American Journal of Information Science and Technology
2023; 7(2): 62-69
http://www.sciencepublishinggroup.com/j/ajist
doi: 10.11648/j.ajist.20230702.13
ISSN: 2640-057X (Print); ISSN: 2640-0588 (Online)

A Stacking-Based Ensemble Model for Prediction of Metropolitan Bike Sharing Demand

Xinxue Lin[1], Chang Lu[2]

[1]School of Resource and Environmental Sciences, Wuhan University, Wuhan, China

[2]School of Urban Design, Wuhan University, Wuhan, China

Email address:
xinxuelin@whu.edu.cn (Xinxue Lin), luchang@whu.edu.cn (Chang Lu)

To cite this article:
Xinxue Lin, Chang Lu. A Stacking-Based Ensemble Model for Prediction of Metropolitan Bike Sharing Demand. *American Journal of Information Science and Technology*. Vol. 7, No. 2, 2023, pp. 62-69. doi: 10.11648/j.ajist.20230702.13

Received: March 14, 2023; **Accepted**: April 14, 2023; **Published**: April 20, 2023

Abstract: Due to the climate crisis and the improvement of public transportation networks, countries around the world are strongly advocating the low-carbon traveling mode. Shared bike as a new business model has a positive impact on the urban environment and transportation. The ability to estimate the hourly demand for bike sharing with high accuracy is essential for metropolis to offer stable bike rental services. Presently, data mining and predictive analysis technology can be utilized to realize the forecast of the hourly demand of shared bicycles. Data used in this article include the Seoul bike rented count dataset and weather information. This paper discusses various machine learning models for rental bike demand prediction, including Linear Regression, Ridge Regression, Lasso Regression, K-Nearest Neighbor, Random Forest, Decision Tree Regression, Support Vectors Machine, and Gradient Boosting Decision Tree. Different parameter tuning methods have been applied to improve the performance of basic predictive models. In addition, the redundant and irrelevant features have been removed to improve the performance of each basic model. After evaluating the individual basic predictors, several competent basic predictors are selected to compose a stacking-based ensemble model. Experimental results show that the stacking-based ensemble model outperforms the basic predictive models in all indicators.

Keywords: Data Mining, Predictive Analytics, Regression Models, Ensemble Models, Bike Sharing Demand

1. Introduction

At present, urban air pollution is becoming increasingly serious, and even in metropolis with comprehensive transportation network coverage such as subways and buses, there is still a problem of "the last mile from the platform to the destination". In this context, the emergence of shared bicycles effectively improved environmental and traffic problems. As a growing and effective supplement to the urban transportation network, the free and flexible rental mode of shared bicycles can be combined with the existing public transportation system to form a "point-to-point" transportation service, improve the accessibility of the existing public transportation system, and is of great significance for alleviating urban traffic congestion.

However, after a period of explosive growth and fierce competition between various sharing bike service providers, many problems have been exposed, which ultimately revolve around the same core, i.e., the imbalance between supply and demand. The first situation is that supply exceeds demand, which will lead to the overcrowding of shared bikes on the road, and produce a waste of space and resources. Another situation is that demand exceeds supply, which will make the needs of users unsatisfied and reduce the ability of shared bikes to be a complement to urban transport.

As a non-motorized mode exposed to the traffic environment, the demand for shared bikes is significantly influenced by the weather. However, there is still a lack of research on the demand prediction of new transport modes such as shared bicycles, particularly on the impact of weather on the demand of shared bicycles.

Therefore, short-term demand forecasting based on big data can help bike sharing service providers to distribute and schedule shared bicycles more reasonably. It better ensures sufficient travelling resources and makes low-carbon shared bicycles better serve urban operations.

In light of the aforementioned, this study employs weather data

and the data of shared bike rentals in Seoul, Korea as the model input. To improve the quality of the dataset and the final predictability, some data preprocessing methods are employed, such as feature selection, format transformation, and standardization. This study proposes an ensemble model based on different predictors in combination with the current mainstream data prediction algorithms, including Linear Regression (LR), Ridge Regression (RR), Lasso Regression (Lasso), K-Nearest Neighbor (KNN), Random Forest (RF), Decision Tree (DT), Support Vectors Machine (SVM), and Gradient Boosting Decision Tree (GBDT). Grid search and gradient boosting were employed for selecting parameters and improving model accuracy.

The sections of the paper have the following structure. Section 2 reviews previous studies related to public bike-sharing systems and predictive algorithms and methods. The data preprocessing methodology and the underlying predictors used are explored in Section 3. Section 4 demonstrates the indicators and results of the experiment. Conclusions are illustrated in section 5.

