Journal of Science and Transport Technology
Journal homepage: https://jstt.vn/index.php/en

Predicting Bike-Sharing Demand Using Random Forest

Thu-Tinh Thi Ngo[1], Hue Thi Pham[1], Juan Acosta[2], Sybil Derrible[2,*]
[1]University of Transport Technology, Hanoi 100000, Vietnam
[2]University of Illinois at Chicago, Chicago, Illinois, United States

Article info
Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2022.en.2.2.13-21

Corresponding author:
E-mail address:
derrible@uic.edu

Received: 24/01/2022
Revised: 09/05/2022
Accepted: 11/05/2022

Abstract: Being able to accurately predict bike-sharing demand is important for Intelligent Transport Systems and traveler information systems. These challenges have been addressed in a number of cities worldwide. This article uses Random Forest (RF) and k-fold cross-validation to predict the hourly count of rental bikes (cnt/h) in the city of Seoul (Korea) using information related to rental hour, temperature, humidity, wind speed, visibility, dewpoint, solar radiation, snowfall, and rainfall. The performance of the proposed RF model is evaluated using three statistical measurements: root mean squared error (RMSE), mean absolute error (MAE), and correlation coefficient (R). The results show that the RF model has high predictive accuracy with an RMSE of 210 cnt/h, an MAE of 121 cnt/h, and an R of 0.90. The performance of the RF model is also compared with a linear regression model and shows superior accuracy.

Keywords: Bike-sharing demand; travel demand forecasting; machine learning; Random Forest.

1. Introduction

The challenges of climate change, global automation, and resource depletion are affecting every nation on the planet, and they are becoming more and more serious, especially for transport systems [1]. In response, governments and authorities are constantly implementing measures to develop more sustainable and resilient transport systems, including clean fuel, electric vehicles, strict regulations of the demand for private vehicle ownership, and the development of efficient public transport systems [2]. Bike-sharing systems are one of the measures that have been adopted to address these challenges [3].

The principle of a bike-sharing system is straightforward. People pay a fee to rent a bike for a short time period. This system is convenient because users can comfortably use it to move around without owning a bicycle, providing health benefits while paying only a small amount of money. In addition, the use of bike-sharing services also brings significant benefits such as greenhouse emissions reduction, zero fuel consumption, congestion reduction, physical exercise (public health), and an increase awareness about the environment [4].

Early bike-sharing systems were invented around 1960. By 2022, they had primarily developed into three models [5]: (a) free bike system, (b) deposit bike rental system with private parking, and (c) bike-sharing system using location technology. The third one often goes along with larger deposits and requests to provide user information in order to overcome the

disadvantages of the previous two forms. Currently, bike-sharing systems are being used and developed worldwide, especially in Europe, America, and Asia. In Vietnam, for the past few years, the public sharing model has been piloted in several urban areas, tourist resorts, and major universities in the form of a bike-sharing system, and it has received many positive reviews from users. One of the most critical aspects that may assist investors in developing a successful implementation model that benefits the community is the hourly number of rental bikes required to operate the system. An additional issue for managers is accurately forecasting demand for bikes at any time and from any location to help with traffic planning in urban areas.

Today, thanks to advance in information technology and improvements in computer infrastructure, data-driven decision-making practices with the assistance of technology has become common. The use of data to estimate bike-sharing (or rental) demand has been adopted in a number of studies. For example, a bike-sharing dataset from India was used to predict the number of bikes rented per hour using decision tree (DT), Random Forest (RF), and Gradient Boosting (GB) [6]. Another study presented a prediction model to forecast user demands and efficient operations for rental bikes using tree-based machine learning models. This nonparametric approach uses a DT model to solve regression problems using the following continuous and categorical variables: holiday status, function day, week status, and the day of the week [7]. The prediction was carried out using Linear Regression (LR), k-nearest neighbor (KNN), GB, and RF. Four performance metrics—namely RMSE, coefficient of variance (CV), and mean and median absolute error—are used to determine model performance. Compared with other models, RF performed best [8].

Because bike sharing is a continuous operation, the primary goal of this study is to utilize RF to forecast the required hourly number of rental bikes (count of hourly rental bikes – cnt/h). This study employs RF thanks to its efficiency and robustness in solving regression and classification problems [9,10]. A comprehensive model evaluation was conducted using k-fold cross-validation, and three statistical measures were used to assess model performance: (a) RMSE, (b) MAE, and (c) Pearson correlation coefficient (R). Compared with previous studies, this work leverages predictive modelling to evaluate the bike-sharing demand in Seoul by using nine continuous input variables.

