2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) | 979-8-3503-3141-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICML60161.2023.10424944

Research on Demand Forecasting Method of Shared Bicycle Based on Ensemble Learning

Qiongshuai Lyu
College of Software
Pingdingshan University
Pingdingshan, China
4354@pdsu.edu.cn

Ruizhe Zhang*
College of Software
Pingdingshan University
Pingdingshan, China
*Corresponding author: richard217@qq.com

***Abstract*—With the continuous acceleration of urbanization, bike-sharing has become a convenient, environmentally friendly, and healthy way of travel. To alleviate the imbalance of shared bicycle demand, this paper designs a stacked random forest support vector machine regression ensemble model (RF-SVR) and a weighted average random forest support vector machine regression ensemble model. To verify the effectiveness of the design method, experiments were conducted on hourly bicycle rental datasets from Washington, D.C., USA from 2011 to 2012. Compared with the random forest regression model (RF) and support vector machine regression model (SVR), the experimental results show that the performance of the ensemble learning method is better than or equivalent to the single regression model under the evaluation index of RMSLE.**

Keywords-component: random forest; support vector machine; ensemble learning; demand forecasting for bicycles

I. Introduction

During the rapid development of shared bicycles, issues such as vehicle damage, maintenance, and imbalanced supply and demand are inevitable [1]. To address the problem of imbalanced supply and demand, predicting the future demand for shared bicycles using machine learning algorithms [2] can greatly help optimize bicycle resource allocation, improve bicycle utilization, and save operating costs [3].

In the related studies on addressing imbalanced supply and demand in shared bicycles, Yu et al. utilized Bayesian networks and association rules to predict bicycle demand correlated with time series [4]. Li et al. proposed a deep residual recurrent neural network method for bicycle demand prediction, which combines residual learning and recurrent neural networks to effectively capture long-term dependencies and complex patterns in the bicycle demand time series [5]. Yan et al. analyzed the correlation between shared bicycle demand data and building environmental data using spatial regression [6]. Zhang et al. combined the advantages of data-driven prediction algorithms such as random forest, XGBoost, and GBDT, and proposed a weighted logarithmic average combination model based on vector projection [7]. Yin et al. applied shared bicycle datasets to the Spark platform and built regression models using four machine-learning methods to predict bicycle demand [8]. Zhang et al. addressed the problem of overfitting in the random forest algorithm when dealing with datasets with a large amount of redundant data by proposing an improved algorithm called FWRF, which divides the feature space into highly correlated and lowly correlated intervals to limit the range of feature selection [9]. The aforementioned studies have achieved certain results in the field of shared bicycle demand prediction, but there is still room for improvement to further enhance the accuracy, robustness, and applicability of the models.

From the perspective of ensemble learning, this paper combines the random forest regression model with the support vector machine regression model and designs the Stacked RF-SVR model and Weighted Average RF-SVR model. Comparative experiments are conducted on a bicycle rental dataset [10]. The experimental results show that the ensemble learning method, especially the Stacked RF-SVR, performs better in predicting the demand for shared bicycles, with higher accuracy and predictive ability compared to other models.

II. Introduction of Relevant Algorithms

A. Random Forest Regression

Random Forest Regression RF is a method that improves the predictive ability of models by combining the predictions of multiple decision trees [11]. The model diagram is shown in Figure 1.

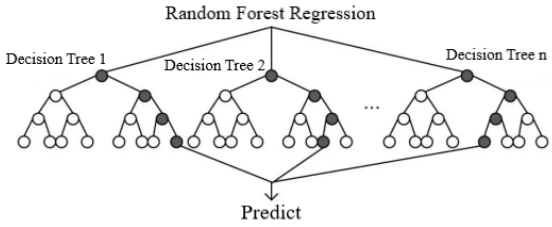

Figure 1. Random forest regression model.

