Optimizing Bike Sharing Demand Prediction: A Comparative Study of Multiple Linear Regression and LSTM Models based on Time and Weather Factors

Xun Zhang [1,*], Zengyi Yu [2], Chen Xie [2]

[1] College of Mathematical Sciences, Zhejiang University of Technology, Hangzhou, China

[2] College of Education, Zhejiang University of Technology, Hangzhou, China

* Corresponding Author Email: zhangxun_20021107@163.com

Abstract. Bike sharing, as an emerging Internet product, utilizes advanced technologies like IoT, cloud computing, and big data to improve urban public transport and offer more travel options. However, poor demand forecasting has resulted in the over-placement of shared bicycles. In this paper, we explore the use of an LSTM-based approach to predict the demand for shared bikes by analyzing time and weather factors to optimize operations. The study begins by underscoring the significance of these factors in demand forecasting. After identifying and addressing outliers through the use of box plots and the DBI index, we find that bicycle usage exhibits a cyclical pattern over time with a lagged weather effect. To tackle this issue, we introduce triangular transformations of the temporal factor and lagged weather variables in our multiple linear regression model. While these adjustments enhance the model's ability to capture complex relationships, they result in poor accuracy. Consequently, we employ a Long Short-Term Memory (LSTM) neural network, which significantly outperforms the traditional multiple linear regression model. The LSTM model achieves RMSE values of 14.63 and 39.41, as well as R-squared values of 93.21% and 95.58% for casual and registered users, respectively. This highlights its superior ability to capture complex patterns in the data, providing insights for better management and optimization of shared bike systems.

Keywords: Bike-sharing demand prediction, Time & Weather factors, Multiple linear regression model; LSTM, Operational optimization.

1. Introduction

Bike sharing, as an emerging Internet product, uses advanced technologies such as the Internet of Things, cloud computing and big data to improve the urban public transport system and provide residents with more travelling options [1]. However, in order to rapidly capture market share, some bicycle-sharing enterprises have not fully considered market demand, resulting in excessive placement of shared bicycles [2].

Recent studies have made significant advancements in predicting bike-sharing demand. Sathishkumar et al. adopts a feature filtering method to eliminate non-predictive parameters. The resulting best model, Gradient Boosting Machine, achieves an R^2 value of 0.96 on the training set and 0.92 on the test set. [3]. Cho et al. applied the rule-based model CUBIST was able to explain about 95 and 89% of the Variance in the testing set of Seoul Bike data and Capital Bikeshare program data [4]. Kashyap et al. use regression models to predict bike-sharing demand, with the results showing an R-Square value of 56.7% [5]. However, these studies do not incorporate time and weather factors in predicting demand for bike sharing. Even if they existed, they only used simple machine learning models rather than deep learning models to make predictions. This study uses multiple linear regression (MLR) and LSTM tools to Forecast the demand for shared bikes at different times of the day and quantitatively assess the reliability of the forecasts. We have three Research Questions (RQs):

RQ1. How do time and weather factors influence the demand for bike-sharing services?

RQ2. What improvements can be achieved by incorporating time cyclic terms and weather lag effects in a multivariate linear regression model for bike-sharing demand prediction?

RQ3. How does the performance of advanced neural network methods like LSTM compare with traditional multivariate linear regression models in predicting bike-sharing demand?

Based on the RQs, the main contributions of this paper are:

(1) Quantitatively analyse the impact of time and weather factors on demand for shared bicycles: By analysing in detail the effects of time (e.g. hours, weekdays, holidays) and weather factors (e.g. temperature, humidity, wind speed, weather conditions) on the demand for shared bicycles, data support is provided for optimising operation and scheduling.

(2) propose and validate an improved method of introducing time-periodic term and weather lag effect in multiple linear regression model: Introducing the periodicity of time and the lag effect of weather factors into the multiple linear regression model significantly improves the predictive generality of the model.

(3) Comparison of advanced neural network methods (e.g., LSTM) with traditional multiple linear regression models in demand prediction of shared bicycles: LSTM models are constructed and trained and compared with traditional multiple linear regression models. The experimental results show that the LSTM model has significant advantages in handling time series data and capturing complex non-linear relationships based on time and weather factor conditions, with higher prediction accuracy.