2. Database description

This study leverages a dataset collected and used in previously published works for bike demand prediction using eight weather parameters and hour information [11,12]. The dataset—available in the UCI Machine learning Repository [13]—includes 8,760 samples. The dataset is divided into two: 6,132 (70%) data points are used for training, and the remaining 2,628 (30%) data points are used for testing. The RF model is built using the following nine continuous variables: hour, temperature, humidity, wind speed, visibility, dewpoint, solar radiation, snowfall, and rainfall, indicated in Fig. 1 as X_1 to X_9, respectively. These inputs influence bike-sharing demand, the output (denoted as Y) of this study.

Fig. 1 shows the distribution histogram and correlations among the input and output parameters used in the study.

In general, most input variables in the database cover a wide range of values. The Pearson correlation coefficient (R) was calculated and highlighted with respect to each pair of variables [14]. Except for X2 and X6, which are directly correlated, the results revealed no strong correlation between the input and output parameters as seen with relatively low R values (i.e., R <0.54). The figure also suggests that all other variables are independent and capture different properties of Y.

As a last note, in order to minimize the errors generated during the simulation process, the dataset is scaled within the range of values [0,1] to limit errors generated by numerical simulations.

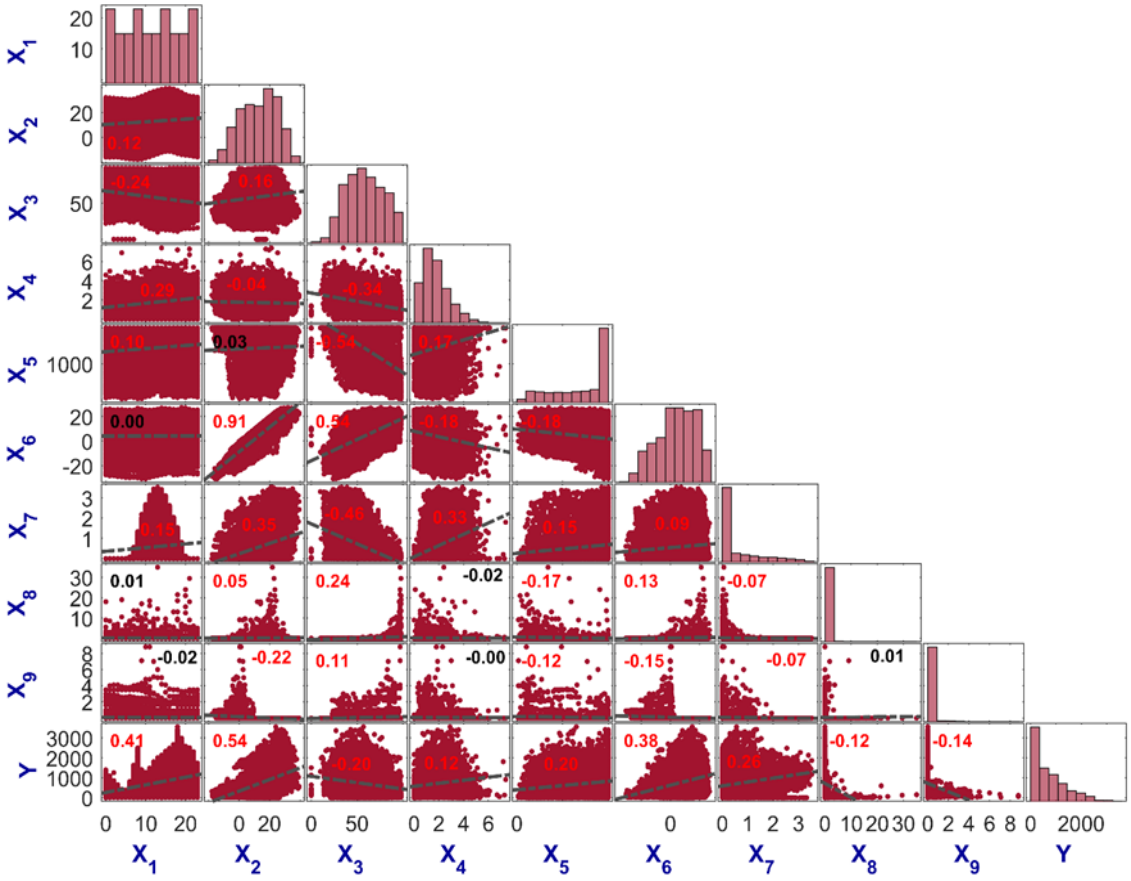


Fig. 1. The distribution chart and correlations between input and output parameters