2. Related Work

2.1. Application of Data Mining

Data mining, as the popular method for large-scale data analysis, is the process of identifying significant, previously unnoticed linkages, trends, and patterns. Data mining was developed thanks to the theories and techniques from various fields such as artificial intelligence, database technology, pattern recognition, machine learning, statistics, and data visualization. Data mining can not only be used to characterize the existing data through descriptive analysis, but also to predict the future based on known data through predictive analysis.

Data mining methods have been applied in many fields. In terms of finance, Ngai et al. found that six types of data mining technologies (classification, regression, clustering, prediction, outlier detection, and visualization) were involved across all four areas of financial fraud (bank fraud, insurance fraud, securities and commodities fraud, and other related financial fraud) [8]. Lessmann et al. used classification and regression algorithms such as SVM and neural-network models to predict the credit score and bankruptcy of the industries [7].

Data mining has also been widely used in the field of ecological environment. Prasad et al. used Classification and Regression Trees (CART), DT, RF and other methods to predict tree species distribution in the eastern United States [10]. Bui et al. compared the efficacy of SVM, Artificial Neural Networks (ANN), Kernel Logistic Regression (KLR), and Logistic Model Tree (LMT) in the spatial prediction models for shallow landslide hazards [2]. To model and predict the spatiotemporal variation of PM2.5 concentrations, Qi et al. proposed a hybrid model based on deep learning methods that integrate Graph Convolutional Networks and Long Short-Term Memory Networks (GC-LSTM) [11].

Agriculture, transportation, medicine and many other fields also benefited from data mining techniques. Ruß et al. predicted wheat yield with ANN [12]. Wu et al. employed a classification-based data mining method to predict vehicle accident [18]. Data mining methods such as SVM and LMR were also employed by Komi et al. to predict diabetes [6]. It is evident that data mining has been widely applied and well developed in all aspects of our lives.

2.2. Application of Data Mining in Bike-Sharing

Because bike-sharing is a relatively new business model in recent years, data mining has fewer applications in this field.

Sun et al. adopted spatial autoregressive models to analyze the bike sharing usage, which collected GPS data from Singapore`s largest bike sharing services providers and took many factors into consideration such as surrounding environment, access to public and weather conditions [15]. Nugumanova et al. provided an analysis of bike-sharing stations using the Chi-square test [9].

Numerous issues, such as site analysis, vehicle scheduling, and personas of users using shared bikes, have been included in the studies on bike sharing [4]. The methodology of published studies ranged from the geographic spatiotemporal analysis to statistical analysis in data mining [5].

In developing cities, the demand for shared bicycle scheduling is considerably greater. Since riding bikes is an open-air mode of transportation, weather has a significant impact on demand, and considered as the influencing factors too.

Sathishkumar et al. found the Gradient Boosting Machine (GBM) gave the best performance with the highest R^2 value of 0.96 in training set and 0.92 in testing set [14]. Sathishkumar & Cho proposed a rule-based regression to predict bike-sharing demand, showing that the optimal model could explain about 95% and 89% of the variance in the training and testing sets respectively [13].

In recent years, ensemble learning for data mining has been developed greatly. Ensemble models provide an effective and robust path to model the process of predictive analytics. However, little research has been done on ensemble models in the topic of shared bicycles. This study will propose a feasible predictor using ensemble models and find the relationship between weather and bike sharing demand in metropolis.

3. Methodology

3.1. Data Preprocessing and Exploratory Analysis

The data analyzed in this study is split into two categories. One is the number of rented bicycles per hour from 2017-12-1 to 2018-11-30. The other is the weather and holiday data collected from the government website. After combining these two datasets, the raw data consists of 8760 instances and 14 characteristics. The whole dataset can be retrieved from the Seoul open data website[1] and the South Korea public holidays URL[2]

The raw data is described in Table 1, as shown below.

1 Available at http://data.seoul.go.kr/

2 Available at https://publicholidays.co.kr

Table 1. *Data description.*

Features	Type	Description
Date	Object	Record dates
Rented bike count	Continuous	The count of bikes rented at each hour
Hour	Continuous	Hour of the day
Temperature	Continuous	Celsius temperature (°C)
Humidity	Continuous	A percentage of air humidity (%)
Wind speed	Continuous	The wind speed at each hour (m/s)
Visibility	Continuous	Perceptible by the eye (m)
Dew point temperature	Continuous	Celsius temperature of dew point (°C)
Solar radiation	Continuous	Solar radiation amount per m^2 (MJ/m^2)
Rainfall	Continuous	Rain height (mm)
Snowfall	Continuous	Snow height (cm)
Holiday	Categorical	Holiday/Non-holiday
Functioning day	Categorical	Nonfunctional/Functional
Season	Categorical	Spring/summer/autumn/winter

The raw data contains ten numeric and four non-numeric features. To improve standardization of the data set, it must undergo preprocessing, which is described below.

