Boruta-algorithm identifies two additional random variables (red box graph) which have no predictive effect on the prediction of rental bike counts. The two added random rv1 and rv2 variables are the shadow variables shown in blue between the Boruta algorithm: shadowMax, shadowMean, shadowMin. Also, the Boruta algorithm ranks each of the features/variables with their respective value of importance i.e., from Hour to Thursday.

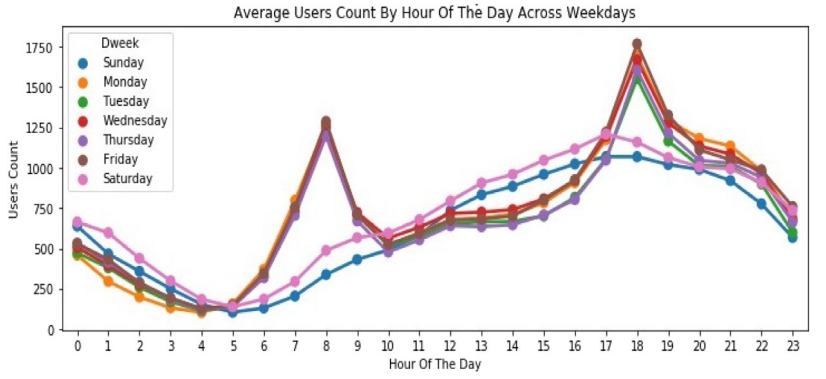
Fig. 7 offers a great deal of insight but does not provide data on the performance of the selected variables concerning RMSE values. A Recursive Feature Elimination (RFE) algorithm is used to find the number of variables needed to lessen RMSE [32]. RFE algorithm predicts the number of independent variables required to lessen the RMSE value. Classification and Regression Training (CARET) package has RFE algorithm and used in this research [33]. RFE uses random forest based on regression and a 10 fold cross-validation is used for training purpose. Fig. 8 shows the outcome of RFE algorithm and the optimal predictors is 21, shown with a filled dot.



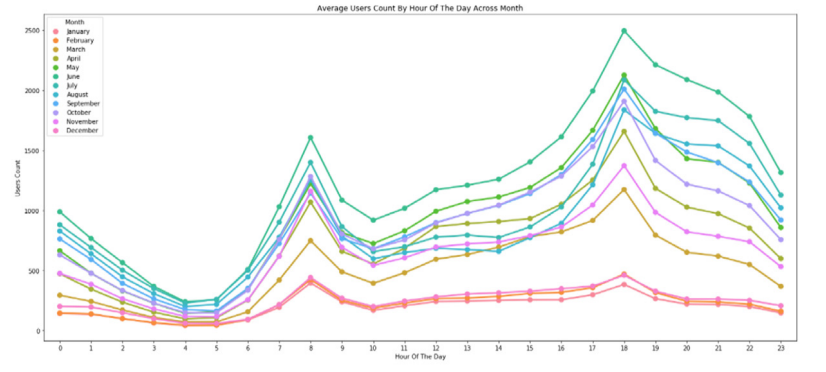
(A)



(B)



(C)



(D)

Fig. 6. (A) Average users count by month (B) Average users count by an hour of the day across the season (C) Average users count by an hour of the days across weekdays (D) Average users count by an hour of the day across month.

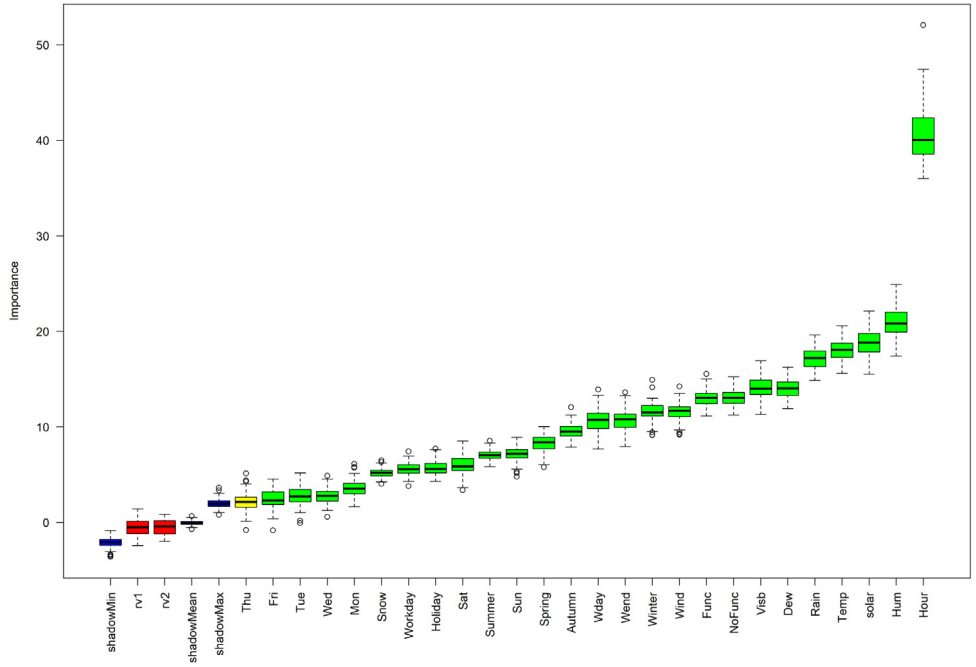


Fig. 7. Feature selection using Boruta algorithm.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