In the specific implementation, random forests utilize bootstrap sampling to randomly extract n samples with replacement from the original dataset, creating a new sample set[12]. Simultaneously, for each decision tree, random forests randomly select m features and find the best feature from these m features to split the subtree [13].

B. Support Vector Machine Regression

Support vector regression SVR is a regression method based on support vector machines [14]. Unlike traditional linear regression approaches, SVR transforms the regression problem into a corresponding optimization problem and utilizes the principles of support vector machines to perform regression modeling[15].

The objective of SVR is to find a specific hyperplane that minimizes the error between the sample points on that hyperplane and the original data points [16]. In SVR, by

introducing the concept of "margin" or "tolerance range", some sample points are allowed to fall within the error range of the hyperplane, which enhances the robustness and generalization ability of the model [17].

III. Ensemble learning method

A. Stacking Integration

The stacking ensemble algorithm is a method that combines multiple models of different types to form a more powerful model [18]. The key idea behind the stacking ensemble algorithm is to leverage the strengths of different models by combining their predictive results to improve overall performance[19]. The base models can capture different features or patterns of the data from various perspectives, while the meta-model further enhances the accuracy and generalization ability of predictions by integrating the predictions of these base models. The principle of a two-layer stacking ensemble model is illustrated in Figure 2.

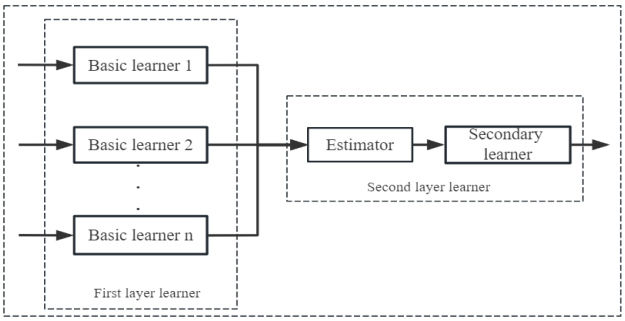

Figure 2. Two-layer stacked integration model.

The stacking ensemble algorithm excels in complex prediction tasks, where a single model may struggle to provide reliable and stable predictions due to the multifaceted nature of the task. Additionally, it proves beneficial in scenarios with noisy datasets, where the presence of noise and outliers can negatively impact the accuracy of a single model. Moreover, the stacking ensemble algorithm addresses the challenge of model selection, enabling the integration of multiple models to yield more accurate and stable predictions. Furthermore, it exhibits effectiveness in situations with limited training data, countering issues such as overfitting or underfitting.

B. Weighted Average

Weighted average is a commonly used method to calculate the average value. It takes into account the weights assigned to each data point and calculates the weighted sum of the data points. In a weighted average, different data points are given different weights to reflect their importance or contribution. The formula for calculating the weighted average is as follows:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_i f_i + \cdots + x_k f_k}{\sum_{i=1}^{k} f_i} \quad (1)$$

In the formula, x_i represents the *i-th* data point, f_i represents the weight corresponding to the *i-th* data point, and Σ represents the summation operation.

Weighted average is a flexible and effective calculation method. Its advantage lies in its ability to handle various types of data and models. Whether it is numerical data, categorical data, linear models, or nonlinear models, weighted averages can be applied. It can also provide more stable and accurate prediction results when dealing with uneven data distributions or the presence of outliers. In addition to simple weighted averages, some variant methods can further improve prediction results. For example, a weighted median can reduce the impact of outliers to some extent, and a weighted exponential average can place more emphasis on the predictions of the most recent data points or models.

IV. Experimental process

A. Dataset Introduction

The experimental data used in this article is selected from the Bike Sharing Demand competition dataset provided on the Kaggle platform [20]. The dataset consists of hourly bike-sharing data in Washington, D.C. for the years 2011-2012. It includes a training dataset and a testing dataset. The training dataset contains 10,886 data entries with 12 fields (attributes). Table 1 lists the names and meanings of each field in the training dataset. The testing dataset contains 6,493 data entries, including the first 9 fields listed in Table 1.