(1) Viewing basic information and detecting outliers

It`s necessary to perform some basic information retrieval when the raw data was organized. Upon inspection, the dataset has no missing values. However, for each feature, the presence of outliers needs to be detected. The detecting techniques used in this study involve statistical description and data visualization. There are essentially no outliers if the data's general distribution and attributes fall within the typical range.

(2) Data exploration and visualization

Data exploration analysis can produce a series of visualizations to form data insights.

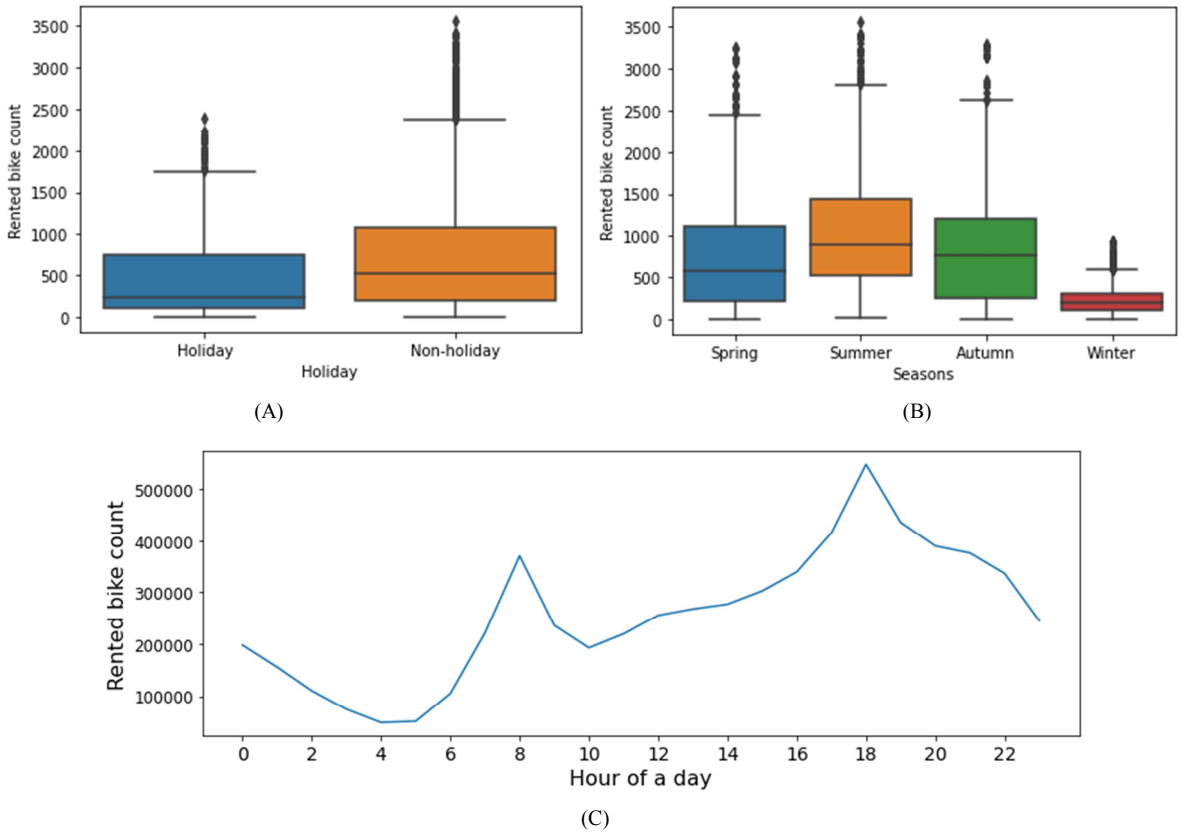
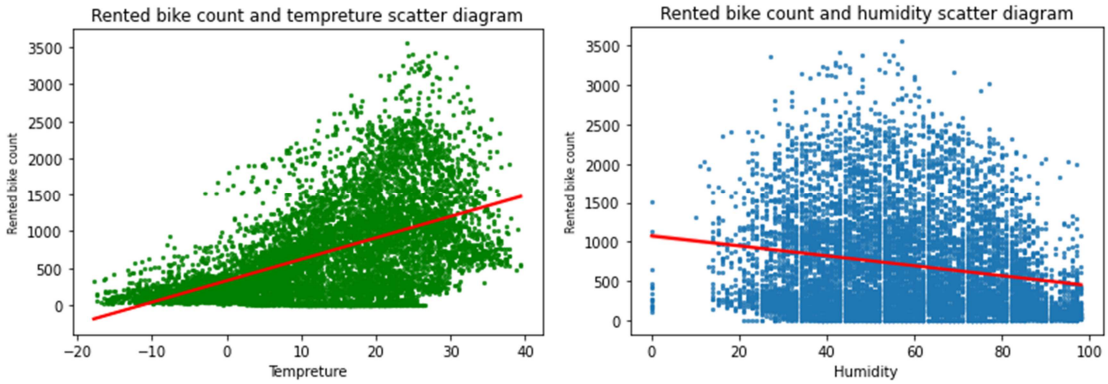


