PREDICTING BUS PASSENGER FLOW AND PRIORITIZING INFLUENTIAL FACTORS USING MULTI-SOURCE DATA: SCALED STACKING GRADIENT BOOSTING DECISION TREES

T.PADMA PRIYA[1], JALLEPALLI VAISHNAVI[2], JANGA VYSHNAVI[3], BYRI MANASA[4]

ASSISTANT PROFESSOR[1], UG SCHOLAR[2,3&4]

DEPARTMENT OF CSE, CMR INSTITUTE OF TECHNOLOGY, KANDLAKOYA VILLAGE, MEDCHAL RD, HYDERABAD, TELANGANA 501401

ABSTRACT- Accurate prediction of bus passenger flow is crucial for making informed decisions and ensuring efficient use of transit resources. Passenger flow is influenced by numerous factors related to the travel environment, and these factors can be collected from multiple sources. To achieve effective prediction, a model must harness the latent insights in this multi-source data while addressing challenges such as multicollinearity, which arises from interdependencies among variables. In response, we propose a novel model called the Scaled Stacking Gradient Boosting Decision Tree (SS-GBDT) to predict bus passenger flow using multi-source datasets. SS-GBDT is composed of two key modules: the feature-generation module and the prediction module. The feature-generation module leverages several foundational models with comparable performance to generate enhanced features from multi-source data through a stacking process. Specifically, a unique scaled stacking technique is introduced, which incorporates a quasi-attention mechanism that includes precision-based and time-based scaling, allowing the model to focus on the most relevant features. The newly generated features are then fed into the prediction module, which utilizes a Gradient Boosting Decision Tree (GBDT) model to forecast passenger flow based on the stacked data, thereby improving prediction accuracy and robustness. The SS-GBDT model was tested on data from two real-life bus routes in Guangzhou, China, demonstrating superior performance in terms of both accuracy and stability, along with an improved ability to manage multicollinearity issues inherent in multi-source data. Furthermore, SS-GBDT effectively prioritizes the influential factors impacting passenger flow, making it adaptable and scalable for integrating various influential elements in big data scenarios, thus providing a flexible tool for transit planning and management.

Index Terms— Big data, Public transit optimization, Machine learning in transportation, Feature generation module, Real-life dataset analysis, Data integration.

I. INTRODUCTION

Public transportation is a fundamental aspect of urban mobility, especially in densely populated cities, where managing the flow of passengers effectively is critical for smooth operations. Accurate predictions of passenger flow are essential for optimizing the management of transit systems, ensuring that resources such as buses, drivers, and infrastructure are used efficiently. These predictions can be divided into long-term and short-term forecasts. Long-term forecasts are typically used for system-level planning, such as designing routes, placing bus stops, and estimating future demand. On the other hand, short-term forecasts focus on operational decisions, such as scheduling and fleet management, ensuring that services meet immediate demand. Short-term, route-level passenger flow predictions are particularly valuable as they provide near-future ridership estimates, enabling transit agencies to deploy dynamic control strategies and optimize fleet allocation. With accurate short-term forecasts, agencies can adjust services in real-time, reducing overcrowding or underutilization of vehicles. Moreover, providing predictive insights into bus schedules can help passengers plan their travel more effectively, improving their overall experience and satisfaction. This not only makes public transport more attractive but also contributes to a more efficient and responsive transit system, capable of adjusting to fluctuations in demand due to factors like weather, special events, or seasonal changes. Ultimately, short-term passenger flow predictions are crucial for creating a seamless, well-managed, and customer-friendly public transportation system.

However, predicting passenger flow is inherently complex due to the multitude of factors that influence ridership. These include environmental conditions such as weather patterns and air quality, temporal dynamics like peak travel hours and seasonal variations, and external factors such as public holidays or special events. With the availability of multi-source data, including sensor readings and internet-based information, these factors can be incorporated into prediction models, but they also introduce challenges like multicollinearity, where multiple attributes are highly correlated, making it difficult to discern the individual impact of each factor. To address this, the Scaled Stacking Gradient Boosting Decision Tree (SS-GBDT) model is proposed. By leveraging the strengths of various machine learning algorithms through a specialized stacking approach, SS-GBDT effectively handles multicollinearity and complex interdependencies among features, leading to more accurate and reliable passenger flow predictions. This advanced model enhances feature extraction, disentangles interaction effects between data attributes, and ultimately provides superior predictive performance compared to traditional methods.

II. LITERATURE SURVEY

A) B. Yu, Z.-Z. Yang, P.-H. Jin, S.-H. Wu, and B.-Z. Yao, “Transit route network design-maximizing direct and transfer demand density,” Transp. Res. C, Emerg. Technol., vol. 22, pp. 58–75, Jun. 2012.

