Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Subway Passenger Flow Prediction Based on XGBOOST with Weather Factors

Daiyu Qian*

School of Economics, Lanzhou University, Lanzhou, China

*Corresponding author: 320220923281@lzu.edu.cn

Abstract. Since the beginning of this century, subways have increasingly occupied a higher share of daily transportation in Chinese cities. The growing network traffic has imposed higher demands on operating companies. Therefore, accurate passenger flow prediction is crucial for operational management and precise service level matching. With the maturity of machine learning methods, there are numerous cases of using these methods for passenger flow prediction. However, the influence of weather factors has been underexplored, despite its impact on passenger travel. This paper uses an XGBoost training model to obtain data on subway passenger flow, weather, maximum temperature, and minimum temperature. It examines the influence of day of the week, weather, and temperature on subway passenger flow, while also scoring the importance of these factors. The results show that the minimum temperature is the most influential factor in the model. The final model's fit is indicated by an R-squared value of 0.79, which can achieve a certain degree of accuracy in passenger flow prediction.

Keywords: XGBOOST, Predict, Subway passenger flow.

1. Introduction

Transportation infrastructure is the foundation of economic development. In recent years, as the urbanization rate in China has increased, city sizes have grown, and road traffic has become increasingly congested. According to the "2023 China Urban Traffic Report" published by Baidu Maps, the top three cities for peak commute congestion are Beijing, Chongqing, and Guangzhou, with an average commute speed of only 25 km/h. In this context, subways have been vigorously developed as a means to accommodate large numbers of passengers. With the construction of more subway lines, China has become the country with the longest subway mileage in the world. By the end of 2023, 59 cities in mainland China had opened rail transit systems, comprising 338 lines with a total mileage of 11,224.54 kilometers, of which subway lines account for 8,543.11 kilometers. In terms of daily maximum passenger flow, the highest recorded in 2023 was on March 8, with Shanghai Metro handling 13.397 million passengers. The massive influx of passengers into stations has sharply increased the pressure on the network, making passenger flow prediction a key task for operating companies. Accurate predictions are crucial for better matching service levels, preparing emergency measures, and precisely allocating personnel to achieve cost reduction and efficiency. Successful predictions are significant for improving urban transportation's overall safety and smoothness and providing passengers with convenient services.

For this topic, Hu uses the matrix method to predict the planning and passenger flow of the Guangzhou subway network [1]. Yang proposes an improved "four-stage" subway passenger flow prediction model, aiming to simplify the model as much as possible while enhancing its practical application [2]. Cai introduced a prediction method based on the ARIMA model, which reduced the prediction error [3]. In recent years, the use of machine learning to predict passenger flow has gained favor among researchers. For instance, Dong uses an improved BP neural network to achieve prediction results, addressing the drawback of traditional neural networks potentially falling into local optima [4]. Li establishes a short-term OD passenger flow prediction model based on GCN-LSTM and proposes an efficiency and spatiotemporal equilibrium-based collaborative flow control optimization model to address the imbalances and unfairness in existing peak hour flow control strategies [5]. Shen Loutao proposes a hybrid model, Hybrid-TCAE-MGM, combining deep learning models with traffic mechanism models to address the sparsity of OD matrices, achieving a balance between utilization and exploration in different feature spaces [6].

With some subway operating companies now publishing daily passenger flow data on social media platforms, analyzing passenger flow using publicly available information has become feasible. However, previous research has rarely explored the impact of weather factors on passenger flow, primarily due to the difficulty in quantifying weather. Nonetheless, the influence of weather on passenger travel cannot be ignored. Building on previous research, this paper employs the XGBoost machine learning method to predict passenger flow and analyzes the importance of selected features to explore their impact. Utilizing web scraping techniques, data on passenger flow, weather conditions, maximum temperature, and minimum temperature for specific dates was gathered. This data then facilitated the construction of an XGBoost model. After fine-tuning various parameters to achieve optimal results, passenger flow was successfully predicted. Finally, feature importances were analyzed to determine the significance of each factor affecting the model, thereby identifying the most influential features.

2. Methods

2.1. MethodsData Processing

Since subway passenger flow is influenced by numerous factors, this paper selects four representative short-term factors: day of the week, weather, minimum temperature, and maximum temperature, aiming to enhance predictive performance.