Figure 1. *The association between period and bicycle rentals.*

Figure 1 exhibits the association between period (i.e., seasons, holidays, and hours) and bicycle rentals. As shown in Figure 1 (A), more people rent bicycles during non-holiday than during holiday, which suggests that holiday is a significant factor for demand, possibly because more people utilize shared bicycles for a tour. The use of shared bikes is largest in the summer, followed by comparable usage in the spring and the autumn, and is lowest in the winter, as shown in Figure 1 (B). The change in bike-sharing demand with the seasons reflects to some extent the weather conditions, in particular the effect of temperature on the use of shared bikes. Figure 1 (C) illustrates how the demand for shared bikes varies throughout the day.

There are peaks in bike use around 8 a.m. and 6 p.m., which may be tied to peak commuter and work hours.



(A) (B)

Figure 2. *Scatter diagram of rented bike count and temperature & humidity.*

As shown in Figure 2 (A) and 2 (B), the data exploration through the scatter diagrams discovered that the number of shared bike users is correlated to temperature and humidity respectively.

(3) Evaluating correlations between features

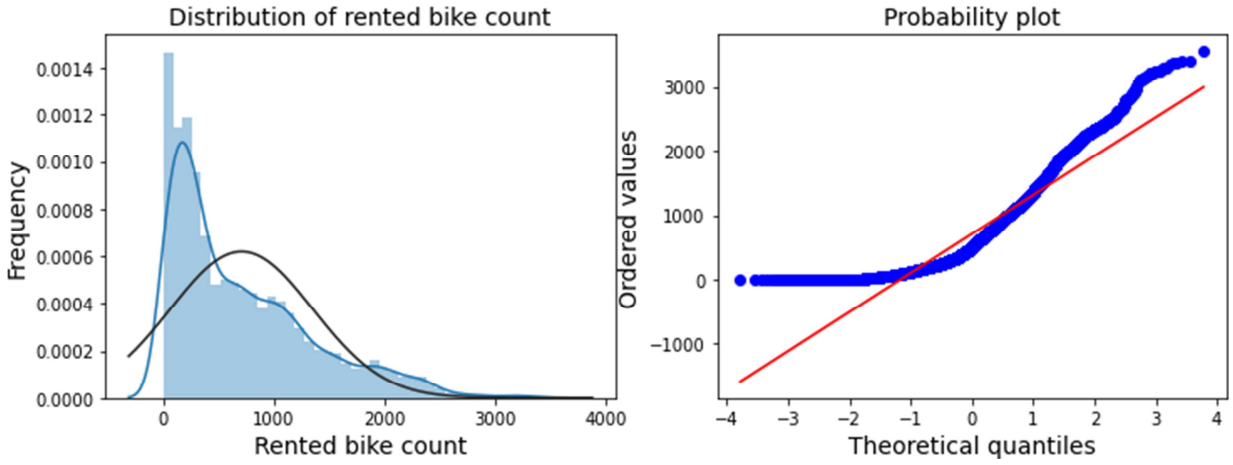
Considering that the redundant and irrelevant features affect the predictive accuracy and complicate the modeling process, they need to be removed through correlation analysis among all features. For independent variables with extra high correlation, only one is chosen as the main factor. The specific feature selection results can refer to the correlation heat map between features, as shown in Figure 3. It is simple to infer that the dew point temperature and temperature are strongly correlated. Therefore, only temperature is left as a salient feature, while dew point temperature is removed. Similarly, solar radiation is related to both temperature and humidity, so it should be excluded. Wind speed is also abandoned because it has very weak correlation with rented bike count (i.e., target variable) and also correlates with humidity and hour relatively. As a result, the number of features is simplified to 11, which theoretically brings higher predictive accuracy to the model performance.

