Short-term Passenger Flow Prediction for Hangzhou Metro based on Machine Learning Algorithms

Zhiyu Li*

Department of Computer and Software, Chengdu Jincheng College, Chengdu, 611731, China

* Corresponding Author Email: lizhiyu@cdjcc.edu.cn

Abstract. Subways have now reached an irreplaceable position in urban rail transportation. The aggregation of people choosing subway travel at the same time can easily lead to passenger congestion, so predicting passenger flow in advance is crucial. This study used LSTM (Long Short-Term Memory) as the data prediction model to forecast the card-swiping data of the Hangzhou subway for 10 consecutive days. The historical data was divided into intervals of 3 minutes, and the random forest was used to compare the correlation among features. The final prediction results showed that both the inbound and outbound flows exhibited a bimodal pattern on weekdays, while on holidays, the prediction results showed a bimodal pattern. Furthermore, by adjusting the parameters appropriately, the simulated trends closely matched the actual values, indicating the applicability of LSTM in simulating short-term passenger flow. Accurate prediction of passenger flow can enable more comprehensive management of the subway system and provide passengers with a higher-quality service experience and travel options.

Keywords: Metro, LSTM, random forest, forecast.

1. Introduction

As cities expand and populations increasingly migrate to urban areas, urban road networks become more complex. Additionally, the impact of short-term events such as extreme weather and large-scale events further exacerbates the uneven distribution of passenger flows, leading to challenging congestion control during peak periods in subway systems. To alleviate this congestion, historical passenger flow information can be leveraged to predict short-term changes in passenger flow at congested stations using machine learning algorithms. Passenger flow prediction plays a crucial role in transportation planning and management, attracting significant attention from industry professionals. For the most popular urban rail transit systems, such as subways, the accuracy of passenger flow prediction is highly stringent. It helps subway operators anticipate passenger flows during different periods, allowing them to schedule subway services effectively, optimize timetables, and meet passengers' travel needs even during peak hours, enhancing overall user satisfaction [1].

In the early stages, traffic flow prediction mostly relied on statistical theory methods. In 1984, the Kalman filtering technique was applied to short-term traffic flow prediction. The Auto-regressive Integrated Moving Average with Exogenous Variables (ARIMAX) model was also used for traffic flow prediction on highways in France [2]. However, due to the fast update frequency, strong randomness, and significant data fluctuations in traffic data, the aforementioned parameter models have limited compatibility and are not suitable for short-term passenger flow prediction [3].

As data prediction methods continue to evolve, machine learning has gradually become the preferred choice for traffic flow prediction, as it can discover complex patterns and relationships within traffic data. Two widely used machine learning techniques in this context are Random Forest and Long Short-Term Memory (LSTM). Random Forest combines multiple decision trees for data prediction, making it suitable for constructing ensemble classification models for traffic flow prediction [4]. Random Forest efficiently handles high-dimensional traffic data and can handle missing values within the data. Some studies have utilized Random Forest for short-term traffic data processing and integrated multiple data sources, including historical weather information. The inclusion of comprehensive data sources improves the accuracy of predictions by making them closer to the ground truth values.

The LSTM model is a type of recurrent neural network (RNN) that has been developed and optimized based on RNN architecture [5]. It incorporates memory cells and forgets gates, effectively addressing the gradient vanishing and long-term dependency [6]. Compared to other conventional neural network models, LSTM performs well in handling long-distance sequential problems, making it widely applied in various short-term traffic flow prediction scenarios [7]. Many researchers have also explored variant models of LSTM by combining them with other techniques. For instance, the combination of Gated Recurrent Unit (GRU) and LSTM has been used to predict passenger flow at railway stations, resulting in more accurate predictions compared to traditional mathematical and statistical models [8]. These studies demonstrate the sensitivity of LSTM in capturing temporal features, making it an excellent predictor in traffic data forecasting [9].

