Optimizing Public Transit Through Machine Learning-Based Passenger Forecasting: A Review

Samuel Sword
March 20, 2025

1. Abstract

Accurate ridership forecasting is vital for optimizing transit operations and policy decisions, yet traditional methods struggle with complex spatial-temporal dependencies. This study reviews machine learning (ML) techniques for predicting transit ridership, comparing classical ML models like Random Forest and Stacking Ensemble with deep learning approaches such as LSTM and Conv-GCN. These models utilize diverse data sources, including smart card and weather data, and are evaluated using RMSE, MAE, and MAPE. Results show that ensemble models excel in long-term forecasting due to interpretability and efficiency, while deep learning models capture intricate patterns for short-term predictions. Methods like AP-SVR address data fluctuations, and Deep-GAN improves accuracy by handling data imbalance. While deep learning generally outperforms classical ML, ensemble methods remain strong for long-term forecasting. The findings highlight ML's ability to enhance ridership prediction, with ensemble models supporting long-term planning and deep learning approaches improving real-time forecasting. Challenges such as data quality, computational costs, and interpretability must be addressed for broader adoption. This research demonstrates the potential of ML to optimize transit systems and support sustainable urban mobility, with future work focusing on hybrid models that combine classical ML and deep learning strengths.

2. Introduction

Public transit systems play a critical role in urban mobility by providing efficient and sustainable transportation for millions of passengers daily. Accurate ridership prediction is essential for optimizing transit planning, resource allocation, and service reliability. Traditionally, transit agencies have relied on conventional statistical and econometric models, such as autoregressive integrated moving average (ARIMA) and ordinary least squares (OLS) regression, to forecast ridership patterns (Zhao et al., 2019). While these models offer interpretable insights and are effective for linear relationships, they often struggle to capture the complex, nonlinear, and dynamic nature of urban transit demand, leading to suboptimal forecasting accuracy (Li et al., 2019).

The limitations of traditional methods stem from their inability to integrate diverse data sources and effectively model spatial and temporal dependencies. For instance, classical regression models typically assume stationary relationships between variables and fail to account for seasonal variations, sudden demand fluctuations, and external influences such as weather conditions, urban land use policies, and socioeconomic factors (Toque et al., 2017). Furthermore, these methods often require extensive manual feature engineering and struggle with large-scale, high-dimensional datasets generated from modern smart card transactions and real-time transit monitoring systems (Fu et al., 2022).

To address these challenges, machine learning (ML) techniques have emerged as powerful alternatives, offering improved predictive accuracy and adaptability to complex ridership patterns. ML methods, including supervised learning algorithms such as Random Forest (Toque et al., 2017), Support Vector Regression (Li et al., 2019), and deep learning architectures like Long Short-Term Memory (LSTM) networks (Liyanage et al., 2022), provide robust solutions for capturing nonlinear dependencies and leveraging multi-source data. Additionally, ensemble approaches, such as stacking models (AlKhereibi et al., 2023), and hybrid techniques, such as Affinity Propagation-based SVR (Li et al., 2019), enhance prediction stability and generalizability. Recent advancements in deep learning, particularly the integration of Graph Convolutional Networks (GCN) with Convolutional Neural Networks (CNN), further enable transit planners to incorporate spatial and temporal dependencies in their forecasts, significantly improving short-term demand predictions (Zhang et al., 2020).

By leveraging ML techniques, transit agencies can make data-driven decisions to enhance operational efficiency, optimize resource allocation, and support transit-oriented development policies. The growing accessibility of high-resolution transit datasets and computational advancements underscore the necessity of ML-based approaches in shaping the future of urban mobility. This paper explores various ML methodologies applied to transit ridership prediction, highlighting their advantages over traditional models and their implications for sustainable urban planning.