(4) Provide practical guidance for the management and optimisation of shared bicycle systems: Based on the research results, some practical applications are suggested, such as considering time and weather factors in scheduling and resource allocation to improve operational efficiency and user satisfaction.

2. Methodology

2.1. Data Collection and Processing

In this section, we will collect and process data related to the "The daily per hour usage of shared bikes in a city during recent years". We eager to strengthen our analysis by explaining our data sources, data pre-processing, and data characteristics.

(1) Data Collection

The dataset for this experiment is sourced from the Bike Sharing Dataset (https://www.kaggle.com/datasets/lakshmi25npathi/bike-sharing-dataset), collected by the Capital bike -share system. It contains the hourly and daily count of rental bikes in the Capital bike share system with the corresponding weather and seasonal information.

(2) Data Pre-processing

First, we use box plots to identify outliers in the hourly data by calculating the upper and lower quartiles and the interquartile range (IQR) for each continuous variable. Variables out-side 1.5 times the IQR from the quartiles are deemed outliers. This process identifies 3,222 outliers, which we replace with the average of surrounding data points to minimize model impact. Due to the data's high dimensionality, we also use the Davies-Bouldin Index (DBI) to find the optimal number of clusters. For normalized data within the [0, 1] range, a grid search over [1, 10] yields an optimal k value of 5. We calculate outlier scores based on distances to cluster centers and identify the top 1% of outlier points within each cluster (174 points in total). To maintain data integrity, we replace these points with the average of adjacent data points.

(3) Factor Analysis

1) Time Factor Analysis

The following chart depicts the variation in bicycle-sharing usage by hour, based on the hour data (a subset of the time period is analyzed):

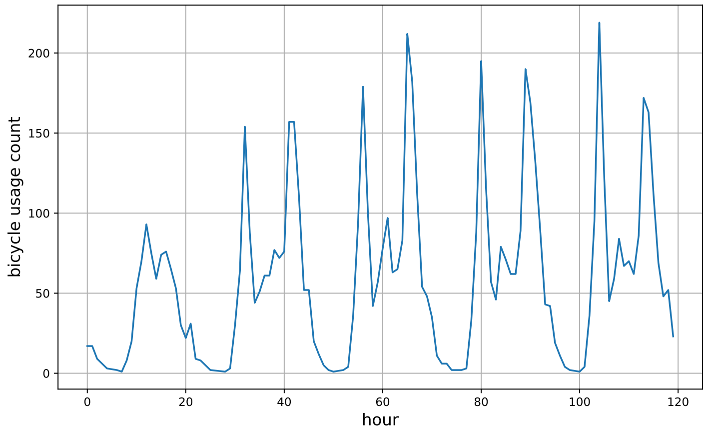

Figure 1. Bicycle usage count by hour

Based on the figure above, we can observe that the bicycle usage count exhibits a certain periodic variation over time. Therefore, in the following sections, we will attempt to use trigonometric transformations on the hourly data for regression analysis.

2) Weather Factor Analysis

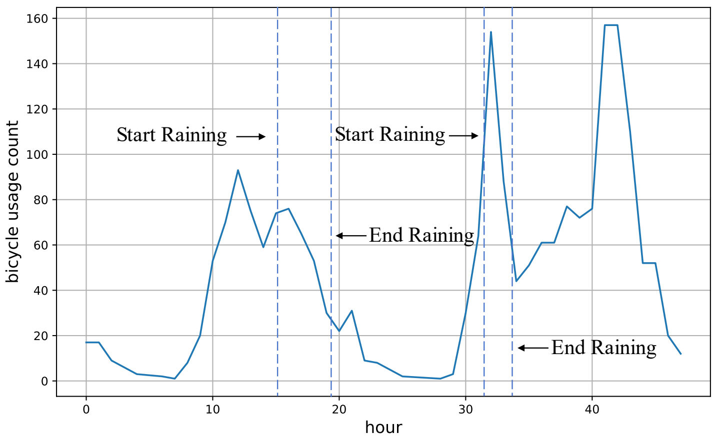

Figure 2. Bike-sharing use varies with the weather

From the figure above, it is evident that weather conditions significantly impact public bicycle usage. However, the effect of rainfall on usage is less clear for the current period. The figure shows an increase in usage during two rainy periods, possibly because some users did not anticipate the weather change, leading to an unexpected rise in usage during sudden rainfall. After the rain, there is a noticeable decline in usage, influenced by both time factors and the fact that most bicycle stations are outdoors. The slippery conditions caused by rain reduce users' willingness to use shared bicycles, indicating that the impact of rainy weather on bicycle usage is somewhat delayed. Similarly, when the rain stops, the usage of bikes does not immediately increase. This further confirms that the impact of weather on bike usage has a certain degree of lag.