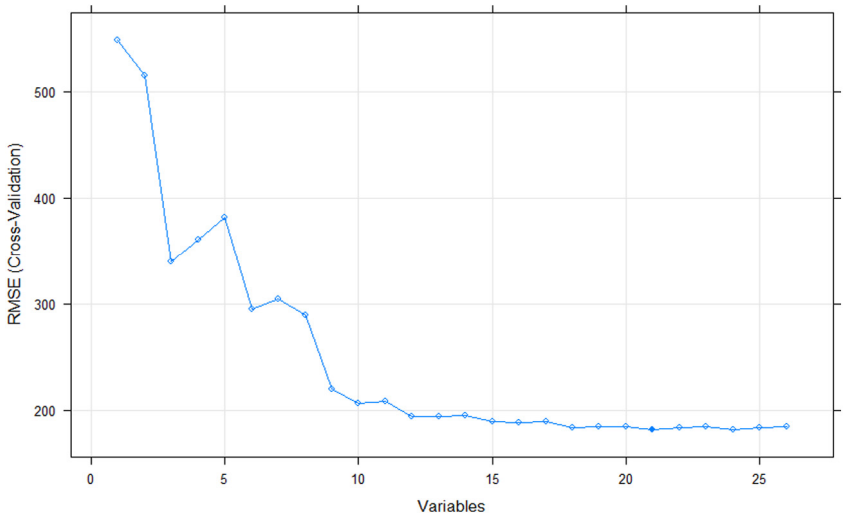


Fig. 8. RFE algorithm feature selection.

Table 3
Feature ranking from rfe algorithm.

1. Hour	2. Hum	3. Temp	4. Solar
5. Rain	6. Wend	7. Wday	8. Visb
9. Func	10. Dew	11. NoFunc	12. Autumn
13. Wind	14. Winter	15. Spring	16. Sun
17. Workday	18. Holiday	19. Sat	20. Mon
21. Summer	22. Snow	23. Fri	24. Wed
25. Tue	26. Thu		

Table 4
Models performance.

Models	Training				Testing			
	R^2	RMSE	MAE	CV (%)	R^2	RMSE	MAE	CV (%)
LM	0.55	431.72	321.84	61.15	0.55	427.71	322.32	61.03
GBM	0.96	117.81	79.77	16.68	0.92	172.73	109.78	24.64
SVM	0.92	173.58	96.32	24.58	0.85	241.94	151.21	34.52
BT	0.92	171.47	108.89	24.28	0.90	195.23	125.80	27.85
XGBTree	0.96	127.63	85.21	18.07	0.91	183.80	119.59	26.22

5. Evaluation indices

Regression models are trained to select the best with a repeated 10-fold cross-validation scheme. The doParallel package [34] is employed to accelerate computations. Various evaluation criteria are employed to test the performance of regression models. The performance assessment indices used here are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Rsquared (R^2) and Coefficient of Variation (CV).

The standard sample deviation between the observed and the predicted values of the residuals is the RMSE metric. Using this method, it is possible to detect major errors and determine the fluctuation of the model response regarding the variance. RMSE is considered as a scale-dependent metric which results in values with the same measurement

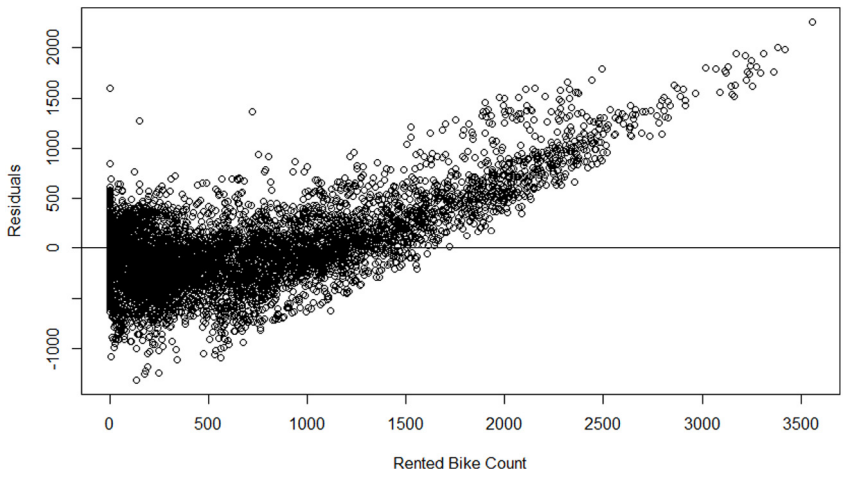


Fig. 9. Residuals and rented bike count plot of the LM.