Figure 3. *Correlation heat map.*

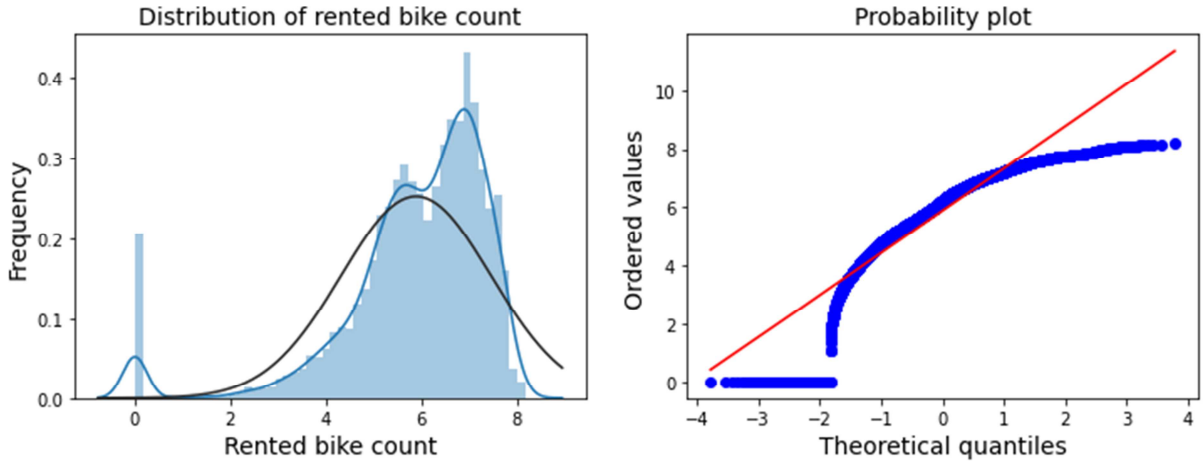
(4) Logarithmic modification and normalization to features

As the target variable, rented bike count should conform to normal distribution to facilitate the appropriate modeling using linear regression. Figure 4 shows the distribution and Q-Q chart of rented bike count both before and after logarithmic modification. Figure 4 (A) finds that the original distribution of rent bike count is not normal and is right skewed. Hence, the logarithmic modification is conducted to it so as to make it regularly distributed. The transformed result is shown in Figure 4 (B).

The final preprocessing step is to use the standard scaler method to normalize the independent features, so that they can have a fixed mean and standard deviation.



(A)



(B)

Figure 4. *The distribution and Q-Q chart of rented bike count.*

3.2. Modeling

The process of modeling is shown in Figure 5.

(1) Individual models training

Before predictive modeling, the data are divided into training, and testing sets with the ratio of 80% and 20%. The testing set is used to assess the model's prediction performance while the training set is used to optimize the model's parameters. As the basic predictive models, three linear regressors (LR, RR and Lasso) and five non-linear regressors (KNN, DT, RF, SVM, and GBDT) are employed.

This study adopts grid search as the main parameter tuning method. All the collocations of the parameter permutations are calculated and modeled for each parameter group. The parameter values that produce the minimum error are selected. Take KNN model as an example. The parameters to consider are {'n_neighbors', 'weights', and 'p'}. The grid search method finds the best group of parameters as {3, 'distance', and 3}. Figure 6 shows the R^2scores of KNN model with different value of 'n_neighbors'.

(2) Stacking-based ensemble model

Stacking-based ensemble model (STACKING) uses the ensemble learning method to stack various basic regressors into the ensemble model via a meta-regressor. In the traditional stacking procedure, a list of basic regressors will fit to the same training set in the first layer and their outputs are used as the input of the second-layer regressor (i.e., meta-regressor). However, this tends to produce overfitting problems.

To address the above issue, a cross-validation method is employed. For each regressor, the training set is divided into *k* parts. In each round, *k*-1 parts are used to train the regressor, and the remaining 1 part will be used to validate the regressor. After *k* rounds, each regressor gets the prediction result on the training set. The results of first-layer regressors are then stacked, whose outputs are used as the input data of the second-layer regressor (i.e., meta-regressor). On the testing set, the average to the *k* rounds of prediction results will be used as the final prediction result. Therefore, the multiple basic regressors have been integrated with the meta-regressor, which enhances the model predictability and decreases the model variance over a single regression model [1, 17].

In this study, three competent basic predictors (i.e., DT, GBDT, and RF) with higher predictive accuracy are selected as basic regressors of STACKING after individual fitting. The meta-regressor RR is selected from the linear regressors due to its highest predictive accuracy compared to others.