2.2. Multiple Linear Regression Model

Based on the above analysis, we can determine the relationships between various variables and the dependent variable, registered (count of registered users), and thereby establish a multiple linear regression model to predict the registered dependent variable. The specific form is as follows:

$$y_{reg} = \widehat{\beta_0} + \sum_{i=1}^{11} \hat{\beta}_i \, \alpha_{i,t} + \sum_{i=0}^{t-1} \hat{\beta}_{temp_i} \, \alpha_{temp_i} + \sum_{i=0}^{t-1} \hat{\beta}_{hum_i} \, \alpha_{hum_i} + \sum_{i=0}^{t-1} \hat{\beta}_{windspeed_i} \, \alpha_{windspeed_i} + \widehat{\beta_s} sin\omega t + \widehat{\beta_c} cos\omega t + \varepsilon \tag{1}$$

Where $\alpha_{i,t}$ represents the value of the i-th factor at time t. Due to the lag effect of weather factors, the weather factors temp, hum, and windspeed need to be weighted and summed from time 0 to time $t-1$.

Similarly, we establish a multiple linear regression model to predict the dependent variable, Casual (count of casual users), based on the relationships between various variables and the Casual dependent variable.

To ensure that there is no multicollinearity among the indicators in the regression model, we use stepwise regression to fit the model. The holdout validation method is used to evaluate the reliability of the prediction model. Ultimately, 10 indicators are retained. The specific regression results are as follows:

Table 1. Regression Accuracy Assessment

	Casual Data	**Registered Data**
Holdout test accuracy	41.36%	45.67%
$\boldsymbol{R^2}$	53.91%	65.30%

As seen from the above table, the model's fit is poor, and its predictive power is weak. Attempts to improve accuracy through various transformations have been limited, suggesting that regression methods are not suitable for this dataset. Given the time and weather attributes of this data, we introduce a Long Short-Term Memory (LSTM) neural network for prediction [6].

2.3. Long Short-Term Memory (LSTM) Model

In the regression section above, we have already conducted a visual analysis. Since hours are essentially just a unit of time measurement, it is assumed that hours also exhibit some form of lag effect. That is, the indicators from the previous hour can influence the figures for the subsequent hour.

We use the indicator data from the $i-th$ hour of the $j-th$ day within a three-day period, along with data from the i-1 and i-2 hours of the same day as the input to a neuron, while the demand data from the i+1 hour serves as the neuron's output. Additionally, three gating functions are introduced to effectively address the issue of inconsistent data characteristics between weekends and weekdays. The input gate, forget gate, and output gate collectively control the impact of abrupt changes between weekend and weekday data. In this model, the Sigmoid function is used as the activation function for all three gates.

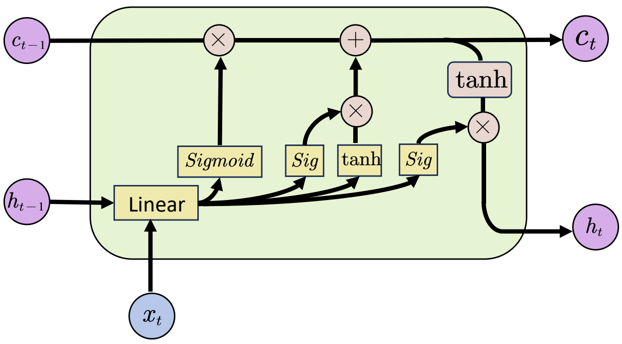

Figure 3. LSTM schematic

When using the neural network, we do not classify the data but use the entire dataset for prediction. This approach prevents potential interactions between different types of data and further avoids the fragmentation of continuous data characteristics that direct classification might cause. For example, the demand data for the first working day after a weekend might have different features compared to the demand data for the last working day before the weekend.