3. Model description

3.1. Random Forest

RF is a supervised learning algorithm used to solve classification and regression problems. It was introduced and developed by Leo Breiman in 2001 [15]. RF is one of the algorithms built on the decision tree modeling technique. Each tree acts as a vote as the basis of the algorithm's decision-making. According to the resampling principle (Bootstrap method), each decision tree is generated based on a random training sample set generated from the original one with the same magnitude according to the resampling principle. Furthermore, each decision tree is based on the newly created sample set with the principle of using only a limited number of input variables at each split node. The final result is the average result obtained from all decision trees. Combining the results of many independent decision trees with low bias and high variance helps RF achieve relatively low bias and low variance.

Prediction trees receive specific numerical values in the regression problem instead of a class in classification problems [15]. In the design of regression trees, each tree is allowed to grow to the maximum depth of the training data without performing any reduction (branching). This is also a significant advantage of this algorithm because tree reduction is a significant factor affecting the model's performance [16]. Breiman [15] also argues that as the number of trees increases, the general error converges even when the tree is not reduced, and the treatment of model overfitting is done based on the law of large numbers. The number of variables (p) used at each node to create a decision tree and the number of decision trees (Q) used are two pre-selected parameters. The number of trees in the forests should be large enough to ensure that all attributes are used several times. Generally, the number of trees used for classification or regression problems varies from a few to 1000 trees, depending on the complexity of the relationship between the input and output variables. The optimal value is determined on a case-by-case basis.

In recent years, RF has been popularly used

thanks to its advantages over other algorithms; namely, its ability to evaluate the internal error, to evaluate the importance of the input variable, and to handle variables with low correlation. As a result, RF has been widely applied in a variety of areas, including water demand [17], banking to forecast client reaction [18], stock market price direction [19], e-commerce [20], and science technology [21,22].

RF includes the following main steps: (i) set of trained regression trees using the training set; (ii) calculate the mean of the yield of individual regressors; (iii) cross-validate prediction data using a validator.

3.2. k-fold cross-validation

k-fold cross-validation (CV) is a basic form of cross-validation (Fig. 2). It is widely used to quantify the prediction performance of machine learning models, especially for the selection of hyper-parameters. The data is first partitioned into k equally sized segments or folds. Subsequently, *k* iterations of training and validation are performed such that, within each iteration, a different fold of the data is held out for validation, whereas the remaining *k-1* folds are used for training. In classification problems, data are commonly stratified prior to being split into *k* folds. For regression problems, 5-fold or 10-fold cross-validation choices are often selected [23,24].

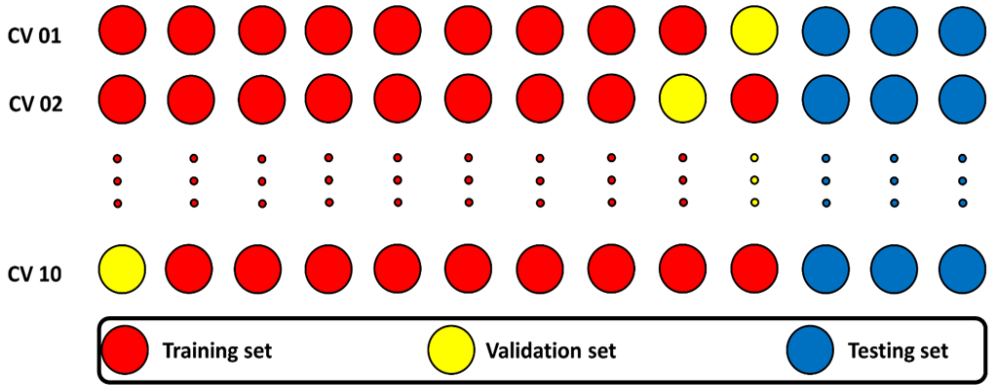


Fig. 2 . Cross-validation technique with 10-fold used in this study

In this work, the original data are split into two sets: training and testing. The training set is randomly divided into k parts; then, the model is trained k times, each time with 1 part as validation data and k-1 parts as training data. Meanwhile, the testing set is set aside because to be used to evaluate the model after the training phase to see how the model handles new data. It is kept separate and reserved only for the final evaluation step to check the performance of the model when encountering completely unseen data.

