Prediction of Shared Bicycle Entry and Exit Flow based on LSTM Model

Aobo Zhang *

Department of Software, Dalian Jiaotong University, Dalian, 116000, China

* Corresponding Author Email: yvonne@tzc.edu.cn

Abstract. As the use of shared bicycles becomes increasingly prevalent in urban areas, the issue of short-distance travel for residents has been significantly alleviated. Moreover, within the 30-minute radius of daily life, bicycles serve as a convenient solution for short-distance travel. However, the deployment strategy employed by shared bike operators lacks scientific rationality. The placement of bicycles in inappropriate locations compromises their convenience. This research aims to establish a predictive model for forecasting the entry and exit flow of shared bicycles in each region of the city. The shared bicycle data for January 2024 were collected from the Divvy online platform in Chicago. Python was utilized to extract and process the trip data, and time features were handled by minute-wise aggregation of entry and exit flows. By selecting data from the past 30 minutes as input features, normalizing the data, and constructing time series data suitable for LSTM input, the predictive model was developed. Following model training, traffic prediction was conducted using the test set, and the model's performance was evaluated. Ultimately, the effectiveness of the model was intuitively understood by plotting comparative graphs between actual and predicted values. This research aims to provide real-time decision support for urban traffic management and shared bike operations by offering insights into the predictive trends.

Keywords: Shared bike; LSTM; usage prediction.

1. Introduction

The emergence of shared bicycles as a popular mode of transportation for public transportation has allowed urban dwellers to enjoy quick and easy access to short-distance travel while simultaneously supporting a green and sustainable city. Shared bicycles also serve as an efficient solution to first- and last-mile problems [1]. However, issues like vacant, disorganized parking spaces and the regional "tidal" supply shortage are also becoming more prevalent. These are caused by residents' commuter behavior and the division of urban functions; these issues typically arise during a fixed time period (the morning and evening rush hour on workdays) and in particular locations (near subway stations). These phenomena have a significant impact on the user experience and the effectiveness of the shared cycling system. Based on the commonly held idea that people often choose the shortest path and that bicycles may no longer choose the quickest route when other factors such as accessibility and congestion interfere, this grade was created. The index improves the way that cycling behavior is characterized in earlier studies by drawing attention to cyclists' differing degrees of motivation to choose a path [2]. It also improved the position update strategies of followers and leaders in the population, reaching a conclusion that the rational distribution of shared bicycles should take pedestrian density, bicycle density, operating costs, and safety costs into account [3].

Researchers from various research units have undertaken collaborative efforts to investigate spatial and temporal patterns in shared cycling usage. For instance, Jiang Xiao employed a community-scale approach, utilizing space-time clustering methods to scrutinize the characteristics of shared cycling tidal patterns. The findings revealed pronounced trends during weekday mornings, characterized by a substantial influx of shared cyclists towards office spaces and shopping malls, while the outflow was notably directed towards residential areas and factories [4]. Moreover, in light of the swift advancements in wireless communication and digital electronics technologies, the landscape of machine learning research leveraging big data has progressed significantly within the realm of transportation [5]. Over the past three years, there has been notable progress in the domain of short-term predictions pertaining to bike-sharing demand. Enrico Collini and his research team employed

a Bidirectional Long Short-Term Memory (Bi-LSTM) model to devise a prognostic methodology aimed at forecasting the requisite number of available bikes in a smart bike-sharing system over short time intervals. Remarkably, this study drew upon a dataset spanning merely three months [6].