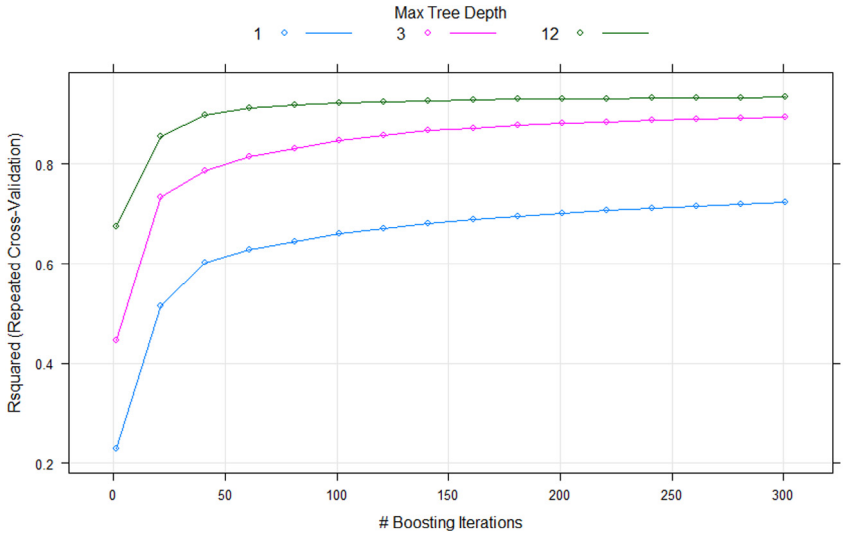


Fig. 10. Grid search results for GBM model.

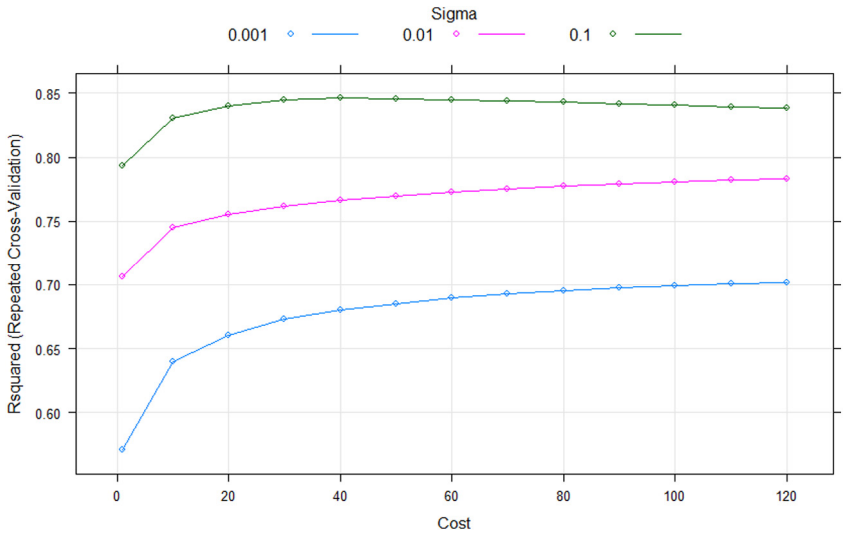


Fig. 11. Grid search results for SVM.

Table 5
Comparison of GBM Model performance with different subsets.

Models	Parameters/Features	Training				Testing			
		R^2	RMSE	MAE	CV (%)	R^2	RMSE	MAE	CV (%)
GBM - no Temp	Hour, Hum, Wind, Visb, Dew, Solar, Rain, Snow, Autumn, Spring, Summer, Winter, Holiday, Workday, NoFunc, Func, Wday, Wend, Sun, Mon, Tue, Wed, Thu, Fri., Sat	0.96	125.24	84.43	17.74	0.91	180.45	113.86	25.75
GBM - no weather data	Hour, Autumn, Spring, Summer, Winter, Holiday, Workday, NoFunc, Func, Wday, Wend, Sun, Mon, Tue, Wed, Thu, Fri, Sat	0.63	392.64	263.13	55.61	0.65	375.24	252.73	53.54
GBM - no categorical variables	Hour, Temp, Hum, Wind, Visb, Dew, Solar, Rain, Snow	0.85	247.65	160.03	35.07	0.77	301.31	188.70	42.99
GBM - no Snow, Fri, Wed, Tue, Thu	Hour, Temp, Hum, Wind, Visb, Dew, Solar, Rain, Autumn, Spring, Summer, Winter, Holiday, Workday, NoFunc, Func, Wday, Wend, Sun, Mon, Sat	0.96	118.80	80.88	16.82	0.92	174.68	109.89	24.92

units and R^2 is the determination coefficient which usually ranges from 0 to 1, representing the goodness-of-fit. A high R^2 value shows the values that are predicted match the observed values perfectly.

Equations for computing RMSE and R^2 is given in Eq. (14) and Eq. (15) respectively,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \tag{14}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2} \tag{15}$$

MAE metric is used to determine the acuteness of the prediction MAE is also scale-dependent metric like RMSE metric, that efficiently represents the error values of prediction by avoiding the offset between the positive and negative errors. MAE can be computed mathematically by using Eq. (16)

$$MAE = \frac{\sum_{i=1}^{n}[Y_i - \hat{Y}_i]}{n} \tag{16}$$

The coefficient of variance (CV) is used to determine the relative variability measure. CV determines the variation of the overall prediction error concerning the target's mean. A high CV score shows a high number of errors in the model.

Equation of CV is given by Formula (17):

$$CV = \frac{\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}}{\overline{y}} * 100 \tag{17}$$

Here, the actual measured value is Y_i, the predicted value is $\hat{Y}_i$, the sample average is, $\overline{y}$ and the sample size is n. Here, Y_i is the actual measurement value, $\hat{Y}_i$ is the predicted value, $\overline{y}$ is the average of the sample and n is the sample size.

6. Model development

In order to find and decrease the error values while fitting a model, it is necessary to find optimal tuning parameters for each of the regression algorithms. The caret package provides a grid search function for determining the optimal possible parameter values for a model. The grid search provides interactive approach to try all combinations of hyperparameters and select the best hyperparameters.