In addition to Random Forest and LSTM mentioned in this paper, there are many other machine learning methods extensively used in traffic flow prediction, such as K-Nearest Neighbor (KNN) and Support Vector Regression (SVR). However, each method has its advantages and limitations, and the prediction accuracy may vary depending on the characteristics of the dataset. Therefore, this paper chooses to first use the Random Forest algorithm to select and optimize the features of approximately tens of millions of records of Hangzhou subway data in January 2019, and then use the LSTM model to predict short-term passenger flow [10]. The goal is to find an accurate and short-term algorithm model, and the obtained analysis results can assist in planning reasonable travel routes, mitigate the risk of heavy passenger flow, provide relevant information to subway personnel, assist them in better scheduling, ensure the normal operation of the subway, and make people's travel safer and more comfortable [11].

2. Methods

2.1 Data Collection and Preprocessing

The data for this project is sourced from the Aliyun Tianchi competition's subway card swiping dataset, specifically from January 1st to January 10th, 2019. The dataset includes records from three subway lines and 81 stations, totaling approximately 100 million entries. In this project, data from these 10 days were selected for short-term traffic flow forecasting experiments. The data field types of the initial required subway card swiping data are depicted in Table 1.

Table 1. Initial Data

column names	Type	explain	column names
time	String	Subway card swiping time	time
lineID	String	Metro line ID	lineID
3	213	654	649

The subway data was filtered based on Line B, subway operating hours, and entry/exit status, excluding certain non-operating hours data such as card swiping records of staff commuting. The data sample is shown in Table 2.

Table 2. The preprocessed data

time	lineID	stationID	deviceID
2019-01-03 06:00:00	B	7	327
2019-01-03 06:00:01	B	29	1436
2019-01-03 06:00:02	B	7	333

During the data preprocessing phase using Python, checks were performed to identify any missing values or empty cells. It was found that the dataset had already undergone preprocessing, and there were minimal dirty data. This significantly reduced the time and computational costs associated with data processing.

In the process of initial data processing, all features are first extracted for random forest voting to select the most relevant features for subsequent analysis. The subway card swiping data is collected 24 hours a day. For analysis and forecasting, the subway operation time of line B from 6:00 to 23:02 was selected, and the time was divided into 3-minute intervals. The preliminary graph of subway entrance passenger flow data is shown in Figure 1.