The paper titled “Transit Route Network Design-Maximizing Direct and Transfer Demand Density,” published in Transportation Research Part C: Emerging Technologies in 2012, focuses on improving the design of transit route networks. The authors aim to optimize transit systems by maximizing the density of both direct and transfer passenger demands. The research acknowledges the challenges of efficiently serving urban populations with varying travel needs, where both direct routes and transfer points (e.g., stations or hubs) play critical roles in shaping ridership patterns. The proposed model integrates these factors into a comprehensive framework to guide network design. The study suggests that by considering both the direct demand on individual routes and the potential for transfers between routes, transit systems can enhance accessibility, reduce travel times, and improve overall service efficiency. The methodology involves analyzing the interplay between demand density and network connectivity to achieve optimal performance. This research provides insights for transportation planners, offering practical strategies for designing transit systems that balance direct routes with transfer opportunities to meet passenger demands more effectively. It emphasizes the need for a systematic approach to network planning that accounts for both local and broader mobility patterns, ultimately contributing to the creation of more sustainable and efficient public transportation networks. The findings are relevant to cities aiming to improve urban mobility by redesigning their transit networks to cater to dynamic passenger needs while optimizing operational efficiency.

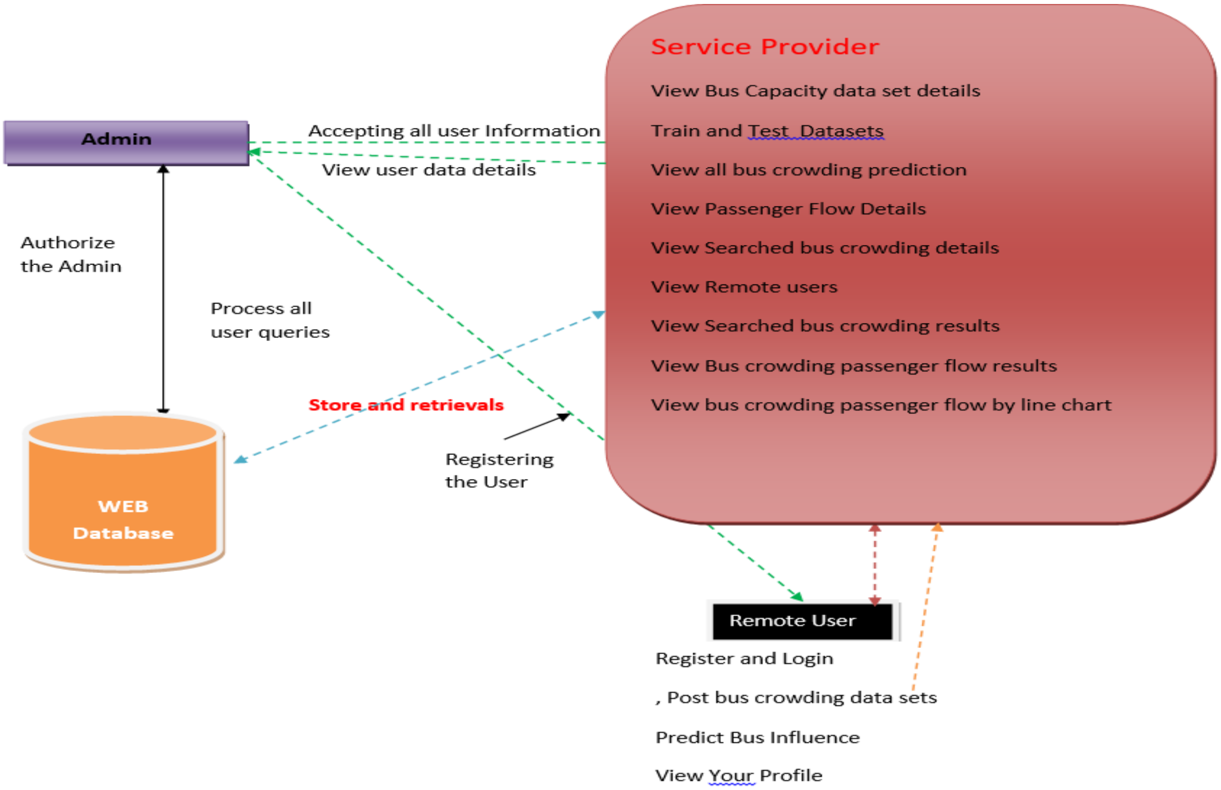
B) W. Wu, R. Liu, W. Jin, and C. Ma, “Stochastic bus schedule coordination considering demand assignment and rerouting of passengers,” Transp. Res. B, Methodol., vol. 121, pp. 275–303, Mar. 2019.