Fig. 9 exhibits a linear regression model residual map. In case of LM, residuals are determined as the difference between the actual value and the predicted values. Linear regression tries to fit the dependent variable in a linear line derived by the regression equation. Fig. 9 shows the relationship between the variables considered in the rented bike count is not very well represented by the linear regression model as the residuals are also not well dispersed around the horizontal axis.

Hyperparameters for GBM include the number of trees (Boosting iterations) and the maximum depth of the tree (Max Tree Depth). As shown in Fig. 10, GBM model attempts to boost prediction from first tree itself and tree depth information 1. The search for the optimal depth of the hyperparameter maximum tree ranges from 1 to 15 and for number of trees ranges from 1 to 500. The optimal value for the hyperparameter number of trees is 301 and maximum tree depth is 12. After finding the best hyperparameters, the models are trained with its best hyperparameters. In addition to the predictors, two tuning parameters are necessary for the SVM model, namely sigma and cost. The search for cost is within the range 0 to 120 and sigma values 0.001, 0.01, 0.1 are used. The optimal values are obtained for the variables sigma (0.1) and cost (40) with a grid search shown in Fig. 11. R^2 value remains constant after cost value 40 and the R^2 values begin to drop after 40 values. So 40 is selected as the best value for the hyperparameter cost.

The maximum tree depth of 10 and mstop(trees) 41 for the BT model are the optimal tuning parameters and the grid search outcomes are shown in Fig. 12. It can be seen that the R^2 value in all Maximum Tree Depth Values remains constant after Trees value 40.

From Fig. 13, XGBTree model raises R^2 values from tree depth 1 and the final optimal parameters include 3 maximum tree depth values and 1 to 1100 rounds (boosting iterations).

7. Results and discussion

Upon training each regression model, each of the regression prediction model has 30 outcomes from the 10-fold cross-validation sets repeated for 3 times. For each model, CARET uses this data to plot R^2, RMSE and MAE values along with the confidence intervals as shown in Fig. 14. The best model is the model with lower MAE, RMSE and CV values as well as higher R^2 values. This is because the error values should be less and R^2 describes the explanation of the fit, so this value should be higher. Table 4 shows that the GBM has the lowest RMSE, MAE and CV values and higher R^2 values than other regression models. XGBTree model also reveals much reduction of RMSE, MAE and CV and higher R^2 value when compared to BT, SVM and LM. XGBTree

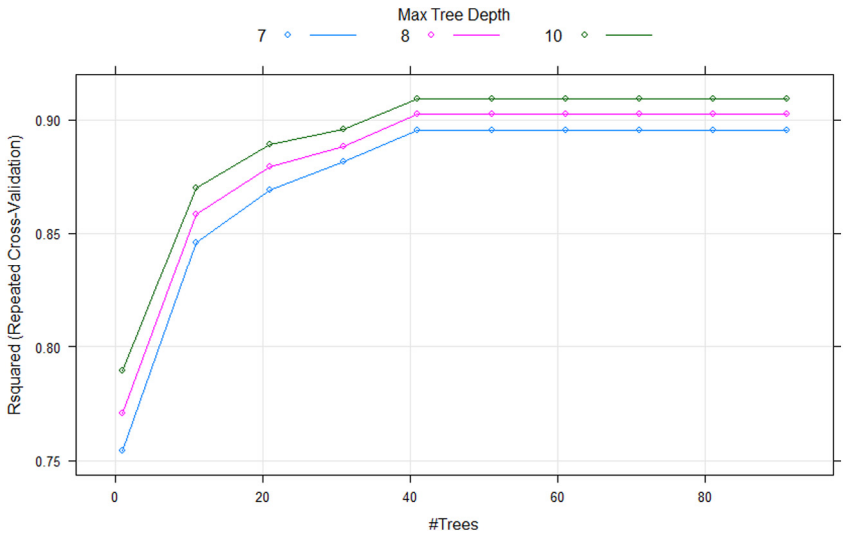


Fig. 12. Grid search results for BT model.

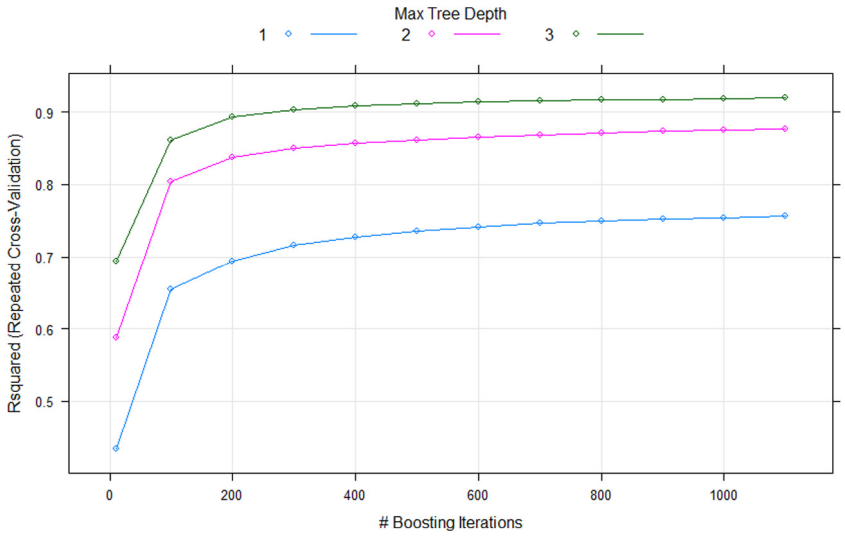


Fig. 13. Grid search results for XGBTree model.

model has R^2 of 0.96 equal to GBM model R^2 value in the training set, but the R^2 value is 0.91 in the testing set which is 0.1 lesser than GBM model. The GBM model has the highest predictive outcomes in the test set. Table 4 shows the efficiency of the trained regression models both in testing and training sets. LM model produces the worst results compared to other models. This shows that the usage count is not linearly related to any of the independent variables.