Based on the observed lag effect of weather over time, we believe that the weather conditions within a three-day period will impact the current demand data. Therefore, 72 sets of data are selected as the input to a neuron. The model was found to approach convergence after 70 iterations, so the number of iterations is set to 70. To select the training batch size, a preliminary grid search was performed, and the best epoch is selected to be 50, with the best batch size of 20 for training.

For the loss function, since this study involves predictive analysis of continuous variables, the root mean square error (RMSE) is used as the loss function. Due to the continuous nature of LSTM, the holdout validation method mentioned earlier is no longer applicable. Instead, the dataset is divided into three parts: the first 50% as the training set, 30% as the test set, and the final 20% as the validation set for predicting future data.

Given the "many to one" network structure used in this study, we can only predict the data for the next hour. Considering the inherent randomness in LSTM optimization algorithms, we repeat the prediction 5 times with the optimal hyperparameters and take the average as the final prediction result. After trying various optimization methods, we finally use the "Adam" optimizer for training the model.

3. Results And Discussions

This section presents the efficacy of the proposed LSTM-based methods for predicting bike-sharing demand. The LSTM models are trained to forecast the demand for both casual and registered users. The evaluation metrics, including RMSE, MAE, and R^2, demonstrate the accuracy and reliability of the predictions.

3.1. LSTM Results

After training, the evaluation parameters of LSTM prediction results are as follows:

Table 2. Evaluation of LSTM Prediction Results

	$RMSE$	MAE	R^2
Casual Data	14.63	8.79	93.21%
Registered Data	39.41	26.32	95.58%

The figures of the predicted data and the original data in the test set are shown below:

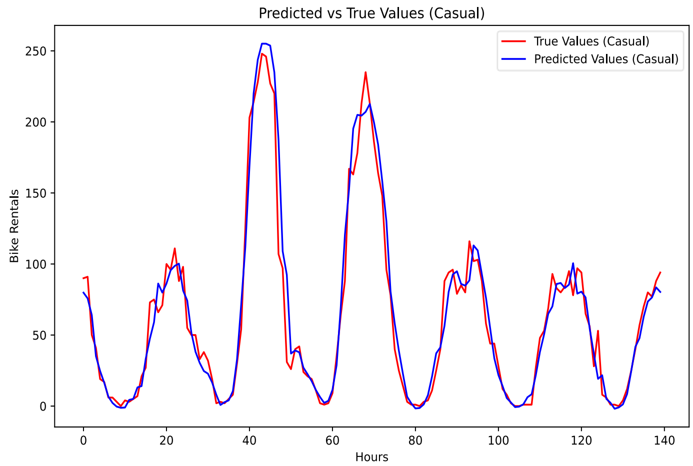

Figure 4. Comparison of predicted data with real data for the test set (Casual Data)

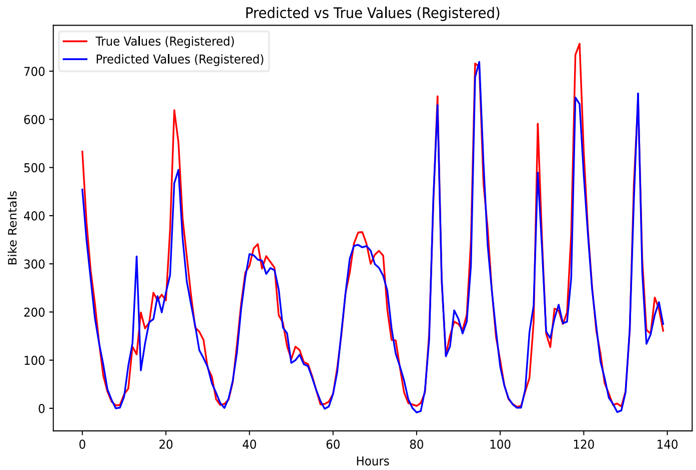

Figure 5. Comparison of predicted data with real data for the test set (Registered Data)

Combining Table II and Figure 4 & Figure 5, it can be seen that it can be found that at this point the prediction effect of the neural network has improved very significantly compared to the regression method, and is able to better predict the demand of Bike Sharing under Time and Weather Factors conditions.

3.2. Discussions

The efficacy of the LSTM-based methods in predicting bike-sharing demand has been demonstrated through quantitative analysis. The LSTM models have been specifically trained to forecast the demand for both casual and registered users, capturing the complex time and weather-related patterns inherent in the data.