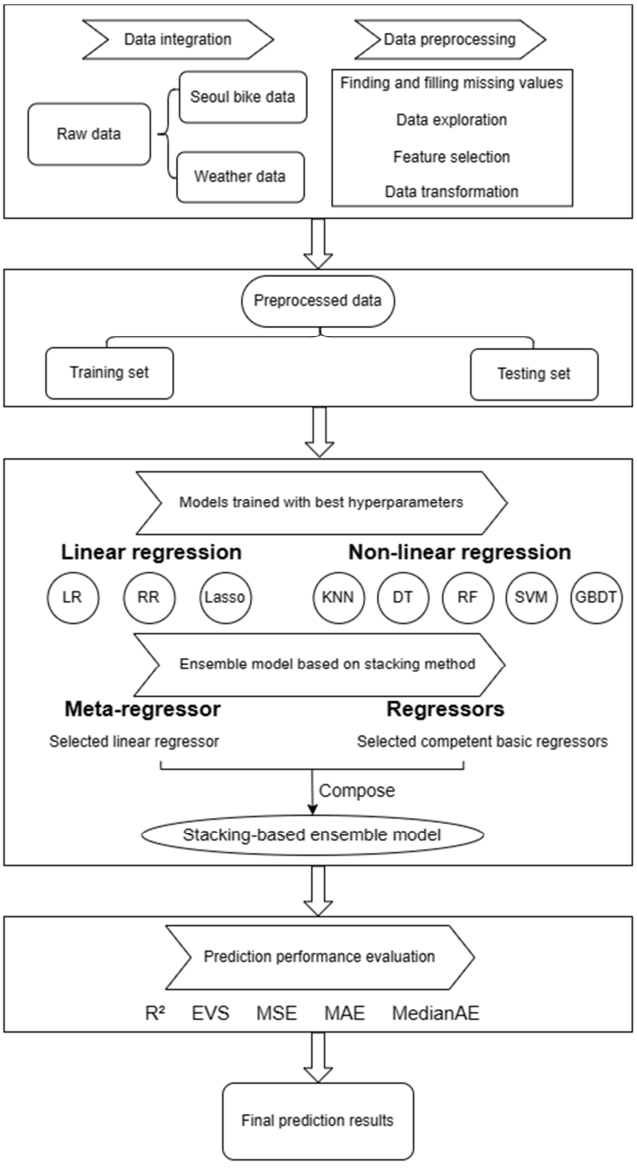


Figure 5. *Flow chart of modeling process.*

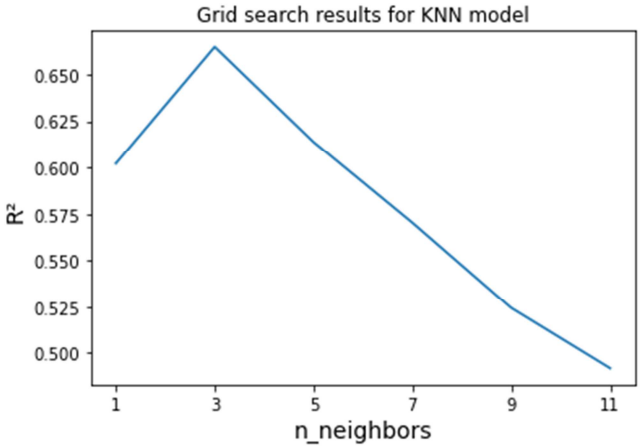


Figure 6. *The R²scores of different 'n_neighbors'.*

4. Experiments

4.1. Evaluation Indices

After training each model and optimizing it with the best parameters, the model performance was assessed using some evaluation indices [3, 16]. The specific meaning of each indicator is illustrated as follows.

Several statistical notions are used. In the following equations, y_i stands for the true observation at the *i*th sample point, $\bar{y}$ for the average of the observation, and $\widehat{y_i}$ for the prediction at the *i*th sample point.

(1) Coefficient of determination

R^2 is the coefficient of determination, reflecting the goodness-to-fit, which ranges from 0 to 1. The higher value of R^2 means the better performance in predicting.

The formula of R^2 is given in equation (1).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

(2) Explained variance score (EVS)

EVS is used to measure the degree to which the predictive model explains the variation in the dataset. If the value is 1, the model is perfect. The smaller the score, the worse the model performs.