The relative significance of the LM, GBM, SVM, XGB and BT models is shown in Fig. 15. The residual sum of squares is used for tree based models GBM, BT and XGB to calculate the significance for each of the variable. For LM, the most linearly fitted variable for predicting the dependent variable is ordered and used as a measure to calculate the variable importance. In case of SVM model, the predictor and the outcome relationship is evaluated and a linear based model is then fitted and the absolute value of the t value for the predictor slope is used [35].

As can be seen, GBM model offers the best prediction results (Higher R^2, Lower RMSE MAE and CV) in the previous research results. So, this model is used with different predictor subsets: excluding temperature variable, eliminating all weather data, removing categorical variables derived using feature engineering, eliminating snow Friday, Wednesday, Tuesday and Thursday to test prediction performance. Table 5 provides information on the performance of the predictors considered both in training set and test sets. Fig. 16 provides R^2, RMSE and MAE confidence interval plots for GBM models built using different subsets.

As can be noted in Fig. 2, rental bike count profile is highly variable, with almost constant demand intervals. The box plot is presented in Fig. 3, which shows the values highly dispersed above the median. The data filtering approaches such a as boruta and RFE shown in Section 4 is vital as it helps to identify and rank the predictors that do not have an impact on the accuracy of predictions. The Boruta algorithm observed two random variables in the dataset and also revealed that all the considered parameters are related to the problem of prediction. As shown in Fig. 8, the RMSE values can be significantly reduced by 21 parameters according to the RFE algorithm. Table 3 lists the parameters.

The plots with R^2, RMSE and MAE with the respective level of confidence are proven to be effective in the process of predicting the efficiency of each regression model using the test set. GBM and XGBTree are the best models based on R^2, RMSE and MAE values as shown in Table 4. In training set, the GBM model and XGBTree model developed have an R^2 value of 0.96, but in the test set, GBM model has an highest R^2 value of 0.92 greater than the R^2 value in the test set of the XGBTree model. In the test set, the RMSE, MAE, and CV values of

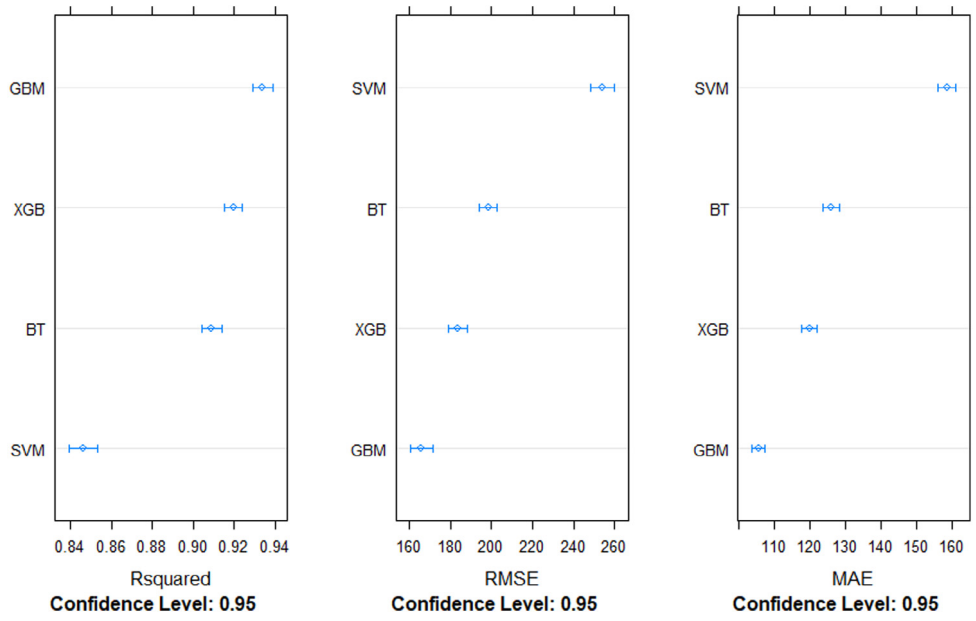


Fig. 14. Rsquared, RMSE and MAE values with confidence interval for trained models.

