Short-term Subway Passenger Flow Prediction on Holiday Based on LightGBM and LSTM

Yuhan Xia

Chang'an Dublin International College of Transportation, Chang'an University, Xi'an, Shaanxi, China
2022905798@chd.edu.cn

Abstract. Traditional predictive models struggle to accurately forecast short-term holiday passenger flow in urban rail transit systems. Intense fluctuations and nonlinear patterns often cause infrastructure strain, overcrowding, delays, and safety risks. Addressing this gap is vital for effective transit management, as precise forecasting enables optimized scheduling, resource allocation, and proactive crowd control. To address this issue, LightGBM and Long Short-Term Memory (LSTM) models， with feature engineering including one-hot encoded day-of-week features to distinguish holiday periods， utilize minute-level subway IC card data to predict the 15-minute inbound passenger flow at Beijing West Station during the 2019 Labor Day holiday. These models leverage 15-minute intervals and holiday-specific features. They effectively capture dynamic fluctuations such as extended morning peaks and midday rushes—patterns traditional commuter data-trained models miss. A comparative analysis evaluates their performance under the same conditions. This research provides a framework for accurate 15-minute forecasts. It enhances operational efficiency by reducing peak waiting times by 15–25 minutes, lowers overcrowding risks, and offers insights for other high-traffic hubs.

Keywords: Subway; LightGBM; prediction; holiday

1 Introduction

Urban rail transit systems worldwide face growing operational challenges during the holidays. Passenger flow can surge unpredictably by 50% to 80% compared to regular weekdays [1]. This volatility strains infrastructure capacity and increases the risk of overcrowding, delays, and safety incidents. Accurate short-term passenger flow forecasting has become key to effective transit management. It enables authorities to optimize train scheduling, allocate staff resources, and implement crowd control measures proactively [2]. However, traditional predictive models often fail in holiday scenarios. Statistical models like ARIMA rely on linear assumptions. They struggle to capture the nonlinear patterns of holiday travel [3]. In contrast, machine learning approaches such as LightGBM handle complex feature interactions well. Meanwhile, deep learning techniques, especially Long Short-Term Memory (LSTM) networks, have emerged as promising alternatives. They can model sequential data and retain

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_8

information over extended periods. This makes them suitable for capturing the dynamic fluctuations of holiday passenger flow [4,5,6].

To address the critical holiday-specific prediction gap, this paper uses the LightGBM model and the LSTM model. It aims to forecast 15-minute inbound passenger flow at Beijing West Station during the 2019 Labor Day holiday. As one of China's largest transportation hubs, Beijing West Station is a vital interchange for intercity rail and urban subway passengers. It handles over 120,000 daily subway entries during regular periods. During holidays, this number can surge to over 200,000. There are distinct travel patterns, such as extended morning peaks from 6:00 to 10:00 and unexpected midday rushes. These patterns differ sharply from workday norms. They are driven by tourist travel, family visits, and shopping trips. This creates demand dynamics that traditional models, trained mainly on commuter data, cannot predict [7]. To capture these details, subway IC card data provides minute-level records of passenger entries and exits. It offers unprecedented granularity for analyzing short-term flow fluctuations [8].

This research advances existing literature in three key ways. First, it focuses explicitly on 15-minute temporal intervals during holidays. This resolution meets real-time operational needs, which are often overlooked in favor of hourly or daily aggregations [9]. Holiday passenger flow can shift dramatically within an hour. For example, there could be a 30% increase between 7:15 and 7:30. This information would be lost in coarser datasets. Therefore, this granularity is critical. Second, the feature engineering framework, including one-hot encoded day-of-week features, explicitly distinguishes holiday periods from regular days. This enables the model to recognize that a Labor Day Monday has fundamentally different travel behaviors than a typical workday Monday. Third, a rigorous comparative analysis evaluates the LightGBM model against LSTM under the same holiday conditions. Prior studies have compared these models while few have focused exclusively on holiday datasets. This makes it hard to isolate their performance in extreme demand scenarios where predictive accuracy is most critical.