The specific value is computed using equation (2).

$$EVS = explained_{variance(y_i, \widehat{y_i})} = 1 - \frac{Var\{y_i - \widehat{y_i}\}}{Var\{y_i\}} \quad (2)$$

(3) Mean squared error (MSE)

MSE calculates the mean of the error square between the fitted data and the corresponding sample points of original data. The smaller the value, the better the model performs. MSE can be computed using equation (3).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \quad (3)$$

(4) Mean absolute error (MAE)

MAE also reflects the closeness between predicted results and the true dataset. It is evaluated by the mean distance between the predicted and actual values as shown in equation

(4).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (4)$$

(5) Median absolute error (MedianAE)

MedianAE metric has similar principles to MSE and MAE, assessing the median of the absolute values of predicted results minus the actual values. The formula is calculated in equation (5).

$$MedianAE = median(|y_i - \hat{y}_i|) \quad (5)$$

4.2. Experiments Results and Discussion

The performance of eight basic regressors on testing set is presented in Table 2. There are two classes of results, one with feature selection and the other without feature selection. They are respectively noted with 'FS' and 'Non-FS' labels. Figure 7 exhibits the bar graph that compares the values of several indicators for various models in a more visual manner.

Table 2. *The indicators of each model.*

Regressor	FS					NON-FS				
	R^2	EVS	MSE	MAE	MEDIAN	R^2	EVS	MSE	MAE	MEDIAN
LR	0.807	0.807	0.497	0.523	0.397	0.768	0.768	0.556	0.544	0.403
RR	0.807	0.807	0.497	0.523	0.397	0.768	0.768	0.556	0.543	0.403
Lasso	0.803	0.803	0.507	0.530	0.367	0.765	0.765	0.562	0.547	0.403
KNN	0.649	0.655	0.901	0.478	0.210	0.643	0.648	0.857	0.476	0.216
RF	0.934	0.934	0.169	0.272	0.178	0.911	0.911	0.213	0.292	0.184
DT	0.874	0.874	0.325	0.350	0.205	0.833	0.833	0.400	0.388	0.223
GBDT	0.920	0.920	0.207	0.305	0.209	0.900	0.900	0.239	0.325	0.221
SVM	0.582	0.612	1.071	0.682	0.549	0.542	0.567	1.097	0.690	0.559
STACKING	0.935	0.934	0.169	0.271	0.174	0.912	0.912	0.210	0.294	0.182

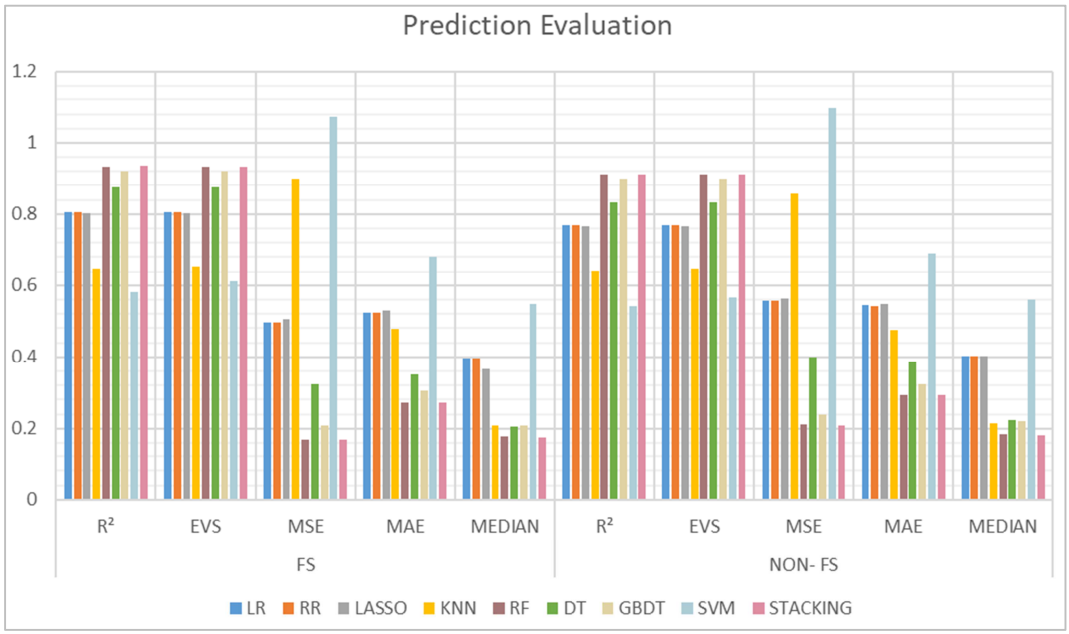


Figure 7. *Bar graph of evaluation indicators.*

It is obvious that almost every predictive model has certain performance improvement after feature selection is employed. On average, R^2 is improved by 3.96% and EVS by 4.02%. Additionally, the average error of the models is decreased by 1.12% in MSE, 4.67% in MAE, and 4.68% in MedianAE.