Table 1. Fields and their meanings in the training dataset

Filed	Meaning
datetime	date and timestamp of each hour.
season	season (spring, summer, fall, winter).
holiday	whether it is a holiday or not.
workingday	whether it is a working day or not.
weather	weather situation category.
temp	temperature in Celsius.
atemp	"feels like" temperature in Celsius.
humidity	relative humidity.
windspeed	wind speed.
casual	number of non-registered users renting bikes.
registered	number of registered users renting bikes.
count	total count of bikes rented.

B. Data Processing

1) Support Vector Machine Regression

By calculating the difference between the observed values of each sample and the mean value, data points that do not meet the 3-sigma rule are marked as outliers and removed. This process eliminates the disturbance caused by these outliers [21] in data analysis, modeling, and statistical inference, resulting in more accurate results.

2) Feature Extraction

Four new fields will be extracted from the datetime field, including date, hour, weekday, and month. These new time features provide a more granular and easily understandable representation of time. The values in the season and weather fields will be encoded and mapped to their corresponding textual descriptions. This encoding will facilitate a better understanding and interpretation of the data, as well as assist in further analysis.

3) Correlation Analysis

The relationships between the temperature, wind speed, humidity, and the number of rented bikes in the dataset were

visualized through data visualization. The generated scatter plot, as shown in Figure 3, illustrates these relationships.

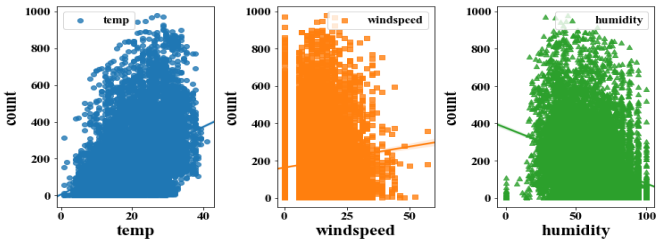

Figure 3. Correlation analysis.

From Figure 3, it can be observed that there is a positive correlation between temperature and the number of rented bikes. This means that as the temperature increases, the number of rented bikes also gradually increases. The highest number of rented bikes is observed at around 22 degrees Celsius. However, when the temperature approaches 40 degrees Celsius, the number of rentals sharply decreases.

On the contrary, when analyzing the relationship between wind speed and the number of rented bikes, we observe a weak correlation. This can be inferred from the regression line, which has a slope close to zero. A slope close to zero indicates that changes in wind speed have little impact on the number of bike rentals.

Additionally, we find a negative correlation between relative humidity and the number of rented bikes. This suggests that as relative humidity increases, the number of bike rentals tends to decrease. In other words, high relative humidity is associated with lower demand for bike rentals.

4) Monthly user count analysis

Figure 4 visualizes the average user count of bike rentals for each month using a bar chart.

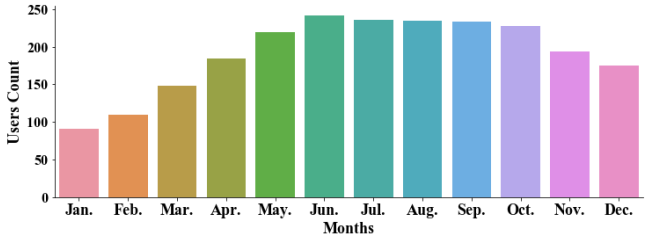

Figure 4. Average user count of bike rentals per month chart.

From Figure 4, it is evident that people prefer renting bikes during the summer season, as it is the most suitable time for cycling. Therefore, the demand for bike rentals is relatively high in June, July, and August.

In addition, Figure 5 visualizes the average user count of bike rentals per hour in different seasons using a scatterplot chart.