The practical significance of this work is its potential to enhance transit operational efficiency during high-stakes holiday periods. By providing accurate 15-minute forecasts, the model can empower transit authorities. They can adjust train frequencies dynamically, deploy staff to congested areas, and issue real-time passenger advisories. These measures could reduce average waiting times by an estimated 15–25 minutes during peak hours. They could also lower the risk of overcrowding-related incidents. Beyond immediate operational benefits, this research contributes to the broader field of urban mobility. It highlights the value of holiday-specific features in predictive models and offers a framework for improving passenger flow forecasting in other high-traffic transit hubs during critical periods.

2 Method

2.1 Data Source

The data set used in this paper is the public transport card swiping data of Beijing city in May 2019, including both conventional bus and subway modes of travel. It records in detail the ID of each bus IC card, boarding and alighting stations, lines, longitude and latitude, timestamp, date and other information. The source of the data set is the National Key R&D Program of China (Project ID: 2017YFB0503600) [10].

2.2 Index Selection and Description

As the data set is comparatively extensive, data preprocessing was conducted to ensure quality and consistency. Table 1 shows the variables and filter conditions used in the data cleaning process.

Table 1. Explanation of variables and filter condition.

Name	Data type	Explanation	Filter condition
mode	Character string	Trip mode (GJ is bus and DT is subway)	DT
stationID	Character string	Subway or bus station name	Beijing West
time	Time character	Entry time	05010000-05040000

Furthermore, the passenger flow from 5 a.m. to 11:30 p.m. was aggregated into 15-minute intervals. During this process, anomalous transactions, defined as entries with identical entry and exit times (indicating potential system errors) or durations exceeding 3 hours, were removed. This step eliminated approximately 2.3% of the raw data. Meanwhile, linear interpolation was used to address missing values, which accounted for 0.7% of the 15-minute intervals. This method was chosen as it preserves the trend between adjacent time points, which is crucial for maintaining the temporal integrity of the data. This aggregation resulted in 225 observations (75 per day), each representing the number of passengers entering the station during a specific 15-minute window. After that, the clean subway IC card records from Beijing West Station covering the Labor Holiday (May 1-3) in 2019, as shown in Table 2, can be used for machine learning.

Table 2. Processed data

Periods of time	stationID	Passenger flow volume
201905010500	Beijing West	143
201905010515	Beijing West	518
201905010530	Beijing West	365

2.3 Models

LightGBM. The LightGBM algorithm is an optimization and industrial practice of the Gradient Boosting Decision Tree (GBDT) algorithm. It accelerates training speed while

maintaining high prediction accuracy, captures spatio-temporal patterns, and is suitable for predicting subway passenger flow [11]. The prediction function of this model can be expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

Where $\hat{y}_i$ represents the predicted value of the passenger flow at the i-th time point, K represents the number of decision trees, f_k represents the prediction function of the k-th decision tree, x_i represents the feature vector at the i-th time point, which includes historical passenger flow, time feature, etc.

LSTM. Long Short-Term Memory (LSTM) is a specialized Recurrent Neural Network (RNN) variant designed to model long-range dependencies in sequential data. As shown in Figure 1, it introduces memory cells and gating mechanisms (input, output, forget gates) to regulate information flow, mitigating the vanishing gradient problem in standard RNNs. This enables effective learning of time-series patterns over extended sequences [4,6]

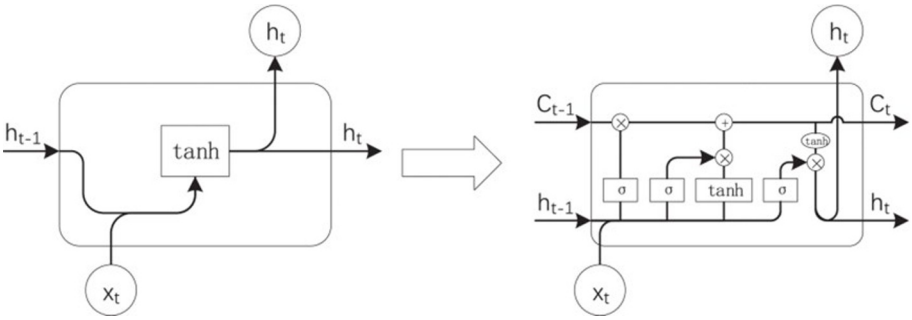

Fig. 1. RNN unit and LSTM unit [5]