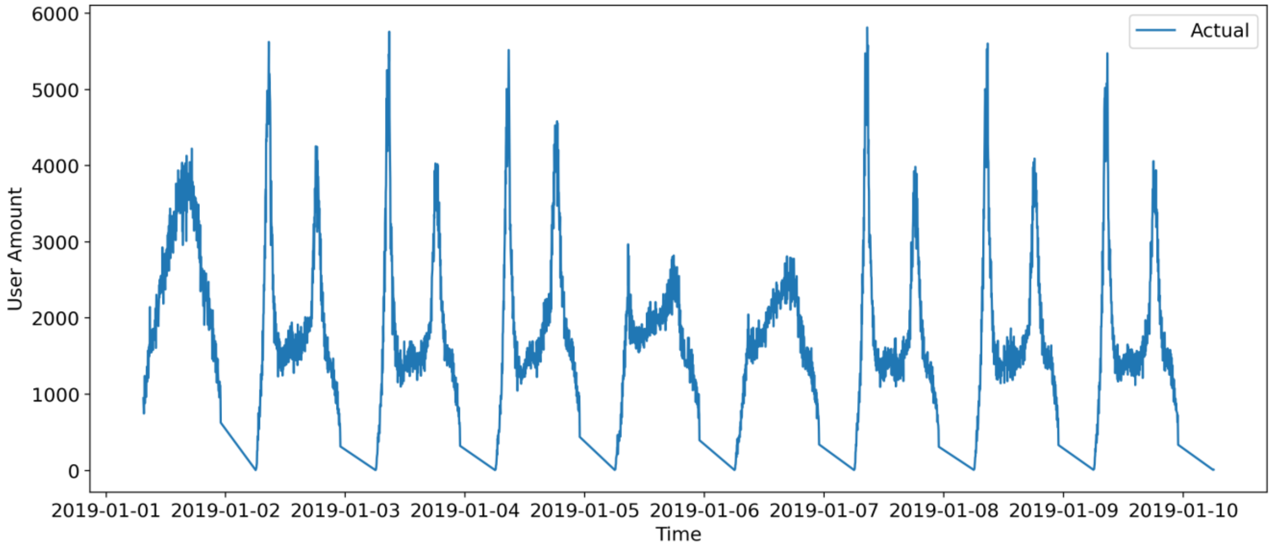

Fig. 1 Preliminary metro entrance passenger flow

Similarly, this study also selected the same time period as the entrance passenger flow data and analyzed exit passenger flow data at 3-minute intervals, the metro card swiping data is collected 24 hours a day. The data on subway passenger flow during exit for these ten days is shown in Figure 2.

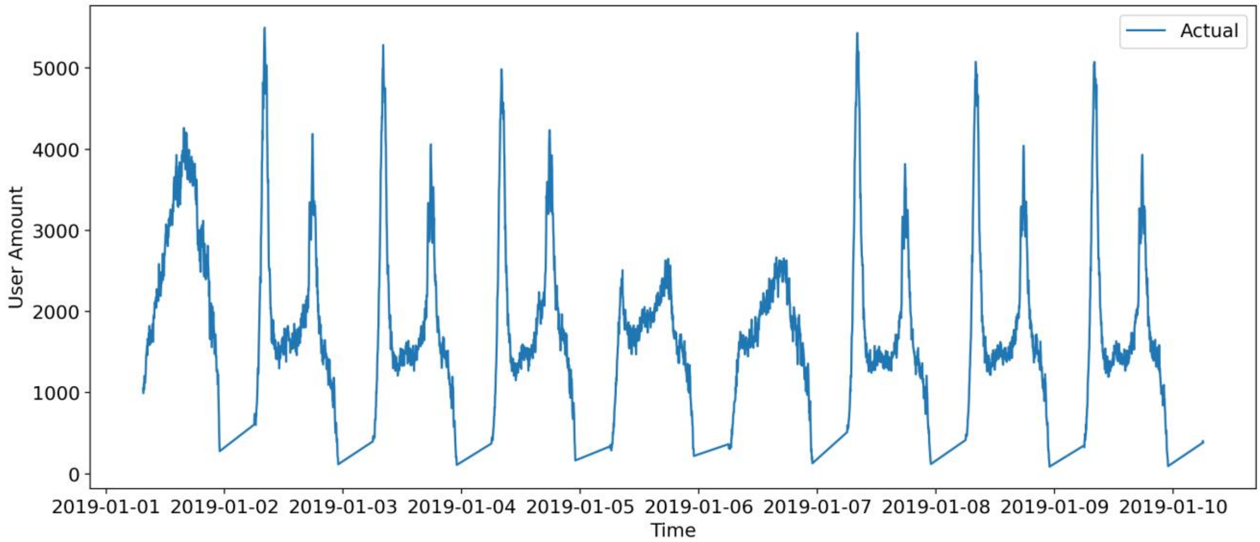

Fig. 2 Preliminary metro exit passenger flow

2.2 Random Forest

Random Forest is composed of multiple decision trees that are combined, and each decision tree does not influence the other. By training on different subsets of features and then voting or averaging their predictions together, the most representative features can be identified. The process can be visualized in Figure 3.

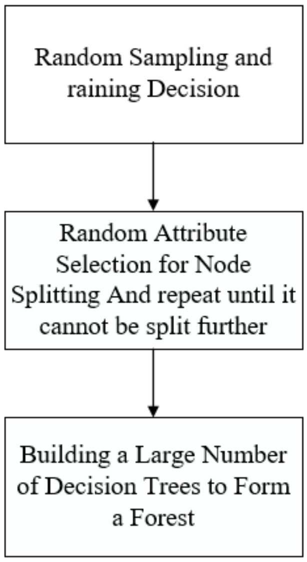

Fig. 3 Random Forest Flowchart

The construction of a random forest model is based on training data, and parameter tuning is performed using validation data. After considering factors such as the maximum depth of the decision trees, the number of trees, and the size of the feature subsets, efforts are made to achieve the best performance.