After the model is evaluated and if the results (e.g., the average performance) are acceptable, one of the following two ways is carried out to create the final model for further use and investigation.

The first one: the best model is taken in the training process. The advantage of this approach is that there is no need to retrain the model, but it might not be able to cover all the range of input data (the training data is fixed), so the model might not work well with new data.

The second one: train the model one more time using the full training data (not separated into folds), then it is used to predict the test set to measure the model's generalization ability.

3.3. Performance assessment

The performance of the prediction models is evaluated using three statistical indices, namely RMSE, MAE, and R. The R value ranges from -1 to 1, and a higher R value (i.e., an absolute value closer to 1) suggests a higher performance. For RMSE and MAE, a lower value (i.e., lower error) suggests a higher performance. Mathematically, RMSE, MAE, and R are defined as

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N}(y_0 - y_p)^2 / N} \quad (1)$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_0 - y_p\right| \quad (2)$$

$$\mathrm{R} = \sqrt{1 - \frac{\sum_{i=1}^{N}(y_0 - y_p)^2}{\sum_{i=1}^{N}(y_0 - \overline{y_i})^2}} \quad (3)$$

where N is the number of input data, $\bar{y}$ is the mean value of the outputs, and y_0 and y_p express the actual and modeled/predicted values, respectively.

4. Results and discussion

4.1. RF prediction results

This section introduces the training and evaluation of an RF model using k-fold cross-validation (CV), which includes two steps. For the first step, the RF model is trained using the training dataset (70% of the samples; 6,132 data points) with a 10-fold CV. This means that the training dataset was divided into ten parts, and the simulation was repeated ten times, as previously mentioned. Notably, this first step is used for hyper-parameter tuning. Namely, a range of hyper-parameter values are tried, and the ones that achieve the highest prediction accuracy are kept (highest R and lowest RMSE, MAE values). The final hyper-parameter values are as follows: number of trees: 200; max depth: unlimited; minimum samples split: 2; minimum samples leaf: 1.

For the second step, the model is retrained on the full training dataset and tested on the testing dataset. It is worth noting that the testing dataset (which contains the remaining 30% of data; 2,628 data points) is only used to evaluate the model's predictive ability. It is not used for model training and hyper-parameter section. Using the hyper-parameters from step 1, the second step is repeated ten times by randomly shuffling the training and testing sets from the full dataset to ensure that the performance accuracies are stable. The main goal here is to ensure the generalizability of the prediction results.

The evaluation results for the ten runs in step 2 are shown in Figure 3. It can be seen that the proposed RF model performs reasonably well for all ten runs. Further, the performances of the models do not fluctuate significantly, which mean they all capture the same relationships from the dataset.

On the training sets, R values oscillate around 0.97, and RSME and MAE values oscillate around 155 cnt/h and 95 cnt/h, respectively. These results demonstrate that the trained RF model performs well and can be selected for further exploration on unseen data. Similarly, the model performance on the testing sets are also stable. Compared to the performance on the training sets, R values decrease by about 10%, and RMSE and MAE values increase roughly by a factor of 2, which is reasonable. The average prediction results of the RF model remain high (i.e., $R \approx 0.88$, RMSE $\approx$ 310 cnt/h, MAE $\approx$ 185 cnt/h).

We note that randomly shuffling the training data can have a certain influence on the prediction performance of the RF model. For example, the sixth run achieved the best performance with respect to the R values, whereas the fifth run had the best performance in terms of RSME, and the third run in terms of MAE. Yet, the difference between the various simulations is negligible, and the RF model performs well in the overall analysis. As a result, the selected RF model may be used to predict the bike-sharing demand with such a high level of accuracy.