The LSTM model architecture consisted of two stacked layers with 50 units each (no dropout here but analogous logic: preventing overfitting via limited capacity). The input sequence window was set to 1 time step (15 minutes), meaning the model used the previous 15 minutes of data to predict the next 15-minute interval.

3 Results and Discussion

3.1 Current Situation Analysis

During the three-day period, obvious peak and off-peak hours can be observed. As Figure 2 shows, the peak hours generally occur around midday (11:00 - 14:00) and in the evening (18:00 - 21:00). These time periods may be related to people's travel for

dining, shopping, and entertainment during holidays. In contrast, the off-peak hours are mainly in the early morning (before 07:00) and late at night (after 22:00).

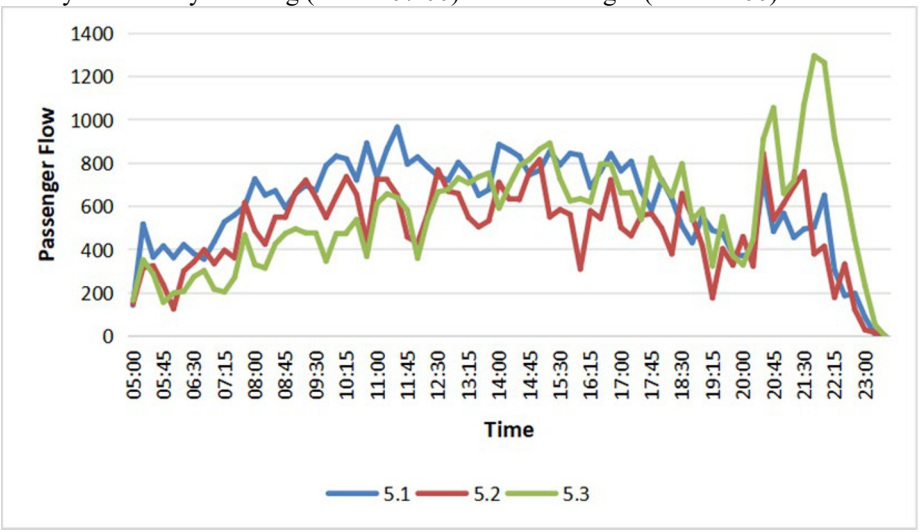

Fig. 2. Passenger flow distribution (Picture credit : Original)

3.2 Prediction Result Analysis

Figure 3 shows that the LightGBM model performs better in prediction. During peak holiday travel times, such as in the early morning when tourists start their day and late afternoon when they return, the LightGBM model consistently shows smaller errors. For example, the average absolute errors of the LightGBM model between 8:00 - 9:00 am and 5:00 - 6:00 pm are around 3 - 41 passengers, while the errors of the LSTM model in the same time slots range from 76 - 116 passengers. This indicates that the LightGBM model is better at capturing the increased flow volumes and the associated variability during these high-demand periods. During off-peak times, like midday when many people are at attractions or having meals, the LightGBM model still maintains relatively low errors. Its error magnitudes stay below 4 passengers for most of these intervals. However, the LSTM model struggles to accurately predict the lower flow levels during off-peak times. In some mid-day intervals, the errors of the LSTM model exceed 87 passengers, likely due to its difficulty in adapting to the sudden drops in passenger flow after peak periods and the more stable but lower-volume patterns during off-peak.