The Long Short-Term Memory (LSTM) network has become a powerful tool for sequential modeling among different prediction approaches, and it has attracted a lot of attention when it comes to bike-sharing trajectory prediction. Because long-term dependencies within time sequences are so well-represented by LSTM, it is especially well-suited to handle the spatiotemporal complexities included in bike-sharing trajectory data. Using Long Short-Term Memory (LSTM) networks for predicting bike-sharing demand offers numerous advantages, especially when compared to traditional machine learning models. To predict the hourly demand of bike sharing more accurately, the trend in algorithm usage has shifted away from machine learning algorithms toward deep learning algorithms. The Backpropagation (BP) neural network exhibits commendable adaptability, effectively addressing nonlinear problems by autonomously learning from the dataset. However, a notable limitation arises in that the trained model is susceptible to local extrema during the prediction phase [7]. This capability is crucial for tasks such as predicting bike-sharing trajectories, where understanding the temporal patterns is of paramount importance. The internal memory units of LSTM can store past information and selectively forget or update it as needed. Furthermore, the integration of an attention mechanism proves beneficial in enhancing both the accuracy and interpretability of deep-learning neural networks [8]. This would significantly assist urban planners and city designers in formulating judicious plans for public transport and non-motorized vehicle capacity [9]. This ensured that the proposed station layout conferred advantages to city administration, vendor revenue, and user convenience [10].

Despite numerous breakthroughs achieved by researchers utilizing a diverse array of machine learning algorithms, a notable challenge persists: the limited scale of data employed in these investigations hinders in-depth exploration. This constraint poses difficulties in identifying factors that may wield significant influence on research outcomes. The utilization of a larger dataset holds promise for advancing the refinement of deep learning algorithms. Consequently, the subsequent research endeavors to leverage an expanded dataset to uncover new possibilities related to factors crucial for enhancing the efficacy of deploying shared bicycles in urban settings. Furthermore, it aims to elucidate the role played by location factors in optimizing the distribution of shared bicycles within urban areas.

2. Methods

2.1. Data Source

In this research, data is sourced from the Chicago Divvy bike-sharing system, encompassing records of bicycle rides conducted within the geographical coordinates of the city of Chicago (41.64-42.07N, 87.54-87.82W) from January 2024 to February 2024. As delineated in Table 1, the data is meticulously organized within CSV files, encapsulating details such as bike identifiers, commencement and conclusion timestamps for rides, spatial coordinates corresponding to initiation and termination points, as well as user categorization.

To ascertain the temporal and spatial attributes of the data, the investigator undertakes filtering and classification procedures on the extensive and unorganized dataset. This method facilitates the segmentation of the data into meaningful subsets, thereby enabling subsequent in-depth analysis.

Table 1. Attribute information for raw data

Field	Instruction	Data Type
Ride_id	Bicycle ID	Object
Rideable_type	Bicycle type	Object
Started_at	Start time	Object
Ended_at	End time	Object
Start_station_name	Start position	Object
Start_station_id	Start position ID	Object
End_station_name	End position	Object
End_station_id	End position ID	Object
Start_lat	Latitude of start	Float64
Start_lng	Longitude of start	Float64
End_lat	Latitude of end	Float64
End_lng	Longitude of end	Float64
Member_casual	Member type	Object

2.2. Indicator Description

In this study, the primary indicator used is the predicted short-term bike inflow and outflow, derived from an LSTM-based time series forecasting model. This indicator provides valuable insights into the anticipated demand for shared bikes in specific urban areas within short time intervals.

Temporal Aspect: The predictive model is designed to predict the expected bicycle inflow and outflow per unit of time, with a time granularity of 1 minute to 30 minutes. By statistically statistically counting the data on a minute-by-minute basis and using a 30-minute time window as an input feature, the model can adapt to the deployment strategy of bike sharing in real time. This choice of time granularity allows the model to capture short-term fluctuations and trends in bicycle usage, providing important decision support for sharing bike operating companies.

Spatial Aspect: The spatial distribution of predicted bike inflow and outflow is indicative of the areas with higher demand for shared bikes. Areas experiencing frequent usage within short time intervals are likely to have a higher population in need of the bike-sharing service.

Population Demand Representation: The number of shared bikes used in different areas within short time intervals serves as a proxy for the population's demand for bike-sharing in those areas. This reflects the spatial distribution of demand for shared bikes among urban residents.

Optimization Implications: Utilizing the predicted short-term bike inflow and outflow, urban planners and bike-sharing operators can dynamically adjust deployment strategies in real-time. Efficient allocation of shared bikes to areas with higher predicted demand enhances the overall service efficiency.