Initially, Chongqing subway passenger flow data spanning from November 2023 to April 2024 is compiles and organizes into a tabular format. The extended time frame and expanded sample size contribute positively to model training effectiveness. The subway passenger flow data originates from the daily reports published on the official Weibo account of Chongqing Metro, ensuring dependable data quality. Weather data was sourced from the Weather Network, which archives historical weather records and temperatures for specific dates.

Utilizing Python's ‘requests’ library, daily corresponding weather conditions, maximum temperature, and minimum temperature are downloaded from the Weather Network. This data is then stores in an Excel file using the Pandas. DataFrame function. To ease model interpretation, weather conditions are quantified using Python's regular expressions and categorized into four classes: no rain, light rain, moderate rain, and heavy rain, as presented in Table 1. Specifically, no rain and heavy rain conditions were assigned numerical values of 0 and 1, respectively.

The passenger flow data was plotted over time, as shown in Fig. 1. The vertical axis depicts the daily passenger flow measured in tens of thousands of trips, while the horizontal axis corresponds to the relevant dates. Excluding the Spring Festival period in February 2024, the passenger flow exhibits regular fluctuations ranging between 3 million and 5 million trips.

Similarly, Fig. 2 illustrates the daily maximum and minimum temperatures. The vertical axis displays the daily temperature, and the shaded region signifies the disparity between the maximum and minimum temperatures. The horizontal axis denotes the corresponding dates for these temperature readings.

Table 1. Weather quantification mapping

No rain	Light rain	Moderate rain	Heavy rain
0	0.33	0.67	1

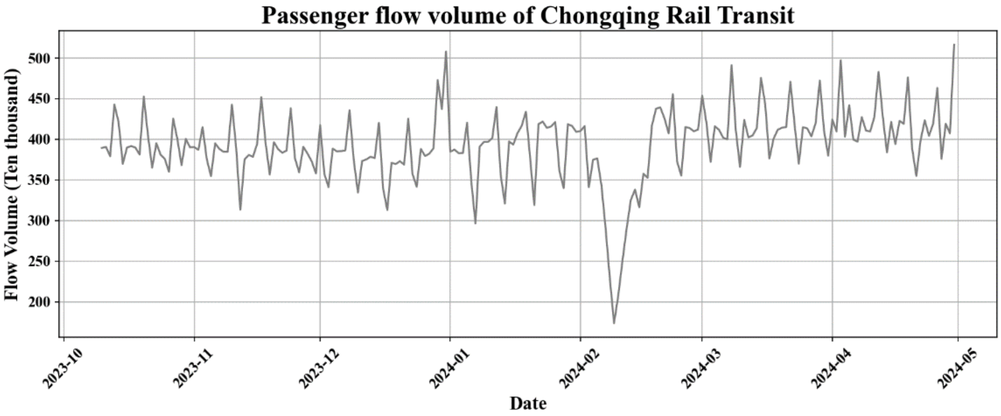

Fig. 1 Passenger flow on selected dates (Photo/Picture credit: Original).

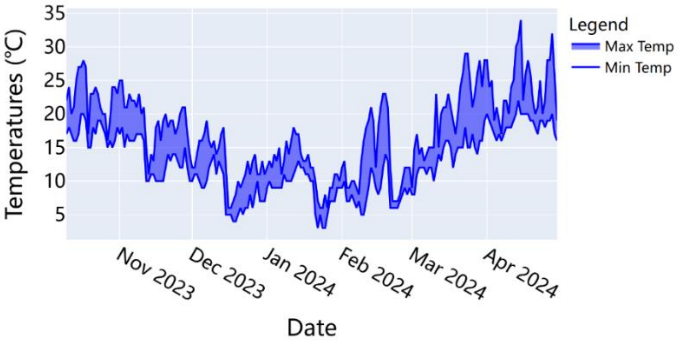

Fig. 2 Temperature on Selected Dates (Photo/Picture credit: Original).

2.2. Model Training

Steps:

Column Mapping and Renaming: Use pandas to map and rename column names from Chinese to English for ease of training the model.

Feature and Target Columns: Define the feature columns as weekday, maximum temperature, minimum temperature, and weather. Define the target column as passenger flow. Use dmatrix to convert the data format and prepare for model training.

