ARIMA Analysis of Metro Ticket Reservation Data Via Machine Learning for Passenger Traffic Forecasting in Metro Systems

K Naresh[1], Uppara SaiKumar[2]

[1]Assistant Professor, Department of MCA, Annamacharya Institute of Technology and Sciences, Tirupati, Andhra Pradesh, India.

[2]Postgraduate, Department of MCA, Annamacharya Institute of Technology and Sciences, Tirupati, Andhra Pradesh, India.

How to Cite this Article:

SaiKumar, U. (2026). ARIMA Analysis of Metro Ticket Reservation Data Via Machine Learning for Passenger Traffic Forecasting in Metro Systems. International Journal of Creative and Open Research in Engineering and Management, <i>02</i>(04). https://doi.org/10.55041/ijcope.v2i4.064

https://doi.org/10.55041/ijcope.v2i4.064

Abstract

The need for effective public transport systems, especially metro rail networks, has greatly expanded due to the fast growth of urbanisation. Reducing congestion, increasing service quality, and optimising resource allocation all depend on accurate passenger flow forecasts. A machine learning-based method for forecasting passenger flow in metro systems is presented in this study.The suggested solution makes use of past passenger data, including time-based characteristics like date, time, peak hours, and seasonal fluctuations. To deal with missing values, normalise data, and identify pertinent features, data preparation techniques are used. To examine temporal patterns and predict future passenger demand, machine learning models like Decision Trees, Long Short-Term Memory (LSTM) networks, and Linear Regression are used. Metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are used to assess the models' performance. The results of the experiments show that the suggested method offers precise and trustworthy passenger flow forecasts, facilitating improved metro operations planning and administration.By facilitating data-driven decision-making, this study demonstrates how machine learning approaches might enhance the sustainability and efficiency of urban transit systems.

Keywords

Passenger Traffic Forecasting, Metro Systems, Machine Learning, Time Series Analysis, Demand Prediction, LSTM, Urban Transportation

I. Introduction

Rapid urban population expansion has increased demand on public transport systems, especially metro rail networks, which are essential for sustainable and effective urban mobility. Transportation officials face a significant difficulty in controlling passenger flow as cities grow. For scheduling to be optimised, traffic to be reduced, the passenger experience to be enhanced, and resources to be used effectively, accurate passenger traffic prediction is crucial. The intricate and dynamic patterns found in real-world transport networks cannot be adequately captured by traditional methods of passenger traffic prediction, which frequently rely on manual observations or basic statistical techniques. Accurate forecasting is difficult because of the numerous factors that affect passenger demand, including time of day, weekdays versus weekends, seasonal variations, special events, and socioeconomic situations. More complex methods for examining vast amounts of historical data and spotting hidden patterns have surfaced with the development of data analytics and machine learning. Temporal dependencies and trends in passenger flow data can be learned using machine learning algorithms, especially time series forecasting methods. In terms of forecasting future demand, algorithms like Linear Regression, Decision Trees, and deep learning models like Long Short-Term Memory (LSTM) networks have demonstrated encouraging outcomes. The goal of this research is to create a framework for passenger traffic forecasting in metro systems that is based on machine learning. The suggested approach seeks to generate precise demand estimates that can help transportation authorities make wise judgements by utilising historical data and sophisticated predictive models. The deployment of such intelligent forecasting systems enhances the sustainability of urban transportation networks, improves crowd control, and increases operational efficiency.

II. Dataset

The historical passenger traffic data gathered from metro systems over a predetermined time period make up the dataset used in this study. Time-series records that document changes in passenger traffic over several stations and time periods are part of it. Attributes including date, time, day of the week, station identifiers, number of passengers arriving and departing, peak and non-peak hour indicators, and seasonal or holiday statistics are usually included in the dataset. These characteristics offer insightful information on passenger behaviour and aid in the comprehension of temporal patterns like weekly cycles, daily trends, and seasonal variations. The information might also contain outside variables that affect passenger demand, like the weather or special occasions. Before using machine learning models, Preprocessing procedures for the dataset include addressing missing values, eliminating inconsistencies, and normalising numerical features. To enhance the model's capacity to capture temporal dependencies, time-based features are extracted, such as the hour of the day, day of the week, and month. After that, the dataset is organised into input-output pairs that are appropriate for forecasting time series. In order to forecast future passenger flow, supervised learning models use past passenger counts as input features. To properly assess model performance, the data is separated into training and testing sets. The machine learning algorithms' ability to identify patterns and predict passenger traffic in metro systems is greatly aided by this dataset.