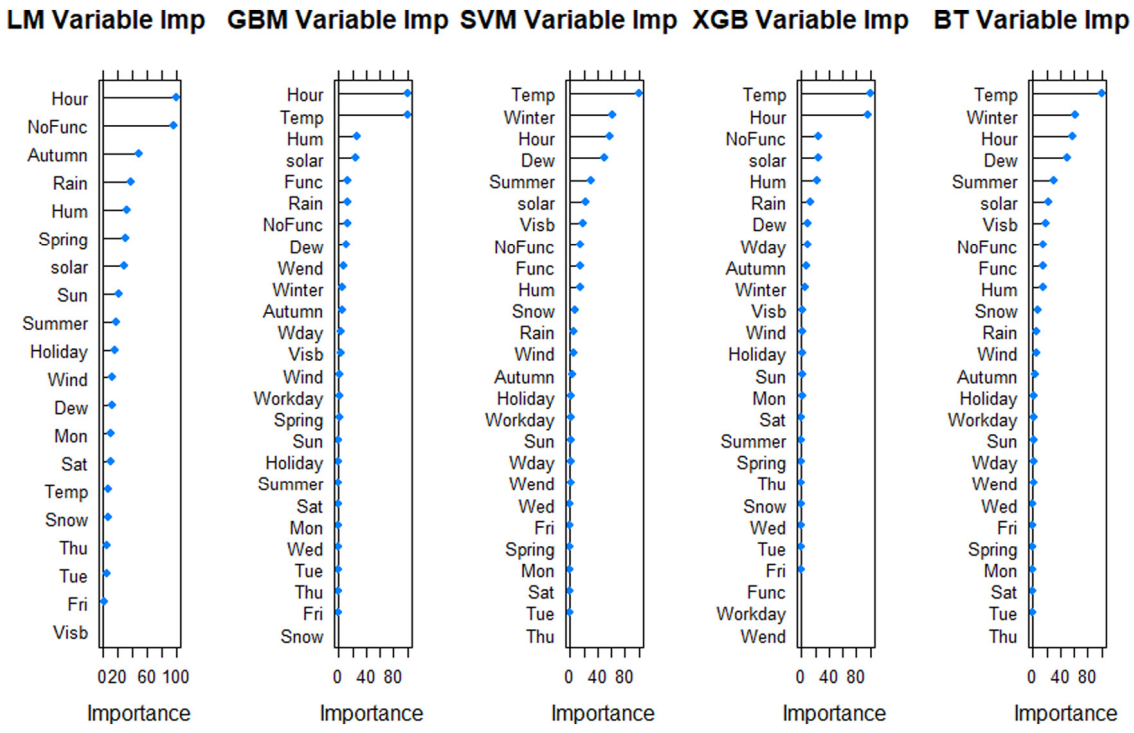


Fig. 15. LM, GBM, SVM, XGB and BT model variable importance.

the GBM model are lower than those of XGBTree. The GBM model is therefore regarded as the best model producing lower error values.

Regarding the significance of each of the independent variables with respect to each of the prediction models displayed in Fig. 15, the most significant predictor for three models (SVM, XGBTree and BT) is Temp. The main predictor is the hour for LM and GBM model. Temp for the GBM model is the second most vital predictor. For all models excluding LM, Hour and Temp are considered to be the main predictors and are put in within the top 3 predictors in these 4 models. Winter is picked as the second most important predictor by SVM and BT models. Temp and Hour are considered the top predictor in the top two performing models XGBTree and GBM. It proves the close association of the weather variable temperature, time-variable Hour with the rental bike count.

As the weather data and time variables are closely linked with the number of rented bikes, studying this data sub-set individually could provide better understanding of the different ranks. Table 5 shows that the GBM model excluding the variables Rain, Fri, Wed, Tue and Thu is as effective as Table 4 variables with R^2 of 0.92 in the test set. The weather and categorical data (derived from the function engineering) enhanced the models performance. The GBM model resulted in smaller R^2 and larger RMSE, MAE and CV values without weather data and categorical variables. This shows the importance of weather and feature engineering variables in predicting bike sharing demand. Looking at

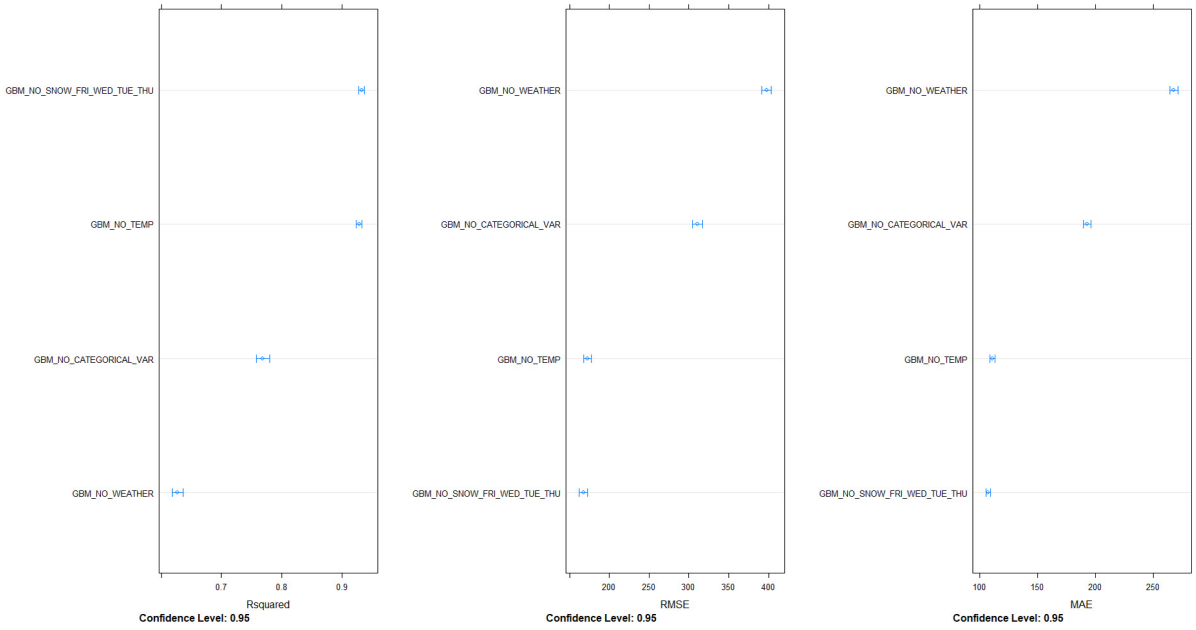


Fig. 16. R^2, RMSE and MAE confidence interval plots for GBM models with different subsets.