3. Methods

3.1. Classical Machine Learning

Predictive Machine Learning Algorithms for Metro Ridership Based on Urban Land Use Policies in Support of Transit-Oriented Development

In this study, AlKhereibi et al. (2023) predict ridership in metropolitan cities, utilizing 12 supervised ML regression algorithms: ridge regression, lasso regression, elastic net, k-nearest neighbor, support vector regression, decision tree, random forest, extremely randomized trees, adaptive boosting, gradient boosting, extreme gradient boosting, and stacking ensemble learner. After hyperparameter tuning and performance metric analyses, the most effective model in terms of predictive power was the Stacking Ensemble, which combined xgBoost, Gradient Boosting Trees, and Adaptive Boosting with a Linear SVR meta-model.

Forecasting Bus Passenger Flows by Using a Clustering-Based Support Vector Regression Approach

Li at al. (2019) introduce a novel forecasting model, Affinity Propagation-based Support Vector Regression (AP-SVR), to predict bus passenger flows by addressing challenges like high fluctuations, nonlinearity, and periodicity. The model combines unsupervised techniques, namely Affinity Propagation (AP) clustering, to partition data into similar groups and supervised Support Vector Regression (SVR) optimized with Particle Swarm Optimization (PSO) to forecast flows for each cluster. Tested on real data from a Guangzhou bus line, the AP-SVR model outperforms other methods (KFCM-SVR, SVR, BPNN, SARIMA) in accuracy and stability, as measured by Mean Absolute Percentage Error (MAPE) and Variance of Absolute Percentage Error (VAPE).

Individual mobility prediction using transit smart card data

Zhao, Koutsopoulosb, and Zhao (2018) propose a supervised ML methodology for predicting individual mobility using transit smart card data, utilizing a regularized logistic regression for trip-making prediction and a Bayesian n-gram model for trip attribute prediction, inspired by language modeling. Tested on pseudonymized smart card data from over 10,000 users in London, the models achieve median accuracies of over 80% for trip-making prediction and varying accuracies for trip attributes: around 40% for start time, 70-80% for origin, and 60-70% for destination.

Short & Long Term Forecasting of Multimodal Transport Passenger Flows with Machine Learning Methods

In this paper, Toque et al. (2017) employ Random Forest (RF) models for both supervised long-term and short-term forecasting of passenger flows in multimodal transport systems. For long-term forecasting, the RF model outperforms basic and enhanced calendar models by leveraging its ability to generalize and avoid overfitting, achieving lower RMSE and MAE values. In short-term forecasting, RF models are used in two variants: one that predicts based on past observations of its own station or stop (RF ST), and another that incorporates past values from other stations or stops on the same network (RF INF ST). The RF INF ST model performs better for train stations and tram stops, while RF ST is more effective for bus stops.

3.2. Deep Learning

Short & Long Term Forecasting of Multimodal Transport Passenger Flows with Machine Learning Methods

Toque et al. (2017) also approach their supervised regression problem by using Long-Short Term Memory (LSTM) neural networks, for short-term forecasting of multimodal transport passenger flows. LSTM is chosen for its ability to capture long-range dependencies and avoid the vanishing gradient problem, making it effective for time series forecasting. The LSTM model is multivariate, predicting passenger counts for all stations or stops simultaneously, and is optimized using grid search and cross-validation. It is compared with Random Forest models, showing competitive performance, particularly for train stations and tram stops.

Multi-graph convolutional network for short-term passenger flow forecasting in urban rail transit

Zhang et al. (2020) introduce Conv-GCN, a deep-learning architecture combining Graph Convolutional Networks (GCN) and 3D Convolutional Neural Networks (3D CNN) for short-term passenger flow forecasting in urban rail transit. This supervised model uses a multi-graph GCN to capture spatiotemporal and topological correlations across recent, daily, and weekly passenger flow patterns, while the 3D CNN integrates inflow and outflow information to extract high-level spatiotemporal features. Evaluated on Beijing subway smart card data, Conv-GCN outperforms seven other models (e.g., LSTM, 2D CNN, ST-GCN) across 10, 15, and 30-minute intervals, improving RMSE by 9.402%, 7.756%, and 9.256%, respectively.