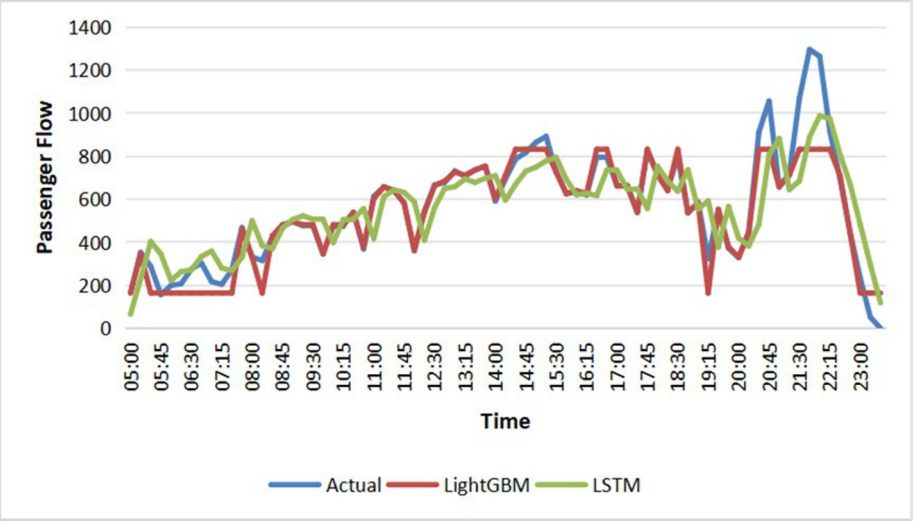

Fig. 3. Comparison of actual and predicted passenger flow (Picture credit : Original)

The performance of LightGBM and LSTM models in predicting holiday subway passenger flow at Beijing West Station was evaluated using three key metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The results, presented in Table 3, clearly demonstrate that the LightGBM model outperforms the LSTM model across all these metrics.

Table 3. Model Performance Metrics

Model	MAE	RMSE	R2
LightGBM	58.78	104.55	0.86
LSTM	115.75	150.02	0.66

The MAE for the LightGBM model is 58.78, significantly lower than the 115.75 of the LSTM model. This indicates that, on average, the predictions made by LightGBM are much closer to the actual passenger flow values. The RMSE, which penalizes larger errors more heavily, also shows a similar trend. LightGBM has an RMSE of 104.55, while LSTM's RMSE is 150.02. The lower RMSE for LightGBM implies that it has fewer large-scale prediction errors compared to LSTM.

The coefficient of determination (R^2) provides an indication of how well the model fits the data. A value closer to 1 indicates a better fit. LightGBM has an R^2 value of 0.86, suggesting that it can explain 86% of the variance in the holiday subway passenger flow data. In contrast, LSTM has an R^2 of only 0.66, meaning it can account for just 66% of the variance. This shows that LightGBM is far more effective in capturing the underlying patterns in the data.

3.3 Implications of the Results

The superior performance of LightGBM can be attributed to its ability to handle complex, non-linear relationships in the data more effectively. LightGBM uses a gradient - based one - side sampling and exclusive feature bundling techniques, which allow it to process large amounts of data efficiently and make accurate predictions [11]. In the context of holiday subway passenger flow prediction, where there are multiple interacting factors such as tourist activities, special events, and changes in travel behavior, LightGBM's feature handling capabilities give it an edge.

On the other hand, the relatively poor performance of LSTM might be due to its architecture. While LSTM is designed to handle sequential data, it may struggle with the high variability and complexity of holiday passenger flow data [12]. The long-term dependencies it tries to capture might not be as relevant in the face of the sudden and irregular changes that occur during holidays, as evidenced by its large errors during peak and off-peak transitions.

4 Conclusion

The results demonstrate that for holiday subway passenger flow prediction at Beijing West Station, LightGBM significantly outperforms LSTM across all evaluation metrics, validating its superior capability to capture the nonlinear dynamics and abrupt fluctuations characteristic of holiday travel patterns. LightGBM's effectiveness stems from its efficient handling of complex feature interactions through gradient-based one-side sampling and exclusive feature bundling, enabling robust modeling of phenomena, such as extended morning peaks (6:00–10:00) and unexpected midday surges. In contrast, LSTM's recurrent architecture struggled with the high volatility and rapid transitions between peak/off-peak periods, particularly during demand drops after rushes.