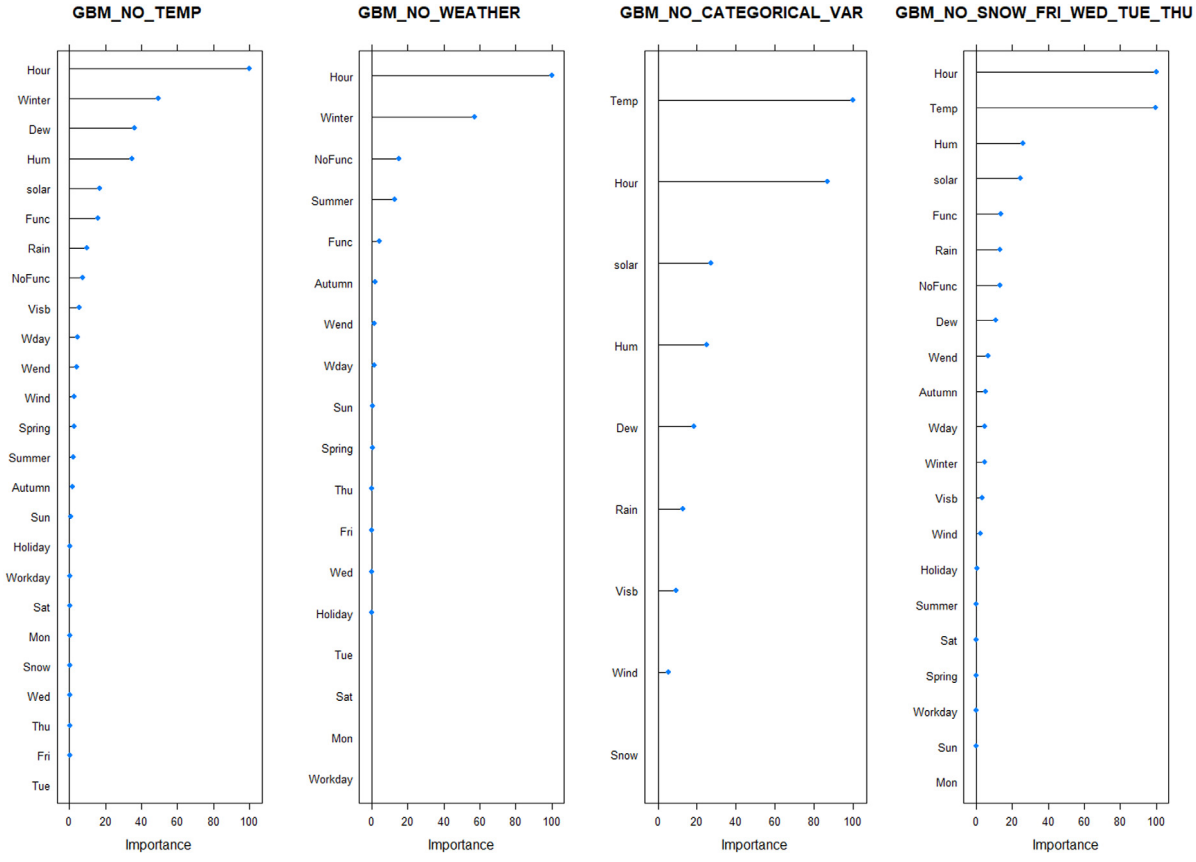


Fig. 17. GBM model variable importance using different subsets.

the GBM model ranking with various subsets in Fig. 17, it is clearly seen that the top ranked predictor is the variable hour for three subsets and only for GBM model with no categorical variable Temp is the top ranked predictor. This ensures that the most important variable or feature for GBM is the time information and the weather variables.

8. Conclusion

The data analysis and prediction provides a thought-provoking outcome for both the data exploratory research and the prediction models. The generated Pairwise plots based on their correlation, certainly show different parameter relationships that can be concealed in the most used

prediction models. GBM and XGBTree models enhance the R^2, RMSE, MAE and CV of predictions rather than SVM, LM and BT. Temp and Hour are considered as the most significant variable for the hourly rental bike count prediction in all models except LM. In all prediction models, weather data is shown to increase prediction accuracy. It shows the impact of weather data on the rental bike users count. This study uncovers the relationship between the weather data and the rental bike users count for a specific hour. Future work includes the prediction of the rental bike demand for district level. This study predicts the rental bike demand for the entire Seoul region. Prediction based on district wise rental bike demand will be much more useful for making the public rental bikes available to the public in a consistent manner.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Sathishkumar V E: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Visualization, Investigation, Validation. **Jangwoo Park:** Supervision, Writing - review & editing. **Yongyun Cho:** Supervision, Writing - review & editing.

References

[1] Paul DeMaio, Bike-sharing: History, impacts, models of provision, and future, J. Public Transp. 12 (2009) 41–56.