2.3. Method Introduction

In this study, Python was employed for the uniform processing of data, extracting the frequency of bicycle usage on specific plots daily. Subsequently, a Sequential model was utilized to analyze data spanning one month, evaluating the actual fitting performance against the fitted values. The forthcoming values for the next time interval are anticipated to be predicted (Figure 1).

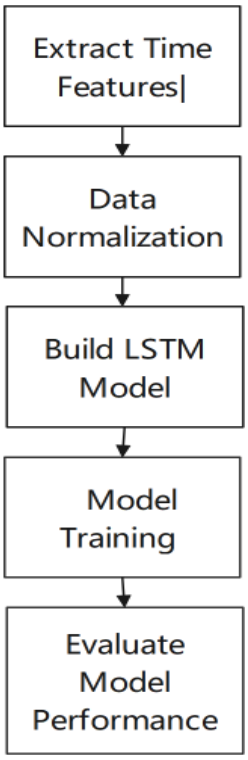

Figure 1. Flow chart of the implementation of the prediction model

3. Results and Discussion

3.1. Data Processing

The researchers used multiple panda data frameworks to organize and collect daily bike-sharing data from January 2024 to February 2024. As shown in Figure 2, the distribution of stations at the start and end points of shared bicycle riding.

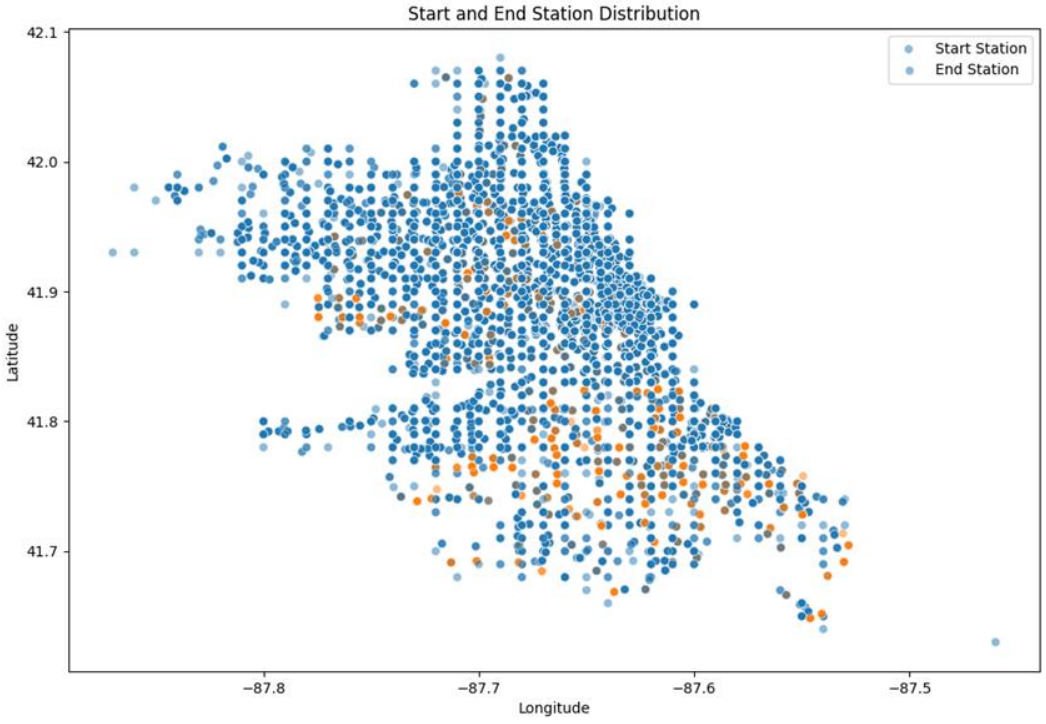

Figure 2. The distribution of stops at the start and end points of the ride

Based on the coordinates of shared bicycles in the dataset (latitude from 41.64 to 42.07 degrees, longitude from 87.54 to 87.82 degrees), the research area is partitioned into a 43 x 28 grid with units of 0.01 degrees. The activation count of bicycles within each unit is computed to construct a comprehensive dataset illustrating the spatial distribution of shared bicycle utilization. This dataset is utilized for data training. Researchers conducted statistical analysis on the starting and ending points of shared bicycle usage each day, aiming to predict the daily usage patterns of shared bicycles within the designated stations.