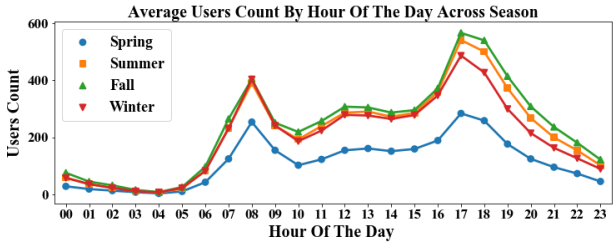

Figure 5. Average User Count per Hour Across Different Seasons.

By analyzing Figure 5, we can clearly observe a noticeable trend: people tend to rent bicycles more frequently during two specific periods. The first peak occurs in the morning hours, precisely from 7 a.m. to 8 a.m. This aligns perfectly with the typical commuting time during weekdays when individuals are heading to work or school. The second peak is observed in the afternoon, specifically from 5 p.m. to 6 p.m. This corresponds to the end of the workday or school hours when people are returning home. These patterns suggest that the demand for bicycle rentals is closely related to the daily commute and underscores the usefulness of shared bicycles as a convenient transportation option during these peak hours.

C. Model construction

Based on the aforementioned data analysis, two ensemble learning methods were designed to predict the demand for bike sharing. These methods are the stacked RF-SVR model and the weighted average RF-SVR model.

1) Establishment of Stacked RF-SVR Model

When designing the stacked RF-SVR model, it is necessary to create a random forest regressor RF and a support vector machine regressor SVR [22]. Then, an estimator is defined as a list containing two tuples, one for the RF tuple and the other for the SVR tuple. Finally, the stacked RF-SVR model for bike-sharing demand prediction is built based on the estimator list.

Firstly, the random forest regressor RF and the support vector regressor SVR are constructed. During the initialization of RF, parameters such as n_estimators=1000 and max_depth=10 are set. The SVR function is called to build the SVR, and parameters such as kernel='rbf', C=1.0, and epsilon=0.1 are set during the initialization. The initialized RF and SVR are then stored in the list as tuples, preparing for the construction of the stacked RF-SVR model.

In the stacking process, the random forest regressor RF and support vector machine regressor SVR are trained individually on the dataset. Then, the predicted outputs from RF and SVR are combined as features for the final SVR model. This stacking technique allows the SVR model to learn from the predictions of both RF and SVR, capturing complementary patterns and improving the overall prediction capability.

By selecting SVR as the final regressor in the stacking model, it is expected to leverage the strengths of SVR in capturing complex relationships and handling nonlinearities in the dataset. This decision aims to enhance the predictive performance of the stacked RF-SVR model. Figure 6 illustrates the principle diagram of the stacked RF-SVR model.

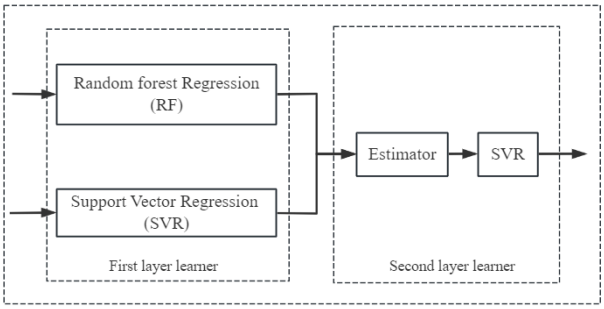

Figure 6. Schematic diagram of the stacked RF-SVR model.

2) Establishment of Stacked RF-SVR Model

In the weighted average RF-SVR model design, we construct two separate models: a random forest regressor (RF) and a support vector machine regressor (SVR). These models are then combined using a weighted average approach, where the relative weights of RF and SVR are controlled by weight factors. The specific method for controlling the weights in the weighted RF-SVR model is as follows:

$$r = \alpha \cdot F_{rf} + (1 - \alpha) \cdot F_{svr} \quad (2)$$

Among them, r represents the prediction result of the weighted average RF-SVR model, F_{rf} represents the RF model, and F_{svr} represents the SVR model. By adjusting the weight factor α, the contribution of the RF model and SVR model to the weighted average RF-SVR model is controlled. In this experimental process, the value of the weight factor α is set to 0.99. This means that almost all of the weight is assigned to the RF model, indicating that it has a higher influence on the final prediction result compared to the SVR model. By fine-tuning the weight factor, we can determine the optimal balance between the two models and achieve more accurate predictions in the weighted average RF-SVR model.