2.3 LSTM-based Model

LSTM is typically composed of a cell, an input gate, an output gate, and a forget gate. The process can be visualized in Figure 4.

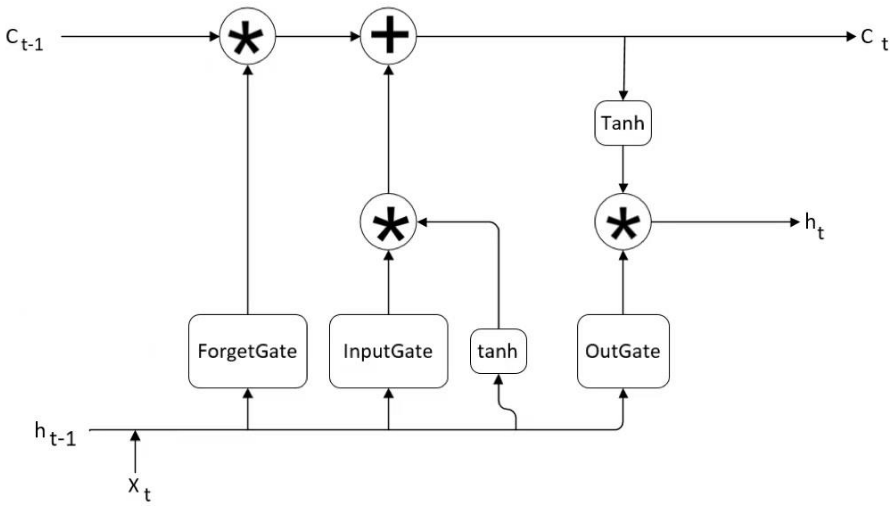

Fig. 4 LSTM Flowchart

The input gate selectively stores new information and replaces forgotten information from the forget gate. The output gate determines which information can be outputted in the current state. The forget gate is responsible for discarding information that is no longer needed. LSTM is particularly suitable for processing and predicting time series data because it can handle uncertain time lags between significant events in the sequence.

The formula for the forget gate in an LSTM:

$$f_t = \sigma\left(W_f\ [h_{t-1}, x_t] + b_f\right) \tag{1}$$

The formula for the input gate in an LSTM:

$$i_t = \sigma\ (W_i\ [h_{t-1}], x_t + b_i \tag{2}$$

$$\tilde{C}_t = \tanh W_c \ [h_{t-1}, X_t] + b_c \quad (3)$$

The formula for the output gate in an LSTM:

$$o_{t=} \sigma \ (W_O \ [h_{t-1}] + b_o \quad (4)$$

$$h_t = o_t \tanh \ (c_t) \quad (5)$$

f_t is the forget gate unit, it is the input gate unit, it is the input gate unit, O_t is the output gate unit, and h_{t-1} is the hidden layer state. W_f, W_i, W_C, W_o are weight matrices, and b_f, b_i, b_c, b_o are bias vectors. The sigmoid function and tanh function are used in the equations. These equations describe the computational process within an LSTM cell, allowing it to capture and process long-term dependencies in sequential data [11].

3. Results and Discussion

The analysis of crowd congestion in the Hangzhou subway system shows that congestion mostly occurs on specific lines or specific stations. From the analysis of the card-swiping data of the three lines, it can be seen that line B has the highest utilization rate and is also the most prone to passenger flow congestion. Its visualization is shown in Figure 5.