3.2. Sequential Model

The Sequential model is based on Long Short-Term Memory (LSTM) networks, which are a type of recurrent neural networks (RNNs). Sequential models are a simple type of model provided by Keras that allows users to build neural networks by sequentially adding layers of layers. The use of this model container is very intuitive and is particularly suitable for linearly stacked layer structures. In the Sequential model, the layers are connected in the order in which they are added, forming a linear network structure. In the Sequential model, LSTM layers can be added to form a layer-by-layer stacked structure. The intuitiveness of sequential models combined with the long-term dependency modeling capability of LSTM networks make them a potent tool for handling time series data.

3.3. Model Results

Figure 3 shows the distribution of different cycling types and the distribution of members and non-members, and the relative frequency of each cycling type, the relative ratio of members and non-members can be clearly seen. Therefore, Figure 3 is useful for understanding the characteristics of user groups.

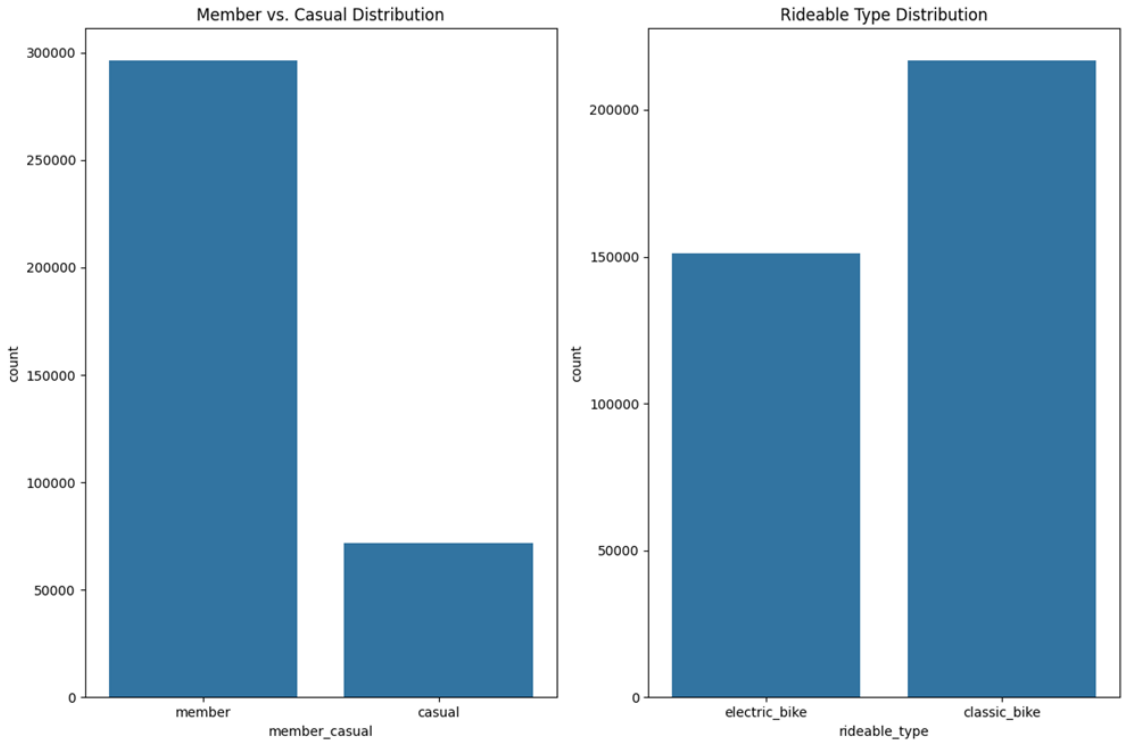

Figure 3. Proportion of members and non-members and the distribution of different riding types

Figure 4 extracts user behavior features, calculates average ride duration and average ride distance, and adds these features to the data: time features (hours and days of the week) are extracted, and ride distance and duration are calculated. In addition, in order to handle missing values, the itinerary data containing the missing spatial information was deleted. Figure 4 shows the distribution of cycling distance, which is of great significance for understanding the spatial characteristics of users' cycling behavior.