A Deep Learning Approach for Predicting Bus Passenger Demand Based on Weather Conditions

In order to predict bus passenger demand, Fontes et al. (2020) propose a Multilayer Perceptron (MLP) artificial neural network to predict bus passenger demand in a medium-sized European metropolitan area, considering weather conditions and temporal variables. Data from 2013, including transit ridership and weather variables (e.g., temperature, wind speed, relative humidity), were used to train the model. The results show that incorporating weather conditions observed two hours before travel improves prediction accuracy, reducing Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) by 6%. The model performs best for normal days, particularly during summer weekends, but struggles with strike days.

Predicting Hourly Boarding Demand of Bus Passengers Using Imbalanced Records From Smart-Cards: A Deep Learning Approach

Tang et al. (2023) address the challenge of imbalanced data in predicting hourly bus boarding demand using smart-card data, where positive instances (boarding events) are rare compared to negative instances (non-boarding events). The authors propose a Deep Generative Adversarial Network (Deep-GAN) to generate synthetic boarding instances, creating a more balanced dataset for training a Deep Neural Network (DNN). The model is tested on real-world smart-card data from Changsha, China, and outperforms traditional resampling methods like SMOTE and random under-sampling. Results show that balancing the dataset significantly improves prediction accuracy, with the best performance achieved at a 1:5 imbalance ratio. The study highlights the importance of addressing data imbalance in public transport demand prediction and provides a robust framework for improving model performance in similar contexts.

AI-based neural network models for bus passenger demand forecasting using smart card data

Liyanage et al. (2022) develop supervised AI-based deep learning models, specifically Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), for short-term bus passenger demand forecasting using real-world smart card data from Melbourne's Myki system. The models, which consider temporal characteristics of travel demand, were trained on data from 18 bus routes and 1,781 bus stops. The BiLSTM model outperformed other conventional models, achieving over 90% accuracy in predicting passenger demand for 15-minute, 30-minute, and 60-minute time horizons.

Short-term prediction of metro passenger flow with multi-source data: A neural network model fusing spatial and temporal features

In this study, Fu et al. (2022) present a novel neural network (NN) model for short-term prediction of metro passenger flow, utilizing multi-source data including smart card data, mobile phone data, and metro network data. The model integrates spatial and temporal features both inside and outside the metro system, addressing the limitation of existing studies that rely solely on internal data like smart card data. The proposed NN model employs long short-term memory layers to capture historical patterns and fully connected layers to understand interactions among features. Validated with data from Suzhou, China, the model demonstrates superior accuracy and stability compared to baseline models, with mobile phone data significantly enhancing prediction accuracy.

4. Results and Discussion

The reviewed studies highlight the strengths and trade-offs of different ML approaches for transit ridership prediction. Classical ML models like Stacking Ensemble (AlKhereibi et al., 2023) and Random Forest (Toque et al., 2017) excel in predictive accuracy while balancing computational efficiency and interpretability. Meanwhile, deep learning models, including LSTM (Toque et al., 2017; Liyanage et al., 2022) and Conv-GCN (Zhang et al., 2020), offer superior performance in short-term forecasting by capturing temporal and spatial dependencies.

Model success was measured using RMSE, MAE, and MAPE, with additional metrics like VAPE (Li et al., 2019) and classification accuracy (Zhao et al., 2018; Fu et al., 2022). Deep learning generally outperforms classical ML, though ensemble methods remain competitive in long-term forecasting. Techniques like AP-SVR (Li et al., 2019) effectively handle fluctuations in ridership data but require computationally expensive optimization. Different ML techniques present unique advantages and challenges. Ensemble models enhance robustness but demand extensive tuning. SVR performs well under high variability but is resource-intensive. Deep learning models, such as Conv-GCN, achieve high precision but require large datasets and significant computational power. Addressing data imbalance, as seen with Deep-GAN (Tang et al., 2023), enhances model reliability but raises concerns about synthetic data realism.

These results emphasize the importance of selecting appropriate ML models based on application needs. Ensemble models support long-term forecasting for urban planning, while deep learning approaches facilitate real-time operational adjustments. Integrating multi-source data (Fu et al., 2022) and refining techniques to manage data imbalances (Tang et al., 2023) are crucial for optimizing transit-oriented development strategies.