In this section, the typical predictive results for the problem are presented. The regression analysis for the training dataset (Fig. 4a), the testing dataset (Fig. 4b), and all data (Fig. 4c) are shown. In each figure, the diagonal (black dashed line) represents an ideal correlation for the problem (R=1). In addition, the RF model's regression line is also shown by the violet line (which deviates from the ideal regression line as is usually the case). For each case, the predictors are calculated and expressed in each figure. Precisely, R = 0.97, RMSE = 158 cnt/h, and MAE = 94 cnt/h for the training data, R = 0.89, RMSE = 298 cnt/h, and

MAE = 186 cnt/h for the testing data, and R = 0.90, RMSE = 210 cnt/h, and MAE = 121 cnt/h for the entire data. These results show a high prediction performance of the proposed RF model.

Finally, Figs. 5a and b describe the RF model's distribution and cumulative distribution of error for the training and testing parts, respectively. As can be seen, 90% of the error is in the range of -300 to 300 cnt/h, and 65% is within the ± 70 cnt/h for the training part. Regarding the testing set, 90% of the error is in the range of -500 to 500 cnt/h, and 65% is within the ±100 cnt/h. This is also confirmed by the higher accuracy of the training set than the testing one.

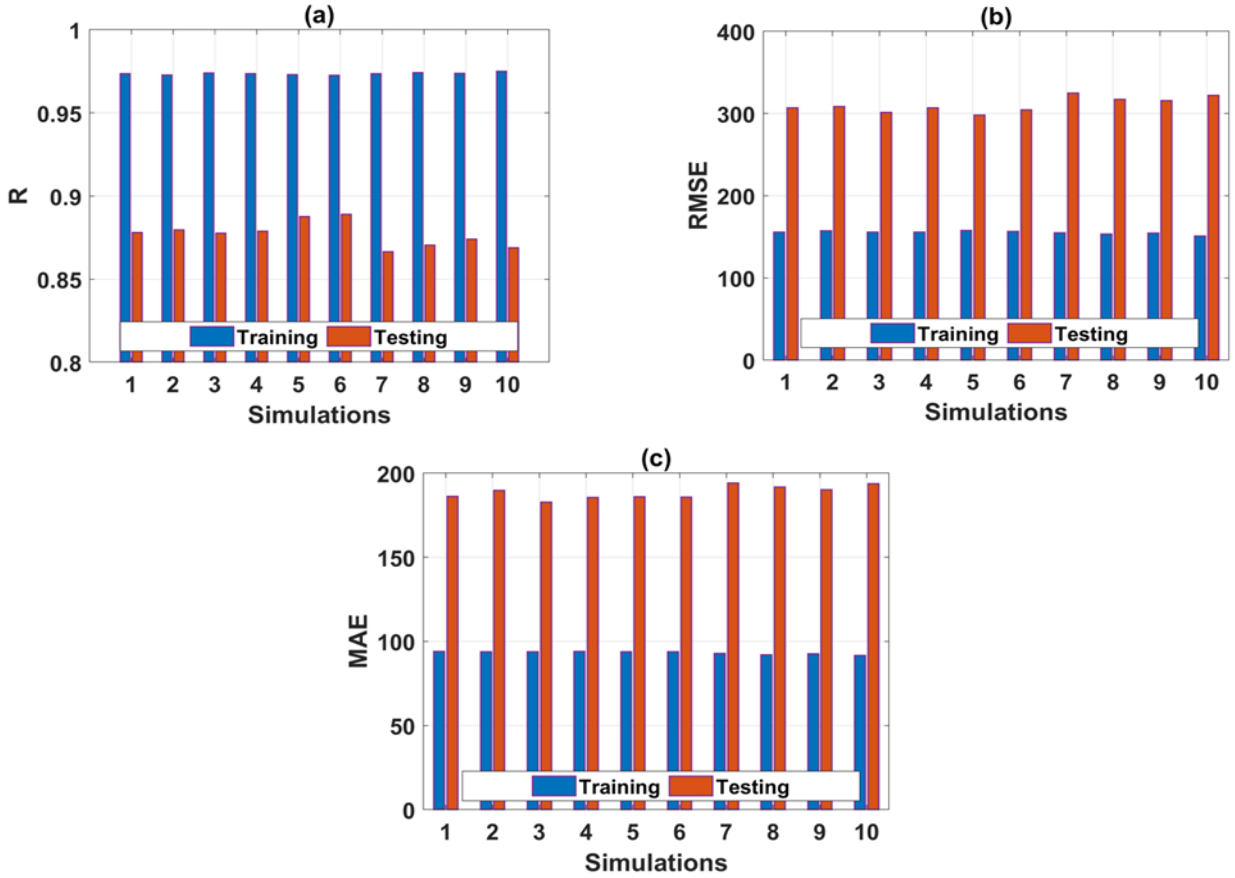
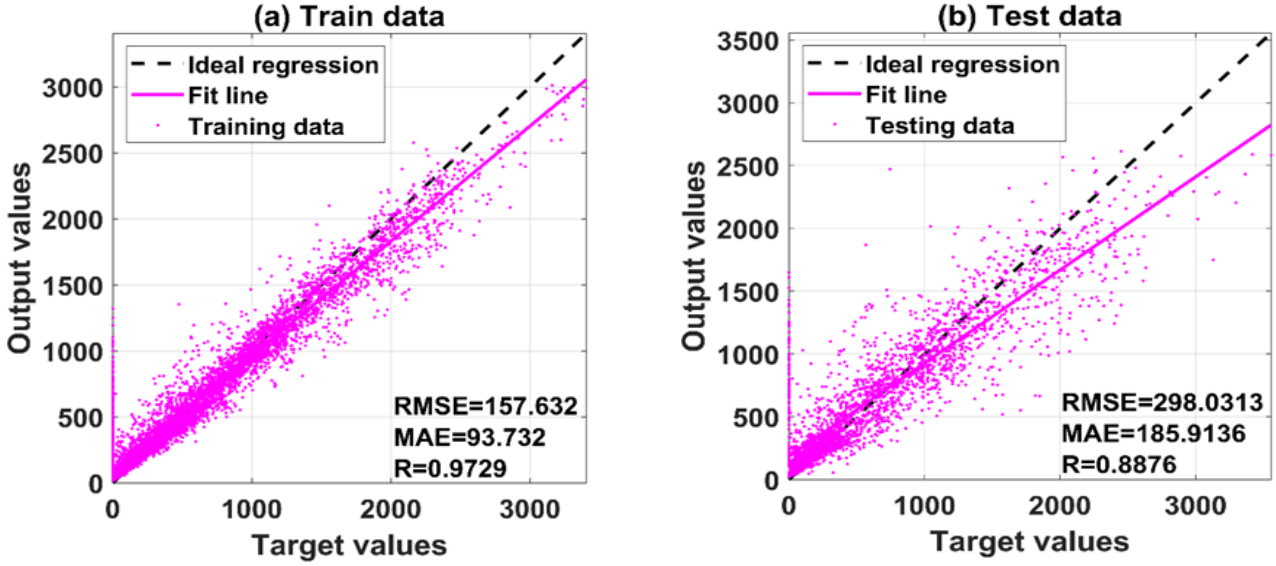