D. Experimental results

1) Experimental Environment and Evaluation Standards

During the validation of model effectiveness, the experimental software environment used was Anaconda3, with a Windows 10 64-bit operating system. The system was equipped with an Intel(R) Core(TM) i5-10200H CPU @ 2.40GHz processor and 16GB of memory. Jupyter Notebook was used as the programming interface, and the programming language used was Python 3.7.

The evaluation criterion used in this study is Root Mean Squared Logarithmic Error (RMSLE). RMSLE is a performance metric used to measure the error of regression model predictions. The specific calculation method is as follows:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(p_i + 1) - log(a_i + 1))^2} \quad (3)$$

Where p_i represents the predicted value of the *i-th* sample point, a_i represents the corresponding true value of that sample point, and n represents the total number of samples. A smaller RMSLE value indicates a smaller error between the model's predicted results and the true values, indicating better performance of the model.

2) Comparison of Prediction Results

To validate the effectiveness of the proposed model in this study, the designed stacked RF-SVR model and weighted average RF-SVR model were compared with the random forest regression model RF and support vector regression model SVR. The results of the comparative experiments on the test dataset are presented in Table II. It can be observed that the performance of the stacked RF-SVR model is the best, while the performance of the weighted average RF-SVR model is comparable to that of the RF model, making it an effective model selection. The SVR model has a relatively lower predictive performance. This indicates that in addressing the bike-sharing demand prediction problem, ensemble learning methods [23] can serve as a preferred solution strategy, with the stacked ensemble learning method (Stacked RF-SVR, 0.0864) outperforming the weighted ensemble learning method (Weighted average RF-SVR, 0.0872).

Table 2. Experimental Results

Model	RMSLE
RF	0.0869
SVR	0.3087
Stacked RF-SVR	0.0864
Weighted average RF-SVR	0.0872

V. Conclusions

To address the problem of imbalanced supply and demand in bike-sharing systems, this study focuses on the investigation of ensemble learning methods and proposes the stacked RF-SVR model and weighted average RF-SVR model. Experimental results demonstrate that the performance of ensemble learning methods surpasses or is comparable to individual regression models such as random forest regression (RF) and support vector regression (SVR), as measured by the RMSLE evaluation metric. Moreover, the stacked ensemble strategy outperforms the weighted ensemble strategy.

Moving forward, the next step would involve exploring the application of ensemble learning strategies in conjunction with GBDT, XGBoost, LightGBM, and CatBoost methods to construct models for tackling the imbalanced supply and demand issue in bike sharing systems, thereby achieving better prediction results.

Acknowledgment

This work was supported by the Key Scientific and Technological Project of Henan Province, China (No.232102210011), Pingdingshan University Doctoral Research Initiation Fund (No. PXY-BSQD-2023019), Pingdingshan College teaching reform research and practice project (No. 2022-JYZD03), the key scientific research projects of universities in Henan Province (No. 24A520034), Henan Province colleges and universities young backbone teacher training program.

References

[1] Chen He. Research on Shared Bicycle Demand Forecasting Method Based on Ensemble Learning. Modern Business, 2018(36):185-186.

[2] Liu Benxing. Research on Short-term Demand Forecasting of Station-based Shared Bicycles. Taoyuan Normal University, 2023.

[3] Li Tiancheng. Analysis of Shared Bicycle Demand Based on Machine Learning Methods. Modern Trade and Industrial, 2020, 41(25):40-41.