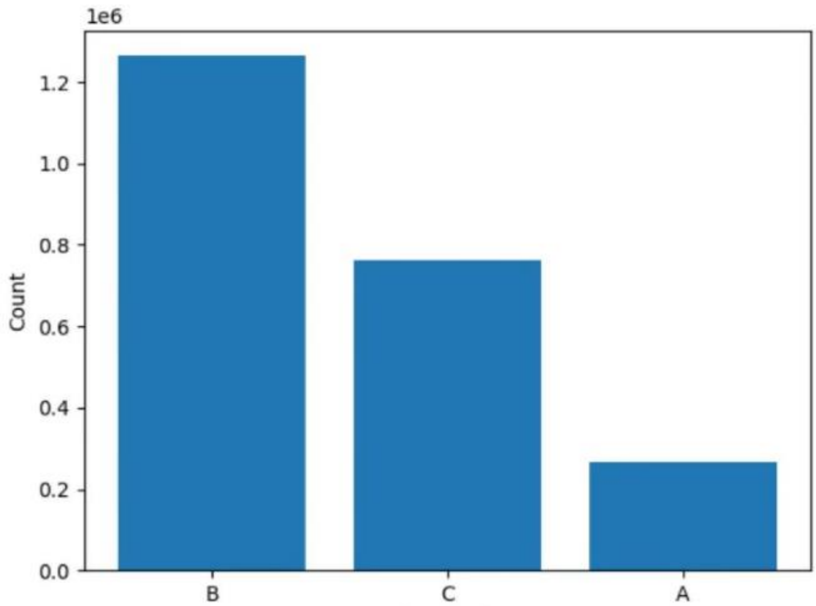

Fig. 5 The usage rates of the three subway lines in Hangzhou

After removing the A and C subway lines as well as non-operational hours during data preprocessing, the data-set size decreased from over 20 million records to over 10 million records. The removal of unnecessary data not only resulted in a significant reduction in data volume but also improved the accuracy of subsequent predictions by making them closer to the actual values.

Due to the presence of seven different features in the source data, and the desire to analyze only the data that is most correlated with the time feature, it is necessary to use random forest for feature relevance evaluation. Since the source data consists of both string and integer types, to make them comparable, some data need to be converted to integers, just like the other features.

For example, the subway line data in the source data is represented as strings 'A', 'B', and 'C', while this study assigns them corresponding numeric values of 1, 2, and 3 for comparison. By applying the random forest model with time as the index feature, a comparison is conducted among lineID, stationID, deviceID, status, userID, and the other six features. The comparison results indicate that lineID is relatively more important, implying a higher correlation with time. The comparison results are shown in Figure 6.

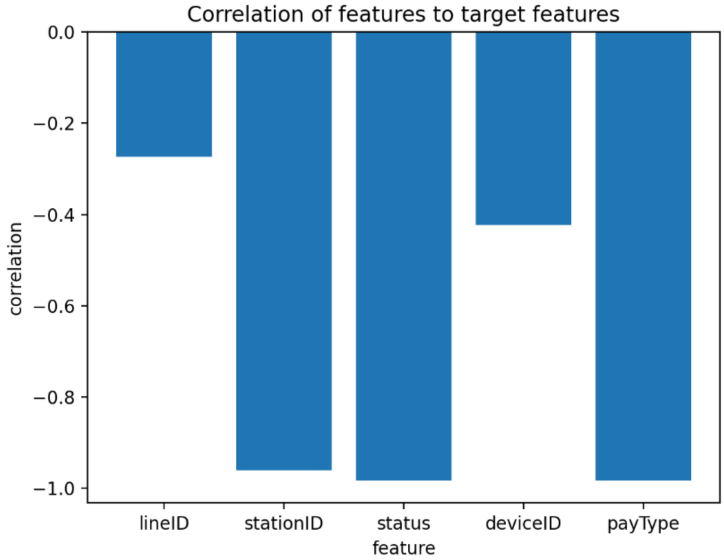

Fig. 6 Comparison chart of correlation with target features

Based on the correlation analysis, the time and lineID were selected as the prediction targets. However, since the time data was originally recorded with second-level precision, which is too granular for general short-term traffic flow prediction, the time data was down-sampled in this study. The time-frequency was reduced to 3 minutes for prediction purposes.

Based on the actual values of the subway data for 10 days, it can be observed that the graph exhibits both single-peaked and double-peaked patterns. This is because people's travel patterns vary depending on the time. On January 1st, 2019, which is the third day of the New Year, there is a significant number of passengers choosing to return in the afternoon. As a result, the passenger flow reaches its peak between 14:00 and 18:00, leading to a single-peaked pattern in the graph. Passenger flow on January 1st is shown in the Figure 7.