Fig. 3. Training and testing results of RF model over 10: (a) R, (b) RMSE, (c) MAE



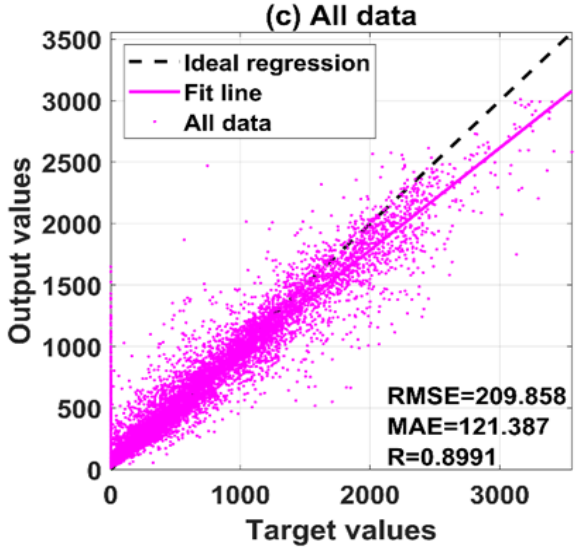


Fig. 4. Regression analysis of the RF model between target and output values with respect to three datasets: (a) training data; (b) testing data; and (c) all data