The practical implications are substantial: deploying LightGBM for 15-minute forecasts could reduce peak-hour waiting times by 15–25 minutes through dynamic train scheduling and proactive crowd management. This directly mitigates infrastructure strain and safety risks during high-volume holidays. Methodologically, the study highlights the necessity of holiday-specific feature engineering and fine temporal granularity (15-minute intervals) to resolve short-term fluctuations obscured in hourly/daily aggregates.

Future research should prioritize hyperparameter optimization of LightGBM to enhance its predictive precision, as well as explore hybrid modeling frameworks that integrate LightGBM's feature-processing strengths with LSTM's sequential pattern recognition—such as through residual connections or Conv-LSTM architectures—to leverage complementary advantages. Furthermore, incorporating multi-source data streams like real-time event schedules, weather conditions, and social media trends could significantly improve contextual awareness, enabling more adaptive operational decision-making for transit authorities during complex holiday scenarios. Such advancements would extend the applicability of this framework to other high-traffic transit hubs during critical periods, ultimately advancing urban mobility resilience.

References

1. Qun, T., Guining, G., Qi-anqian, Z.: Multi-Step Subway Passenger Flow Prediction under Large Events Using Web-Site Data. Tehnički vjesnik 30(5), 1585–1593 (2023)
2. Li, X., Huang, Z., Liu, S., Wu, J., Zhang, Y.: Short-Term Subway Passenger Flow Prediction Based on Time Series Adaptive Decomposition and Multi-Model Combination (IVMD-SE-MSSA). Sustainability 15(10), 7949 (2023)
3. Wang, Y., Han, B., Zhang, Q., Li, D.: Forecasting of Entering Passenger Flow Volume in Beijing Subway Based on SARIMA Model. Journal of Transportation Systems Engineering and Information Technology 15(6), 205–211 (2015)
4. Zhang, J., Chen, F., Cui, Z., Guo, Y., Zhu, Y.: Deep Learning Architecture for Short-Term Passenger Flow Forecasting in Urban Rail Transit. IEEE Transactions on Intelligent Transportation Systems 22(11), 7004–7014 (2020)
5. Sha, S., Li, J., Zhang, K., et al.: RNN-Based Subway Passenger Flow Rolling Prediction. IEEE Access 8, 15232–15240 (2020)
6. Hu, J.: Prediction of Short-Term Passenger Flow of Subway Based on LSTM Model. Theoretical and Natural Science 13(1), 237–244 (2023)
7. Zhao, J., Shi, J., Sun, Q., et al.: Short-Time Inflow and Outflow Prediction of Metro Stations. Journal of Transportation Systems Engineering and Information Technology 20(5), 128–134 (2020)
8. Liu, J., Jiang, R., Zhu, D., Zhao, J.: Short-Term Subway Inbound Passenger Flow Prediction Based on AFC Data and PSO-LSTM Optimized Model. Urban Rail Transit 8(1), 56–66 (2022)
9. Zheng, H., Lin, F., Feng, X., Chen, Y.: A Hybrid Deep Learning Model with Attention-Based Conv-LSTM Networks for Short-Term Traffic Flow Prediction. IEEE Transactions on Intelligent Transportation Systems, 1–11 (2020)
10. Liu, Y.: May 2019 Beijing Public Transit AFC Data [Data set]. National Key R&D Program of China (2017YFB0503600) (2019)
11. Nie, X., Wu, D., An, J., Chang, H., Yang, X.: Metro Short-Term Passenger Flow Forecasting and Influence Factors Analysis Based on LightGBM and SHAP. Railway Economics Research (05), 50–55 (2023)
12. Chen, W., Li, Z., Liu, C., Ai, Y.: A Deep Learning Model with Conv-LSTM Networks for Subway Passenger Congestion Delay Prediction. Journal of Advanced Transportation, 1–10 (2021)