The tree-based non-linear regressors (i.e., RF, DT and GBDT) perform better. Linear regressors (LR, RR and Lasso) perform fair. The non-tree-based non-linear regressors (KNN and SVM) perform worse. Among the basic predictive models, RF has the best predictive accuracy.

The STACKING consisting of three competent basic predictors (i.e., DT, GBDT, and RF) and meta-regressor (i.e., RR) with higher predictive accuracy gets the highest $R^2$and EVS and the lowest MSE and MAE, proving that the combination of non-linear regressors and linear regressor in a stacking-based ensemble model can improve the flexibility and robustness of model.

5. Conclusion

This study successfully uses the stacking-based ensemble model to raise the performance of basic individual models, proving the effectiveness of ensemble learning method. It is also found that feature selection plays a significant role in improving the predictive accuracy.

The proposed ensemble learning model enables the bike-sharing services providers to make smarter decisions, enhances the intelligence level of shared bicycle services, and

enriches the applications of data mining in digital city.

However, the proposed ensemble model still has some drawbacks. First, the generalization ability of model can be further enhanced because only one dataset and weather information are taken into account. In practice, the transportation accessibility, road environment, and the distribution of bike parking spots are also worth considering in the future. Secondly, more data preprocessing and data mining methods can be explored to improve the predictive accuracy and robustness in dealing with regressive problems. Thirdly, this study only employs the stacking-based ensemble model, but does not explore the usefulness of other ensemble methods such as bagging, voting, and boosting. Hence, it is necessary to enrich the research data, improve the data processing effect, and explore the performance of other ensemble models in the future work.

References

[1] Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24 (1), 49-64.

[2] Bui, D. T., Tran, A. T., Klempe, H., Pradhan, B., & Revhaug, I. (2016). Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, 13 (2), 361-378.

[3] Chang, W., Ji, X., Wang, L., Liu, H., Zhang, Y., Chen, B., et al. (2021), A machine-learning method of predicting vital capacity plateau value for ventilatory pump failure based on data mining. *Healthcare*, DOI: 10.3390/healthcare9101306.

[4] Eren, E. & Uz, V. E. (2020). A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable Cities and Society*, DOI: 10.1016/j.scs.2019.101882.

[5] Fishman, E. (2016). Bikeshare: A review of recent literature. *Transport Reviews*, 36 (1), 92-113.

[6] Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017). Application of data mining methods in diabetes prediction. In *Proceedings of 2017 2nd International Conference on Image, Vision and Computing (ICIVC),* Chengdu, China, June 2-4, 2017, pp. 1006-1010.

[7] Lessmann, S., Baesens, B. U., Seow, H. V., Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: a ten-year update. *European Journal of Research*, 247 (1), 124-136.

[8] Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50 (3), 559-569.

[9] Nugumanova, A., Maulit, A., Mansurova, M., & Baiburin, Y. (2021). Understanding bike sharing stations usage with Chi-Square statistics. In *Proceedings of 13th International Conference on Computational Collective Intelligence*, Kallithea, Rhodes, Greece, September 29-October 1, 2021, pp. 425-436.

[10] Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9 (2), 181-199.

[11] Qi, Y., Li, Q., Karimian, H., & Liu, D. (2019). A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory. *Science of the Total Environment*, 664, 1-10.

[12] Ruß, G., Kruse, R. R., Schneider, M., & Wagner, P. (2008). Data mining with neural networks for wheat yield prediction. In *Proceedings of the 8th Industrial Conference on Advances in Data Mining: Medical Applications, E-Commerce, Marketing, and Theoretical Aspects*, Leipzig, Germany, July 16-18, 2008, pp. 47-56.

[13] Sathishkumar, V. E. & Cho, Y. (2020). A rule-based model for Seoul bike sharing demand prediction using weather data. *European Journal of Remote Sensing, 53* (sup1), 166-183.

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

[15] Sun, Y. (2018). Sharing and riding: how the dockless bike sharing scheme in China shapes the city. *Urban Science*, 2 (3), 68.

[16] Tjur, T. (2009), Coefficients of determination in logistic regression models - a new proposal: The coefficient of discrimination. *American Statistician*, 63 (4), 366-372.

[17] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5 (2), 241-259.

[18] Wu, C., Kuo, S., & Kao, S. C. (2019), Classification-based data mining applied in vehicle accident prediction. *Fuzzy Systems and Data Mining*, 320, 218-223.