Model Training: Train the model using the prepared data. Save the trained model locally for future use.

Training Iterations: 100 iterations.

Data Split: Use data from October 2023 to April 2024 as the training set, and data from May 2024 as the test set.

Parameter Tuning: Perform iterative parameter tuning to identify the best model parameters. See Table 2 for details of the optimal model parameters.

Table 2. Parameter settings

Parameter	Value
scale_pos_weight	0.6
max_depth	12
gamma	0.5
lambda	0.1
alpha	1

colsample_bytree	1
subsample	0.9
min_child_weight	1
random_state	1
colsample_bytree	0.9
eta	0.3
eval_metric	rmse

Due to the limited amount of publicly available data, this paper uses passenger flow data prior to May 2024 as training data. The model is then used to predict passenger flow for May 2024, and the predictions are compared with the actual passenger flow to evaluate the model's performance. First, pandas is used to read the Excel file, and the column names are mapped. The trained passenger flow prediction model is then loaded, and dmatrix and predict are used to forecast passenger flow, storing the results in flow_preds. The R2 score, calculated using sklearn's r2_score, is 0.729, indicating a good fit. Finally, Plotly is used to visualize the comparison of predictions and actual results. As shown in Table 5, the x-axis represents the date sequence and the y-axis represents passenger flow. The blue line represents the model's predicted passenger flow, while the green line represents the actual passenger flow for the month. Fig. 3 provides a clear visual representation of the prediction results.

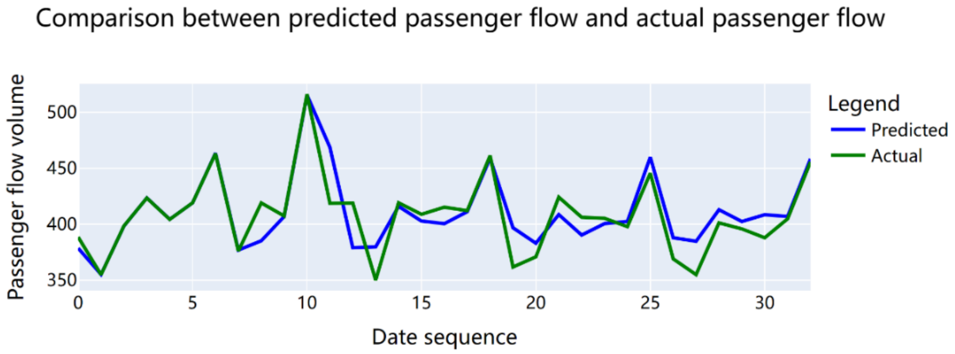

Fig. 3 Predicted vs actual passenger flow comparison (Photo/Picture credit: Original).

3. Feature Importances

Feature importance analysis is conducted to determine which features contribute the most to reducing the loss function during the model construction process. The larger the value corresponding to a feature, the greater its contribution to the model. Calculating feature importance helps understand which features have the most significant impact on the model's prediction results, providing a more intuitive view of the information gain from each feature.

Steps:

1. Set the model parameters.

2. Use `get_fscore` to obtain feature importance.

3. Use the `sorted` function to sort the importance values.

4. Use the `plt` library to output the graph.

This process scores and ranks the contribution of each feature to the model. In this analysis, a higher score indicates that the feature contributes more to the prediction results in the XGBoost model. As shown in Fig. 4, the minimum temperature and maximum temperature have the highest contributions to the model, indicating that temperature significantly affects passenger travel volume. The day of the week has a moderate impact, due to differences in passenger flow between weekdays and weekends, as passengers have different travel needs, travel intentions, and destinations. Weather has the lowest contribution value, suggesting that weather impacts passenger travel intentions less than the other three features. However, this does not mean that weather has no impact on passenger travel.