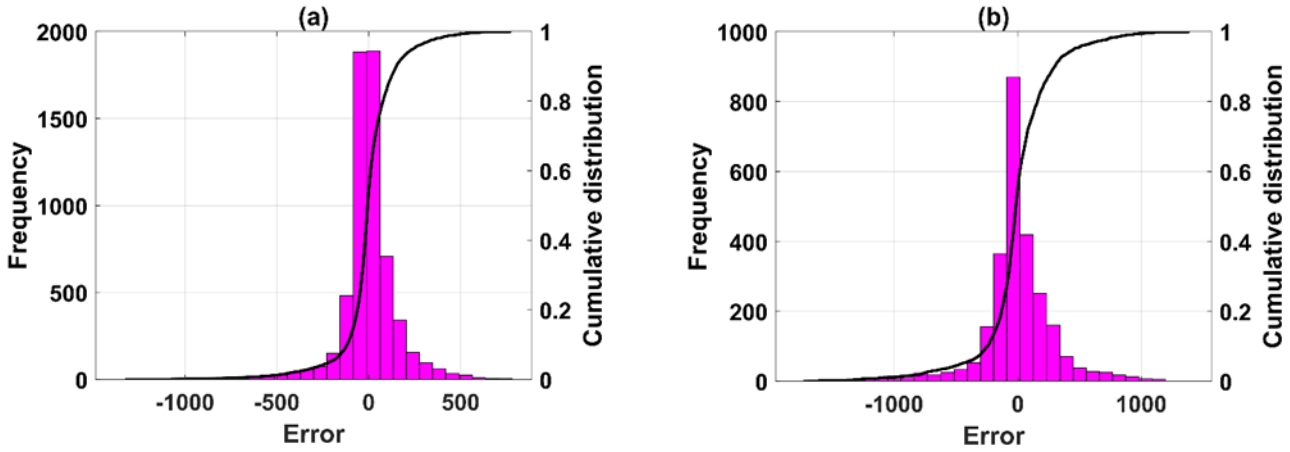


Fig. 5. Error analysis of the RF model: (a) Training set; and (b) Testing set

4.2. Comparison of RF and LR

In this section, the performance of the RF model in predicting bike-sharing demand is compared with a multivariable Linear Regression (LR) model. LR models are widely used thanks to their simple implementation process. As is the case here, they are often used as a benchmark to highlight the performance of the proposed machine learning models. In order to compare the performance of the two models, the training and testing phases of LR model are conducted on the same datasets as used during the development of RF model. The final form of LR model is given in Eq. 4.

$$Y = 27.31X_1 + 26.58X_2 - 8.81X_3 + 6.92X_4 + 0.02X_5 + 5.41X_6 - 79.34X_7 - 58.81X_8 + 21.08X_9 + 548.85 \quad (4)$$

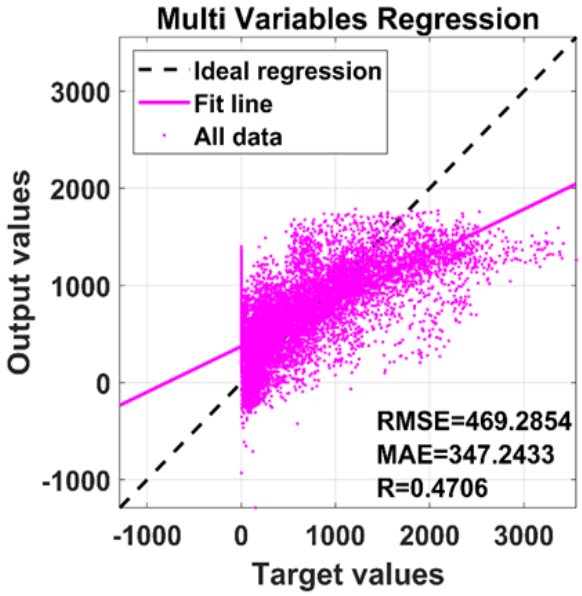


Fig. 6. Regression analysis of the LR model

Fig. 6 illustrates the results in a regression plot of the LR model between the actual and predicted values. The figure shows that the LR model does not perform as well as the RF model. In fact, the LR model only achieves an R value of 0.47, which is about 50% lower than the RF model. Moreover, LR's RMSE and MAE values are about twice as large as those derived from the RF model. Overall, the RF model is superior for predicting bike-sharing demand from the statistical analysis and prediction errors calculated.