(1) Impact of Time and Weather Factors:

The detailed analysis of time factors (hours, weekdays, holidays) and weather factors (temperature, humidity, wind speed, weather conditions) reveal their significant impact on the demand for bike-sharing services. The LSTM models effectively integrate these factors, providing robust support for optimizing bike-sharing operations and scheduling.

(2) Improved Prediction with LSTM Models:

By incorporating the periodicity of time and the lag effect of weather factors into the multiple linear regression model, the predictive generality is significantly improved. However, the LSTM models further enhance this by leveraging their ability to handle time series data and capture complex non-linear relationships. The RMSE values for the casual and registered data are 14.63 and 39.41, respectively, indicating a high level of accuracy. Additionally, the R^2 values of 93.21% for casual data and 95.58% for registered data demonstrate the models' effectiveness in explaining the variance in the data.

(3) Practical Implications for Bike-Sharing Systems:

The research findings suggest practical applications for managing and optimizing bike-sharing systems. Considering time and weather factors in scheduling and resource allocation can significantly enhance operational efficiency and user satisfaction. The LSTM models' high prediction accuracy makes them valuable tools for real-time demand forecasting, enabling better planning and decision-making for bike-sharing service providers.

(4) Potential for Broader Applications:

The methodologies and insights gained from this study can be extended to other shared mobility services and urban transportation systems. For instance, ride-hailing services, car-sharing programs, and public transit systems can also benefit from incorporating time and weather factors into their demand forecasting models. Additionally, the use of LSTM and other advanced neural network models can help address similar challenges in other domains, such as energy consumption forecasting, traffic flow prediction, and inventory management. By adopting these advanced predictive techniques, various industries are able to achieve improved operational efficiency, cost savings, and enhanced user experiences.

4. Conclusion

This paper investigates the advantages of using an LSTM-based approach to predict demand for shared bikes. By analysing the impact of time and weather factors and incorporating them into the prediction model, the study provides strong support for optimising shared bike operations.

Addressing RQ1, the study shows that temporal characteristics are critical for understanding demand patterns, while weather factors can have a significant impact on demand fluctuations. These findings emphasise the importance of considering both time and weather factors in forecasting models to accurately predict demand for shared bikes.

In the case of RQ2, the inclusion of time-periodic terms and weather lag effects in the multiple linear regression model enhances the model's ability to capture complex relationships in the data, resulting in greater generalisation but less desirable accuracy than a traditional model without these factors.

In the study of RQ3, the LSTM model performs significantly better than the traditional multiple linear regression model, with RMSE values of 14.63 and 39.41, and R^2 values of 93.21% and 95.58% for the casual and registered user data, respectively, demonstrating its superior ability to capture complex patterns and dependencies in the data. This resulted in more accurate and reliable demand forecasts, with the LSTM model achieving higher R^2 values and lower RMSE values.

These results offer valuable insights for enhancing the management and optimization of bike-sharing systems, ultimately improving operational efficiency and user satisfaction. The methodologies developed in this study have the potential for broader applications across various shared mobility services and urban transportation systems, further amplifying their impact and utility. Future work will involve expanding the dataset to include additional urban environments and optimizing LSTM algorithms to enhance predictive accuracy, providing more exploration directions for shared mobility services and urban transportation systems.

Acknowledgments

This research was partially supported by Zhejiang students' technology and innovation program (Fund No. G24201080035).

References

[1] Croall R, Jonsson Lundqvist D. Here I go: A prediction model for e-bike and e-scooter positioning inside a CCAM environment [J]. 2024.

[2] Zhao J, Wang H, Huang Y, et al. Does massive placement of bicycles win the market for the bicycle-sharing company in China? [J]. Sustainability, 2020, 12 (13): 5279.

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

[4] VE S, Cho Y. A rule-based model for Seoul Bike sharing demand prediction using weather data [J]. European Journal of Remote Sensing, 2020, 53 (sup1): 166-183.

[5] Kashyap A S, Swastik S. Regression Model to Predict Bike Sharing Demand [J]. Int. J. Innov. Sci. Res. Technol., 2021, 6 (3): 1024-1028.

[6] Ma X, Yin Y, Jin Y, et al. Short-term prediction of bike-sharing demand using multi-source data: a spatial-temporal graph attentional LSTM approach [J]. Applied Sciences, 2022, 12 (3): 1161.