[2] Bikesharing spreads in Korea, LINK: http://www.korea.net/NewsFocus/Society/view?articleId=107208#.

[3] John Pucher, Charles Komanoff, Paul Schimek, Bicycling renaissance in North America?: Recent trends and alternative policies to promote bicycling, Transp. Res. A 33 (7–8) (1999) 625–654.

[4] John Pucher, Ralph Buehler, Why Canadians cycle more than Americans: a comparative analysis of bicycling trends and policies, Transp. Policy 13 (3) (2006) 265–279.

[5] YouLi Feng, ShanShan Wang, A forecast for bike rental demand based on random forests and multiple linear regression, in: 2017 IEEE/ACIS 16th International Conference on Computer and Information Science, ICIS, IEEE, 2017.

[6] Bo Wang, Inhi Kim, Short-term prediction for bike-sharing service using machine learning, Transp. Res. Proced. 34 (2018) 171–178.

[7] Xinhua Wu, et al., Station-level hourly bike demand prediction for dynamic repositioning in bike-sharing systems, in: Smart Transportation Systems 2019, Springer, Singapore, 2019, pp. 19–27.

[8] Zidong Yang, et al., Mobility modeling and prediction in bike-sharing systems, in: Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services, ACM, 2016.

[9] Lei Lin, Zhengbing He, Srinivas Peeta, Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach, Transp. Res. C 97 (2018) 258–276.

[10] Lihuan Zhang, et al., Data analysis and visualization in bike-sharing systems, in: Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services Companion, ACM, 2016.

[11] Yan Pan, et al., Predicting bike-sharing demand using recurrent neural networks, Proced. Comput. Sci. 147 (2019) 562–566.

[12] Dimitrios Tomaras, Ioannis Boutsis, Vana Kalogeraki, Modeling and predicting bike demand in large city situations, in: 2018 IEEE International Conference on Pervasive Computing and Communications, PerCom, IEEE, 2018.

[13] Feihu Huang, et al., A bimodal gaussian inhomogeneous Poisson algorithm for bike number prediction in a bike-sharing system, IEEE Trans. Intell. Transp. Syst. (2018).

[14] Zengwei Zheng, Yanzhen Zhou, Lin Sun, A multiple factor bike usage prediction model in bike-sharing system, in: International Conference on Green, Pervasive, and Cloud Computing, Springer, Cham, 2018.

[15] Maricica Nistor, André Dias, Bike distribution model for urban data applications, Int. J. Transp. Dev. Integr. 3 (1) (2019) 67–78.

[16] Byron Graham, Raymond Bond, Michael Quinn, Maurice Mulvenna, Using data mining to predict hospital admissions from the emergency department, IEEE Access 6 (2018) 10458–10469.

[17] Luis M. Candanedo, Véronique Feldheim, Dominique Deramaix, Data driven prediction models of energy use of appliances in a low-energy house, Energy Build. 140 (2017) 81–97.

[18] P.B Schiilkop, Chris Burgest, Vladimir Vapnik, Extracting support data for a given task, in: Proceedings of the 1st International Conference on Knowledge Discovery & Data Mining, 1995, pp. 252–257.

[19] B. Dong, C. Cao, S.E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy Build. 37 (5) (2005) 545–553.

[20] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794.

[21] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Statist. 29 (5) (2001) 1189–1232.

[22] V. Pashazadeh, F.R. Salmasi, B.N. Araabi, Data driven sensor and actuator fault detection and isolation in wind turbine using classifier fusion, Renew. Energy 116 (2018) 99–106.

[23] N. Sheibat-Othman, S. Othman, M. Benlahrache, P.F. Odgaard, Fault detection and isolation in wind turbines using support vector machines and observers, in: Proc. IEEE Amer. Control Conf., 2013, pp. 4459–4464.

[24] SEOUL OPEN DATA. URL: http://data.seoul.go.kr/.

[25] SOUTH KOREA PUBLIC HOLIDAYS. URL: publicholidays.go.kr.

[26] William R. Revelle, psych: Procedures for personality and psychological research, 2017.

[27] M.B. Kursa, W.R. Rudnicki, Feature selection with the Boruta package, J. Stat. Softw. 36 (11) (2010) 1–13.

[28] T.T. Nguyen, J.Z. Huang, T.T. Nguyen, Two-level quantile regression forests for bias correction in range prediction, Mach. Learn. 101 (1–3) (2015) 325–343.

[29] S. Strempel, M. Nendza, M. Scheringer, K. Hungerbühler, Using conditional inference trees and random forests to predict the bioaccumulation potential of organic chemicals, Environ. Toxicol. Chem. 32 (5) (2013) 1187–1195.

[30] M. Belgiu, L. Drăguţ, Random forest in remote sensing: A review of applications and future directions, ISPRS J. Photogramm. Remote Sens. 114 (2016) 24–31.

[31] C. Brown, Dummies: create dummy/indicator variables flexibly and efficiently. R package version 1.5. 4, 2011.

[32] Cheng Fan, Fu Xiao, Shengwei Wang, Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Appl. Energy 127 (2014) 1–10.

[33] M. Kuhn, Caret: Classification and Regression Training, Astrophysics Source Code Library, 2015.

[34] R. Analytics, S. Weston, doParallel: Foreach parallel adaptor for the parallel package. R package version, 2014, 1(8), p. 2014.

[35] M. Kuhn, K. Johnson, Applied Predictive Modeling, Vol. 26, Springer, New York, 2013.