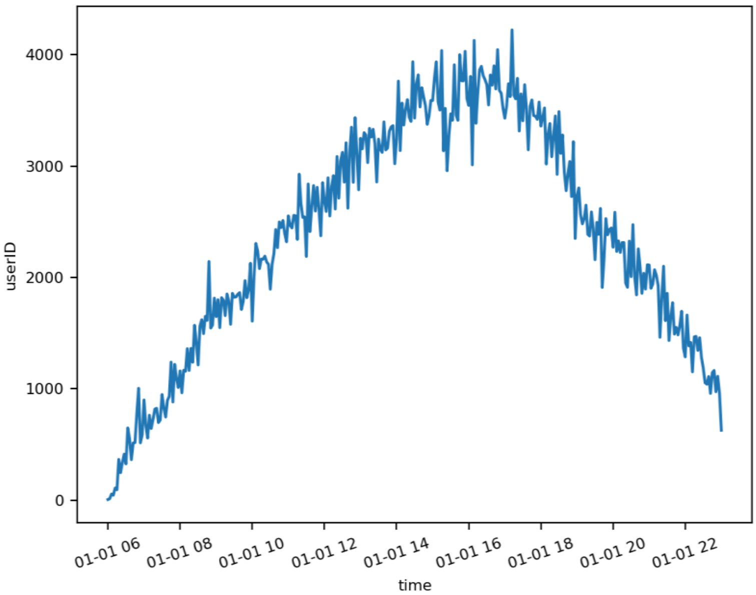

Fig. 7 Passenger flow on January 1st

Furthermore, it can be observed that the actual passenger flow curve on January 5th and January 6th is also a single-peaked graph. This is because it is the weekend and a rest day, so it can be inferred that the passenger flow on those holidays exhibits a single peak characteristic. Looking at January 3rd, which is a typical working day, most people have regular commuting patterns during working weekdays. Therefore, it can be observed that around 8 AM and 6 PM are the peak hours for subway travel. The passenger flow on January 3rd is shown in the Figure 8.

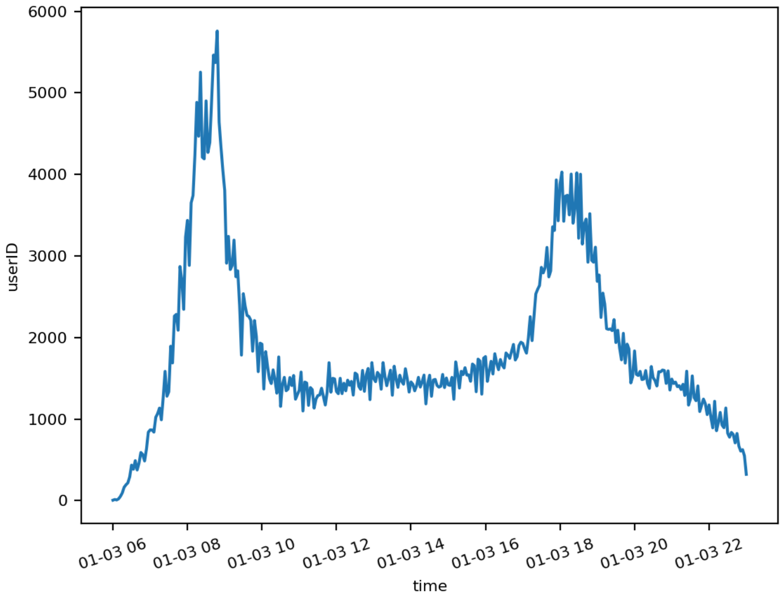

Fig. 8 Passenger flow on January 3rd

After understanding the general characteristics of the graph, only the inbound passenger flow during working hours on Subway Line B for 10 days is selected for short-term passenger flow prediction using LSTM. The parameters adjusted after importing the data into LSTM are shown in Table 3.