5. Conclusion

This study proposes an RF model to predict bike-sharing demand with the use of nine continuous input variables. For this purpose, a dataset for the city of Seoul collected from the UCI Machine Learning Repository was utilized for model training and testing purposes. Nine input parameters are used: rental hour, temperature, humidity, wind speed, visibility, dewpoint, solar radiation, snowfall, and rainfall. Three criteria—RMSE, MAE, and R—were used to evaluate the performance of the proposed RF model, coupled with k-fold cross-validation for hyper-parameter tuning. The results show that the proposed model achieves high accuracy in predicting bike-sharing demand, with performance evaluation measures of RMSE = 210 cnt/h, MAE = 121 cnt/h, and R = 0.90. The RF model also outperformed a conventional LR model and is therefore preferable to use to predict bike-sharing demands here.

References

[1] S. Derrible. (2019). Urban engineering for sustainability. MIT Press.

[2] T.M.T. Truong, H.-B. Ly, D. Lee, B.T. Pham, S. Derrible. (2021). Analyzing travel behavior in Hanoi using Support Vector Machine. *Transportation Planning and Technology*, 44(8), 843-859.

[3] A. Rawat. (2020). Demand prediction in a bike-sharing system using machine learning techniques. *PhD Thesis, Dublin. National College of Ireland*.

[4] V.E. Sathishkumar, J. Park, Y. Cho. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, 153, 353-366.

[5] P. DeMaio. (2009). Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation*, 12(4), 41-56.

[6] A. Patil, K. Musale, B.P. Rao. (2015). Bike share demand prediction using RandomForests. *IJISET International Journal of Innovative Science, Engineering & Technology*, 2(4), 1218-1223.

[7] Sathishkumar V.E., Y. Cho. (2020). Season wise bike sharing demand analysis using random forest algorithm. *Computational Intelligence*.

[8] V.E. Sathishkumar, J. Park, Y. Cho. (2020). Seoul bike trip duration prediction using data mining techniques. *IET Intelligent Transport Systems*, 14(11), 1465-1474.

[9] A. Hapfelmeier, K. Ulm. (2013). A new variable selection approach using random forests. *Computational Statistics & Data Analysis*, 60, 50-69.

[10] A. Liaw, M. Wiener. (2002). Classification and regression by randomForest. The R Journal, 2/3, 18-22.

[11] S. VE, J. Park, Y. Cho. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. Computer Communications, 153, 353-366.

[12] S. VE, Y. Cho. (2020). A rule-based model for Seoul Bike sharing demand prediction using weather data. *European Journal of Remote Sensing*, 53, 166-183.

[13] Seoul Bike Sharing Demand, UCI Machine Learning Repository. (2020). https://archive-beta.ics.uci.edu/ml/datasets/seoul+bike+sharing+demand.

[14] H. Zhou, Z. Deng, Y. Xia, M. Fu. (2016). A new sampling method in particle filter based on Pearson correlation coefficient, *Neurocomputing*, 216, 208-215.

[15] L. Breiman. (2001). Random forests. *Machine Learning*, 45, 5-32.

[16] M. Pal, P.M. Mather. (2003). An assessment

of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment*, 86(4), 554-565.

[17] D. Lee, S. Derrible. (2020). Predicting residential water demand with machine-based statistical learning. *Journal of Water Resources Planning and Management*, 146(1), 04019067.

[18] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43(6), 1947-1958.

[19] J. Patel, S. Shah, P. Thakkar, K. Kotecha. (2015). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, 42(4), 2162-2172.

[20] A.M. Prasad, L.R. Iverson, A. Liaw. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9, 181-199.

[21] T.A. Pham, H.-B. Ly, V.Q. Tran, L.V. Giap, H.-L.T. Vu, H.-A.T. Duong. (2020). Prediction of pile axial bearing capacity using artificial neural network and random forest. *Applied Sciences*, 10(5), 1871.

[22] H.-B. Ly, T.-A. Nguyen, B.T. Pham. (2021). Estimation of Soil Cohesion Using Machine Learning Method: A Random Forest Approach. *Advances in Civil Engineering*, 2021(1), 1-14.

[23] J. Franklin. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27, 83-85.

[24] T. Hastie, R. Tibshirani, J. Friedman. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, *Springer-Verlag, New York*. https://www.springer.com/la/book/9780387848570 (accessed March 21, 2019).