[4] Yu, Y., Li, Z., and Liu, X. Exploration of Bicycle Sharing System: Time Series Forecasting Using Bayesian Network and Association Rules. Transportation Research Part C: Emerging Technologies, 2016, 69: 208-219.

[5] Li, G., Du, W., & Liu, D. A deep residual recurrent neural network for demand prediction in bike-sharing systems. Transportation Research Part C: Emerging Technologies, 2020, 113:308-329.

[6] Yan, S., Ding, Z., & Wang, D. Z. Understanding the relationship between bike-sharing demand and built environment: A spatial regression approach. Transportation Research Part D: Transport and Environment, 2019, 67:101-114.

[7] Zhang Jiantong, Sun Jiaqing. Shared Bicycle Rental Demand Forecasting Based on Combined Prediction Method. Operations Research and Management Science, 2021, 30(10):146-152.

[8] Yin Lifeng, Li Zhao. Study on the Demand of Shared Bicycle Based on Spark Regression Analysis. Electronic Design Engineering, 2023, 31(08):5-9.

[9] Zhang Xu. Research and Application of Shared Bicycle Demand Forecasting and Dispatch Optimization Algorithm. Jiangsu University, 2020.

[10] Zhong Yingshan, Han Xiaoming. Shared Bicycle Station Demand Forecasting Based on Random Forest and Spatio-temporal Clustering. Science Technology and Engineering, 2018, 18(32):89-94.

[11] Zhou Yi, Feng Zhaoxiang, Bai Xizhuo, et al. Design of Data Analysis Software based on Random Forest Algorithm. Journal of Heilongjiang Institute of Engineering, 2017, 31(03):38-41.

[12] Rathika N,Abolfazl M,Pitambar K R, et al. Intelligent gravitational search random forest algorithm for fake news detection. International Journal of Modern Physics C,2022,33(06).

[13] Dong Na, Chang Jianfang, Wu Aiguo, et al. Forest Prediction Method Based on Bayesian Model Combination. Journal of Hunan University (Natural Sciences), 2019, 46(02):123-130.

[14] Cai Lianxiang. Research on Robust Support Vector Machine based on Zero Order Optimization. Hebei University, 2020.

[15] Vincent B,Duhamel C,Ren L, et al. A PCA and SVR based method for continuous industrial process modelling. IFAC PapersOnLine, 2018,51(11).

[16] Lin Fangdou, Zhao Weihua, Zhang Riquan. Bayesian Support Vector Regression and Its Application. Statistics and Decision, 2023, 39(03):49-54.

[17] Guo Miao. Application Research on Passenger Flow Prediction of Large Passenger Stations Based on Support Vector Regression. Railway Computer Application, 2021, 30(03):15-18.

[18] Sun Yiwen, Luo Ronglei. Clothing Network Live Sales Prediction Based on Stacking Ensemble Learning. Textile Dyeing and Finishing Technology, 2023, 45(04):1-5+21.

[19] Vanessa G A, F. G N,Jaime J G. Ensemble random forest filter: An alternative to the ensemble Kalman filter for inverse modeling. Journal of Hydrology,2022,615(PB).

[20] Tang Peipei, Wu Minghui. Exploration and Practice of "Data Mining Technology" Course Construction Based on Kaggle Competition Data. Industrial and Informationization Education, 2021(03):85-88.

[21] Hu Miao. Research on Anomaly Detection Algorithm Based on Random Forest. Fujian Normal University, 2019.

[22] Song Peng, Huang Tongyuan, Liu Yuqiao. Shared Bicycle Demand Forecasting Based on SVM. Journal of Chongqing University of Technology (Natural Science), 2019, 33(07):187-194.

[23] Luo Changwei, Wang Shuangshuang, Yin Junsong, et al. Research Status and Prospect of Ensemble Learning. Journal of Command and Control, 2023, 9(01):1-8.