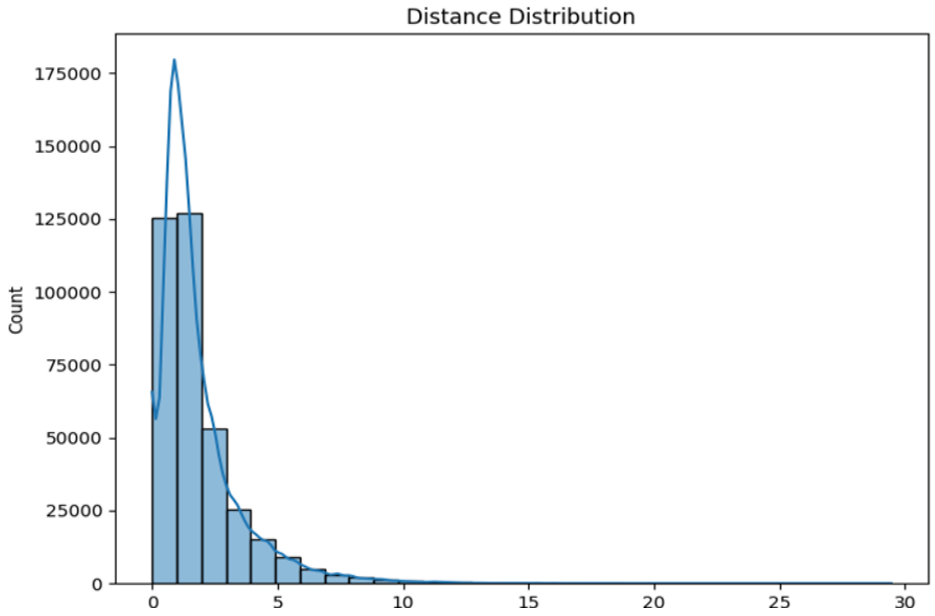

Figure 4. The distribution of riding distances

Ultimately, Figure 5 conducts traffic prediction using an LSTM model by selecting data from the past 30 minutes as input features. Specifically, the data is normalized, and time series data suitable for LSTM input is constructed. Upon completing the model training, traffic prediction is performed using the test set, and the model's performance is evaluated. Finally, by plotting a comparative graph between actual values and predicted values, the model's effectiveness can be intuitively assessed. Based on the results of the model evaluation, the Root Mean Square Error (RMSE) of the prediction errors is calculated. This metric provides a measure of the disparity between actual traffic and model-predicted traffic. The comparative graph between actual and predicted values allows for the visual observation of the model's predictive trends. This chart serves not only to assess the model's performance but also aids in a better understanding of the fluctuation trends in the shared bike entry and exit traffic.