Table 3. Parameter Tuning

status	learning_rate	Batch_Size	Epochs
Entry	0.001	15	200
Exit	0.001	50	200
status	learning_rate	Batch_Size	Epochs

After parameter configuration, the predicted curve closely aligns with the actual data trend, and the simulation results for the inbound data are shown in Figure 9.

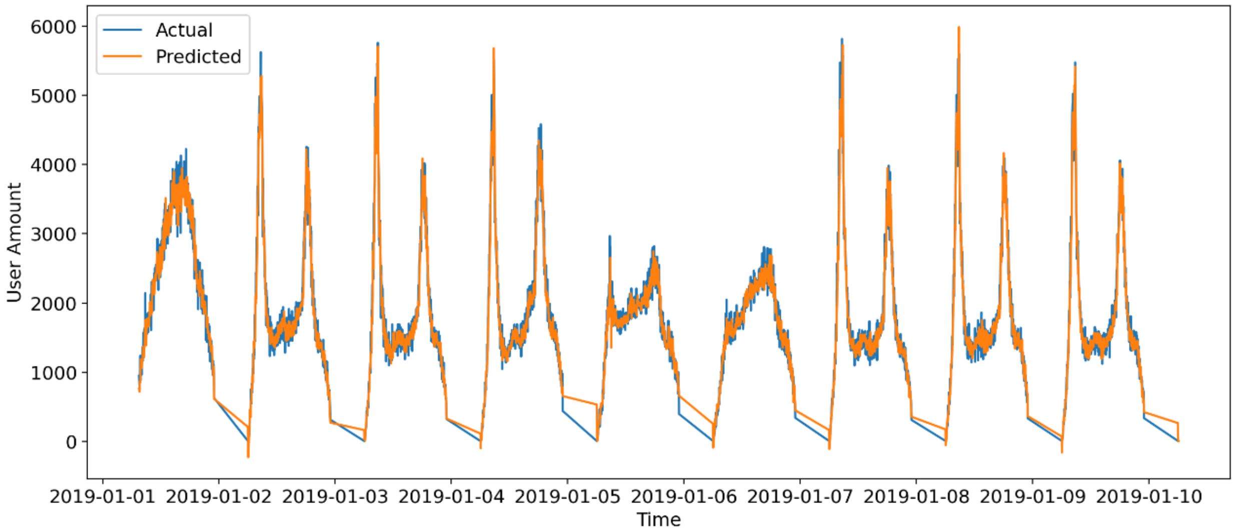

Fig. 9 Entry prediction results

Similarly, after importing the exit data into LSTM for modeling and parameter tuning, predictions can be made for the outbound passenger flow. As shown in Figure 10, the actual trends of both inbound and outbound passenger flows are nearly identical, which also indicates that the alternating frequency of passengers entering and exiting the station is normal. Moreover, the prediction results for both types of flows are highly accurate.