III. Methodology

In order to accurately forecast demand, the suggested framework for passenger traffic forecasting in metro systems is built as a complete time-series prediction pipeline that combines data collection, preprocessing, feature engineering, model building, and deployment. First, time-stamped records of passengers' arrivals and departures from various stations are gathered from metro systems. The gathered data may have discrepancies, noise, and missing numbers because it is sequential by nature. In order to increase the quality of the data, outliers are found and eliminated during the preprocessing stage, and missing values are addressed using interpolation or imputation techniques. To guarantee consistency across numerical attributes, feature scaling strategies like Min-Max normalisation are used. To identify time-based patterns, temporal features such the hour of the day, day of the week, month, and seasonal indicators are retrieved during the feature engineering stage. Furthermore, prior time-step passenger counts are used to construct lag features,

International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 04 April-2026 | Impact Factor: 3.5

which allow the model to learn dependencies on prior data. To identify patterns and seasonality in the data, rolling statistics like standard deviations and moving averages are also calculated.

Predictive models are then trained using the transformed dataset. Both deep learning models and conventional machine learning algorithms are used. While sophisticated models like Long Short-Term Memory (LSTM) networks are used because of their capacity to represent long-term temporal dependencies in sequential data, baseline approaches like Linear Regression and Decision Tree models are employed. To maximise model performance, hyperparameter tweaking methods like grid search and random search are used.To maintain temporal integrity, the dataset is divided into training and testing sets according to time order.

To guarantee robustness and prevent data leakage, cross-validation methods unique to time-series data are employed, such as rolling or walk-forward validation. Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), which measure the discrepancy between expected and actual passenger counts, are among the error metrics used to assess model performance. Ultimately, the top-performing model is used for predicting in real-time or almost real-time. The technology continuously predicts future traffic by using recent passenger data as input. Metro authorities can use these forecasts for resource allocation, crowd control, and dynamic scheduling. This approach offers a clever and scalable way to raise the effectiveness and dependability of urban transit networks.

IV. Output

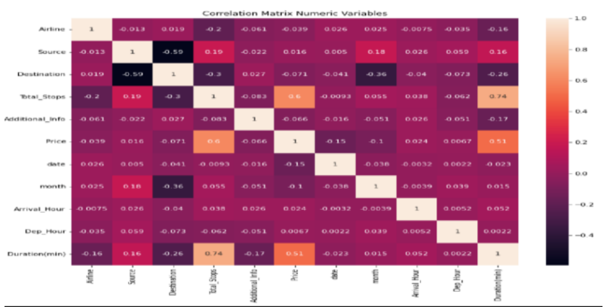

Fig: Correlation Matrix

The correlation matrix of features utilised in the passenger traffic forecasting system is shown in the heatmap above. It illustrates the relationships between several variables, including time-related traits, operational aspects, and traffic-related elements. The heatmap's values all fall between -1 and +1, where:

• Positive values show that both characteristics rise together. • An inverse association is shown by negative values.

• Weak or nonexistent relationships are indicated by values near 0.

Certain characteristics have a significant impact on the target variable (passenger traffic or similar statistic), according to the heatmap. For instance, variables like Total Stops and Duration (0.74) show a high positive correlation, which means that when one rises, the other rises as well. This illustrates how passenger flow is impacted by longer journey times or peak operational conditions in the context of your project. Similar to this, moderately favourable correlations like Price with Duration (0.51), which your project interprets as demand-related patterns, show that passenger traffic rises during specific time periods, such as busy times or peak hours. Inverse associations, such as lower passenger flow during off-peak or low-demand times, are indicated by negative correlations (e.g., -0.59). The majority of other characteristics have weak correlations, indicating that while their individual direct impact is minimal, they may nonetheless contribute to the forecasting model as a whole.

V. Conclusion

Using historical and time-related data, this study proposed a machine learning-based method for predicting passenger flow in metro systems. Passenger flow is greatly influenced by temporal and operational elements including duration, time intervals, and related properties, as demonstrated by the correlation heatmap analysis, which offered insightful information about the relationships between various features.The findings show that some characteristics have significant positive connections, suggesting their significance in precisely forecasting passenger demand. The machine learning model can more successfully identify patterns and trends in the data thanks to these connections, which improves forecasting performance. When incorporated within the model, features with weaker correlations nevertheless contribute to the total prediction even though they are less significant on their own.All things considered, the suggested approach effectively identifies the underlying trends in passenger traffic and generates accurate forecasts. This can assist metro authorities in better allocating resources, controlling traffic during peak hours, and optimising timetables. To improve forecast accuracy and system robustness, future research can concentrate on adding more real-world elements including weather, special events, and real-time data.

Reference

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] S. Raschka and V. Mirjalili, *Python Machine Learning*, 2nd ed. Birmingham, U.K.: Packt Publishing, 2017.

[5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2012.

[6] M. Zhao, X. Liu, and Y. Wang, "Passenger flow forecasting in metro systems using machine learning methods," *Transportation Research Part C: Emerging Technologies*, vol. 95, pp. 109–125, 2018