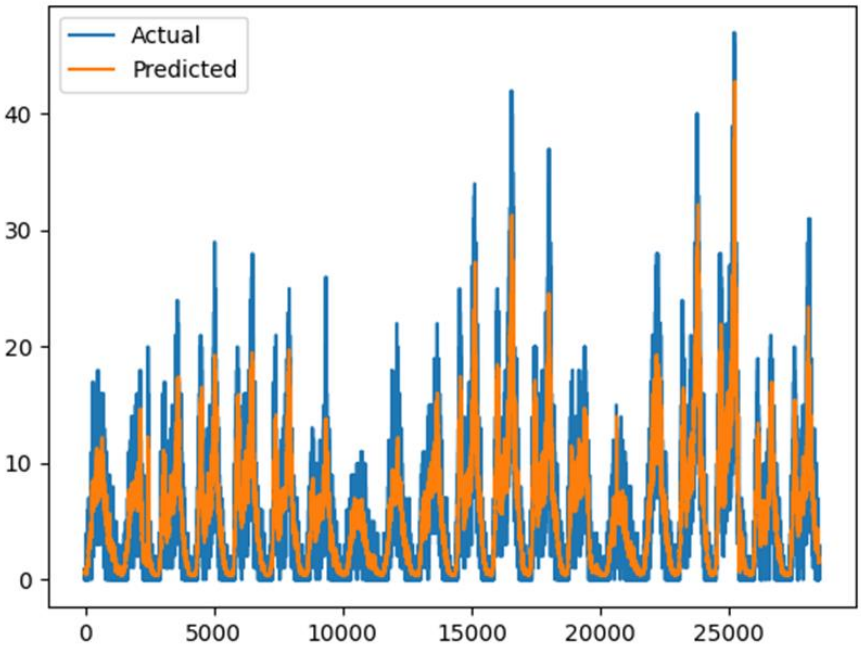

Fig. 5 A comparison chart of actual and predicted values

4. Conclusion

In this study, a sequential model was constructed based on existing data to predict the daily usage of shared bicycles in specific locations, yielding relatively accurate results. By merging two months of shared bicycle data, a more comprehensive dataset was obtained, which helped improve the model's generalization ability and prediction accuracy. The adopted model structure is an improved LSTM model, which includes multiple LSTM layers and Dropout layers to enhance the model's learning and generalization capabilities. After training and testing, it was found that the model performed well on the test set. Evaluation based on Root Mean Squared Error (RMSE) showed that the error between the model's predictions and the actual results was relatively small. This indicates that the model can accurately capture the trends and changes in shared bicycle usage to some extent. Further analysis revealed that the model's predictions were more accurate in the early stages, which is related to the scale of the data input. The higher the degree of match between the model and the real data in the early stages, the greater the accuracy of the prediction results. This provides important reference information for shared bicycle operators, enabling them to better plan bicycle placement and maintenance, thereby improving the efficiency and service quality of the shared bicycle system. In conclusion, by establishing an analytical prediction model, it is possible to effectively predict the usage of shared bicycles and provide decision support for shared bicycle operators, thereby optimizing the operational efficiency and user experience of the shared bicycle system.

References

[1] Kuang Jiaheng, Wu Qunyong. Spatial-temporal Equilibrium Analysis and Attraction Area Optimization of Dockless Sharing Bicycles Connected to Subway Stations. Journal of Geo information Science, 2022, 24(7): 1337-1348

[2] Zhao B, Deng Y, Luo L, et al. Preferred streets: assessing the impact of the street environment on cycling behaviors using the geographically weighted regression. Transportation, 2024, 1-27.

[3] Zhou Chuan. Optimization of sharing bicycle density distribution based on improved salp swarm algorithm. Computer Science, 2021.

[4] Jiang Xiao, Bai Lubin, Lou Xiayin, et al. Usage Patterns Identification and Flow Prediction of Bikesharing System Based on Multiscale Spatiotemporal Clustering. Journal of Geo-information Science, 2022, 24(6): 1047-1060.

[5] Bao H, Zhou X, Hamann C, et al. Understanding children's cycling route selection through spatial trajectory data mining. Transportation research interdisciplinary perspectives, 2023.

[6] Collini E, Nesi P, Pantaleo G. Deep learning for short-term prediction of available bikes on bike-sharing stations. IEEE Access, 2021, 9.

[7] Liu, M.; Shi, J. A cellular automata traffic flow model combined with a BP neural network based microscopic lane changing decision model. J. Intell. Transp. Syst., 2019, 23, 309-318.

[8] 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. Applied Sciences, 2022, 12(3): 1161.

[9] Aqib M, Mehmood R, Alzahrani A, et al. Rapid transit systems: smarter urban planning using big data, in-memory computing, deep learning, and GPUs. Sustainability, 2019, 11(10): 2736.

[10] Chen J, et al. Dynamic planning of bicycle stations in dockless public bicycle-sharing system using gated graph neural network. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(2).