5. Conclusion

This paper examines the application of ML techniques for public transit ridership forecasting, emphasizing the strengths and limitations of classical ML and deep learning models. The reviewed studies demonstrate that ensemble methods, such as Stacking Ensemble and Random Forest, provide robust and interpretable solutions for long-term forecasting, while deep learning approaches, including LSTM and Conv-GCN, effectively capture complex spatial-temporal dependencies for short-term predictions. Performance evaluations highlight that deep learning models generally outperform classical ML techniques, particularly in real-time forecasting, but require large datasets and high computational resources. Despite the advancements in ML-driven forecasting, several challenges remain. Model performance is highly dependent on data quality, and issues such as data imbalance (Tang et al., 2023) and limited external data sources (Fu et al., 2022) can impact predictive accuracy. Additionally, computational costs and model interpretability remain barriers to large-scale adoption, particularly for deep learning techniques. Future research should focus on integrating diverse data sources, improving model explainability, and developing hybrid approaches that combine the interpretability of classical ML with the predictive power of deep learning. Addressing these gaps will enhance transit agencies' ability to make precise, data-driven decisions, ultimately improving public transportation efficiency and sustainability.

6. References

AlKhereibi, A. H., Wakjira, T. G., Kucukvar, M., & Onat, N. C. (2023). Predictive Machine Learning Algorithms for Metro Ridership Based on Urban Land Use Policies in Support of Transit-Oriented Development. Sustainability, 15(2), 1718. https://doi.org/10.3390/su15021718

Altıntaş, A., Davidson, L., Kostaras, G., & Isaac, M. (2022). The day-ahead forecasting of the passenger occupancy in public transportation by using machine learning. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 3–12. https://doi.org/10.1007/978-3-030-97603-3_1

C. Li, X. Wang, Z. Cheng and Y. Bai, "Forecasting Bus Passenger Flows by Using a Clustering-Based Support Vector Regression Approach," in IEEE Access, vol. 8, pp. 19717-19725, 2020, doi: 10.1109/ACCESS.2020.2967867.

Fontes, T., Correia, R., Ribeiro, J., & Borges, J. L. (2020). A deep learning approach for predicting bus passenger demand based on weather conditions. Transport and Telecommunication Journal, 21(4), 255–264. https://doi.org/10.2478/ttj-2020-0020

F. Toqué, M. Khouadjia, E. Come, M. Trepanier and L. Oukhellou, "Short & long term forecasting of multimodal transport passenger flows with machine learning methods," 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 2017, pp. 560-566, doi: 10.1109/ITSC.2017.8317939.

Fu, X., Zuo, Y., Wu, J., Yuan, Y., & Wang, S. (2022). Short-term prediction of metro passenger flow with multi-source data: A neural network model fusing spatial and temporal features. Tunnelling and Underground Space Technology, 124, 104486. https://doi.org/10.1016/j.tust.2022.104486

Liyanage, S., Abduljabbar, R., Dia, H., & Tsai, P.-W. (2022). AI-based neural network models for bus passenger demand forecasting using Smart Card Data. Journal of Urban Management, 11(3), 365–380. https://doi.org/10.1016/j.jum.2022.05.002

T. Tang, R. Liu, C. Choudhury, A. Fonzone and Y. Wang, "Predicting Hourly Boarding Demand of Bus Passengers Using Imbalanced Records From Smart-Cards: A Deep Learning Approach," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 5, pp. 5105-5119, May 2023, doi: 10.1109/TITS.2023.3237134.

Zhang, J., Chen, F., Guo, Y. and Li, X. (2020), Multi-graph convolutional network for short-term passenger flow forecasting in urban rail transit. IET Intell. Transp. Syst., 14: 1210-1217. https://doi.org/10.1049/iet-its.2019.0873

Zhao, Z., Koutsopoulos, H. N., & Zhao, J. (2018). Individual mobility prediction using transit smart card data. Transportation Research Part C: Emerging Technologies, 89, 19–34. https://doi.org/10.1016/j.trc.2018.01.022