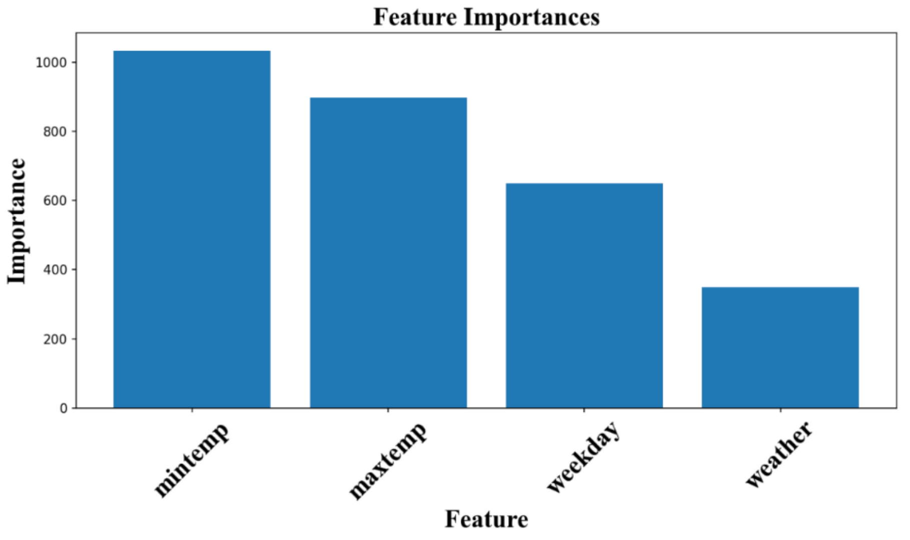

Fig. 4 Feature importance scores (Photo/Picture credit: Original).

4. Conclusion

This paper analyzes the Chongqing subway's passenger flow and constructs a model based on XGBoost. The model underwent training utilizing four key factors: day of the week, weather, maximum temperature, and minimum temperature. Through parameter adjustment, the final model attained an R-squared value of 0.729. To verify the model's effectiveness, its predictions were compared against the actual passenger flow recorded in May 2024. The predictions proved reasonably accurate, potentially attributed to the inherent periodic fluctuations observed in subway passenger flow. As illustrated in Fig. 1, the passenger flow of the Chongqing subway follows a weekly cycle, contributing to the generally satisfactory predictive performance.

In comparison to similar investigations, the incorporation of weather factors in this study had a notable impact on the ultimate training outcomes. Nevertheless, a feature importance analysis indicated that temperature significantly influences passenger travel intentions, whereas weather exerts a lesser influence. In reality, a substantial portion of commuter traffic remains unaffected by weather conditions, and the proportion of passengers influenced by weather is relatively small, explaining weather's limited role in the model. The day of the week also exhibited a minor influence, possibly due to weekdays, which constitute the majority, being predominantly composed of commuter traffic whose overall volume experiences minimal short-term variation.

XGBoost generally demonstrates its effectiveness in predicting subway passenger flow by applying machine learning to enhance prediction validity and accuracy. The model's R-squared value underscores its potential to offer valuable insights and support to subway operating companies.

However, the factors selected in this study remain somewhat limited. Elements such as occupancy rates in residential areas along subway lines and the ownership rate of private cars could also considerably impact subway passenger flow. Furthermore, the model lacks an analysis of unexpected passenger surges during holidays.

Additionally, passenger flow patterns differ across various cities, necessitating specific model optimization for each city's distinct conditions. There is also a requirement to compare diverse models or enhance the XGBoost model to yield superior prediction results.

References

[1] Hu Huaying, Li Guanhua, Guo Zhisheng. Research on guangzhou metro network planning and passenger flow forecasting. Journal of Sun Yat-sen University (Natural Science Edition), 1991, (03): 91-95.

[2] Yang Yongkai, Song Rui, Li Hairong. Analysis and research on subway passenger flow forecasting model. chinese operations research society. Proceedings of the Fourth Conference of Young Chinese Operations Research and Management Scholars. Global-Link Publishing Company, 2001, 7.

[3] Cai Changjun, Yao Enjian, Wang Meiying, et al. forecasting of urban rail transit inbound and outbound passenger flow based on the product ARIMA model. Journal of Beijing Jiaotong University, 2014, 38(02): 135-140.

[4] Dong Shengwei. Research on short-term passenger flow forecasting method for rail transit based on improved bp neural network. Beijing Jiaotong University, 2013.

[5] Li Minghui. Collaborative peak limiting strategy for subway lines based on od passenger flow prediction. Beijing Jiaotong University, 2023.

[6] Shen Loutao. Short-term od passenger flow prediction for subways based on deep learning. Zhejiang University, 2024.