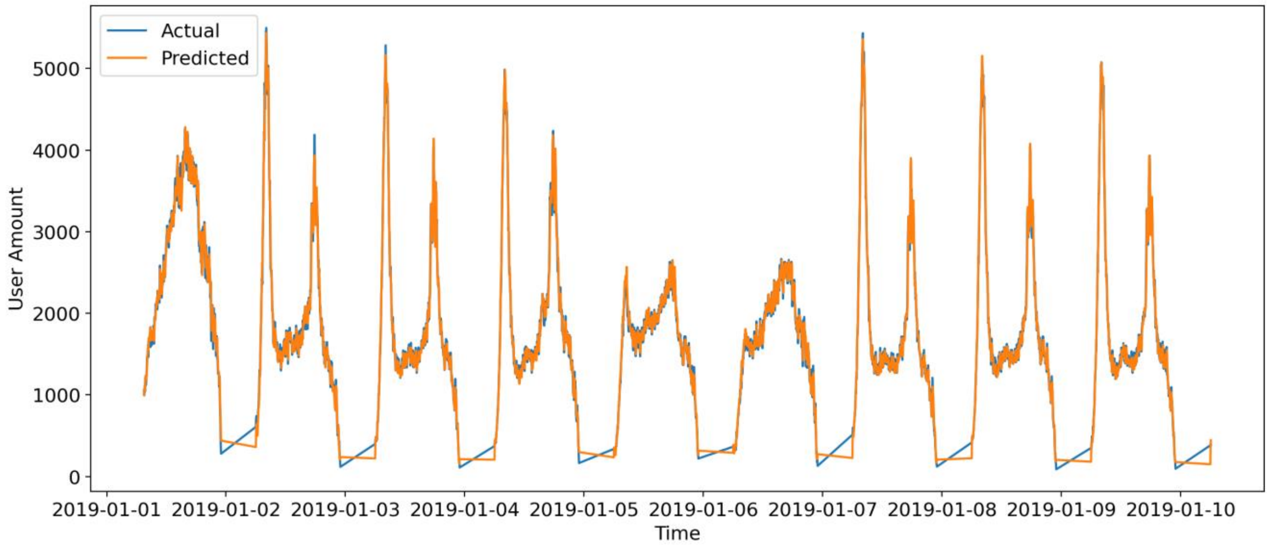

Fig. 10 Exit Prediction Results

Therefore, it can be observed from the subway entry and exit passenger flow prediction graphs that on weekends and holidays, the passenger flow exhibits a single-peak pattern, while on workdays, it shows a double-peak pattern. Moreover, it can also be seen from the graph that the highest passenger flow during workdays occurs around 8 AM and 6 PM, which is consistent with the previous estimation of actual passenger flow. The simulated data trend generated by the LSTM model exhibits peak values at almost the same locations, indicating that the model effectively captures the overall pattern of subway traffic flow. This reflects the initial correctness of selecting the LSTM model for passenger flow prediction.

4. Conclusion

This study aimed to explore the prediction of short-term subway traffic flow using an LSTM model. The dataset used for this research was obtained from the Hangzhou subway system. The data preprocessing phase primarily involved identifying and handling missing or null values. The preliminary analysis involved utilizing a random forest model to determine the most correlated feature with time, which was found to be lineID. Based on this finding, specific lineIDs were selected as filtering criteria. After preprocessing and preliminary analysis, the relationship between time and userID was examined, and the data were down-sampled for further analysis and prediction. According to the prediction results, the use of lineID classification combined with the LSTM model effectively simulates the actual subway passenger flow.

References

[1] Cheng Qiyun, Sun Caixin, Zhang Xiaoxing, et al. Short-Term load forecasting model and method for a power system based on complementation of neural network and fuzzy logic. Transactions of China Electrotechnical Society, 2004, 19(10): 53-58.

[2] Chang, H. Short-Term Passenger Flow Prediction for Subway Based on LSTM Neural Network. Xi Jing University, 2022.

[3] Wang, J. Research on Short-Term Passenger Flow Prediction for Urban Rail Transit Based on Deep Learning Models. (Doctoral dissertation). Beijing Jiaotong University, 2020.

[4] Lv, H., Feng, Q. A Review of Random Forest Algorithm Research. Journal of Hebei Academy of Sciences, 2019, 36(03): 37-41.

[5] Wang, Q. Research on Smart Tourism Based on Big Data Analysis. (Doctoral dissertation). Changchun University of Technology, 2021.

[6] Feng, B. Research on Short-Term Passenger Flow Prediction for Urban Rail Transit Based on CNN-LSTM Combination Model. East China Jiaotong University, 2020.

[7] Zhang, Z. Research on Short-Term Traffic Flow Prediction Based on Chaos Analysis and MEA-LSTM Model. Chang'an University, 2021.

[8] Peng K, et al. Research on Passenger Flow Prediction of Railway Stations Based on Improved LSTM Model. Railway Transport and Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[9] Zhang, J. Research on Short-Term Passenger Flow Prediction Methods for Urban Rail Transit Network. Beijing Jiaotong University, 2021.

[10] BeAmjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[11] Liu W, et al. Research on Holiday Passenger Flow Prediction in Suzhou Rail Transit. Urban Rapid Rail Transit, 2021, 34(05): 66-73.