The paper by W. Wu, R. Liu, W. Jin, and C. Ma, titled "Stochastic Bus Schedule Coordination Considering Demand Assignment and Rerouting of Passengers," published in *Transportation Research Part B: Methodology* in 2019, presents a model for optimizing bus schedules in urban transportation systems. The authors address the challenges posed by fluctuating passenger demand and the need for coordinated scheduling across bus routes. The study introduces a stochastic approach, incorporating random variations in passenger demand and travel times, to improve bus schedule synchronization. It focuses on efficiently assigning passenger demand to buses and rerouting passengers to minimize delays and overcrowding. The model also accounts for the possibility of rerouting passengers between different buses or routes, enhancing system flexibility and resilience. By considering both demand assignment and rerouting, the model seeks to optimize operational efficiency while improving passenger service levels. The paper demonstrates how stochastic coordination can help reduce waiting times, improve travel reliability, and increase the overall efficiency of bus networks. This approach offers valuable insights for transportation planners aiming to create more responsive and adaptive bus systems. The methodology proposed in the paper is useful for addressing real-world issues in urban transit systems, such as demand uncertainty and operational constraints, ultimately contributing to better scheduling strategies in complex transit networks.

C) Z. Dai, X. C. Liu, Z. Chen, R. Guo, and X. Ma, “A predictive headwaybased bus-holding strategy with dynamic control point selection: A cooperative game theory approach,” Transp. Res. B, Methodol., vol. 125, pp. 29–51, Jul. 2019.

The paper by Z. Dai, X. C. Liu, Z. Chen, R. Guo, and X. Ma, titled "A Predictive Headway-Based Bus-Holding Strategy with Dynamic Control Point Selection: A Cooperative Game Theory Approach," published in *Transportation Research Part B: Methodology* in 2019, proposes a novel strategy for improving bus schedule reliability through predictive headway control. The authors introduce a dynamic control system that adjusts bus schedules based on real-time demand and traffic conditions, aiming to reduce delays and improve passenger experience. The strategy focuses on holding buses at designated control points to maintain optimal headways, thus preventing bunching and ensuring even spacing between buses. A key feature of the proposed approach is the dynamic selection of control points, which are chosen based on current conditions and the expected future demand. The model applies cooperative game theory to optimize the decision-making process among buses, ensuring that the coordination between buses minimizes delays and improves system efficiency. By integrating predictive models with dynamic control strategies, the paper demonstrates how real-time information can be leveraged to enhance the coordination of bus movements. This method allows for better management of bus fleets, reducing congestion and improving overall service reliability. The results suggest that the predictive headway-based bus-holding strategy, with its dynamic control point selection, can significantly improve the operational efficiency and effectiveness of urban transit systems.

III.PROPOSED SOLUTION


Service Provider
View Bus Capacity data set details
Train and Test Datasets
View all bus crowding prediction
View Passenger Flow Details
View Searched bus crowding details
View Remote users
View Searched bus crowding results
View Bus crowding passenger flow results
View bus crowding passenger flow by line chart
Admin
Accepting all user Information
View user data details
Authorize the Admin
Process all user queries
Store and retrievals
Registering the User
WEB Database
Remote User
Register and Login
, Post bus crowding data sets
Predict Bus Influence
View Your Profile

Modules

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Browse Data Sets and Train & Test, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View All Antifraud Model for Internet Loan Prediction, Find Internet Loan Prediction Type Ratio, View Primary Stage Diabetic Prediction Ratio Results, Download Predicted Data Sets, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT PRIMARY STAGE DIABETIC STATUS, VIEW YOUR PROFILE.

CONCLUSION

In conclusion, the proposed scalable short-term bus passenger flow prediction model, SS-GBDT, effectively addresses the challenges posed by multi-source data in transit planning and operations. By leveraging a stacking method enhanced with precision-based average weighting and time weighting, the model can generate new features, fuse data from diverse sources, and disentangle their interaction effects. This approach not only outperforms traditional machine learning models but also addresses multicollinearity, a common issue when dealing with complex multi-source datasets. The model's validation with real-world bus route data from Guangzhou, China, highlights its ability to accurately forecast passenger flow and identify the key influential factors affecting ridership, enabling better transit operations and planning.

Furthermore, the study provides valuable insights into the dynamic nature of passenger flow, revealing both similarities and differences in influential factors across different bus routes. The identification of slow-changing behaviors and effective ranges of key factors enhances the understanding of how various elements impact transit systems. The generic nature of SS-GBDT means it can be easily applied to other cities and transit systems, making it a versatile tool for improving public transportation efficiency worldwide. Future research could extend this model to other domains such as subway passenger flow prediction or bicycle demand forecasting and explore the inclusion of geographic factors to further refine predictions.

REFERENCES

[Type here]

1] B. Yu, Z.-Z. Yang, P.-H. Jin, S.-H. Wu, and B.-Z. Yao, “Transit route network design-maximizing direct and transfer demand density,” Transp. Res. C, Emerg. Technol., vol. 22, pp. 58–75, Jun. 2012.

[2] W. Wu, R. Liu, W. Jin, and C. Ma, “Stochastic bus schedule coordination considering demand assignment and rerouting of passengers,” Transp. Res. B, Methodol., vol. 121, pp. 275–303, Mar. 2019.

[3] W. Wu, R. Liu, and W. Jin, “Modelling bus bunching and holding control with vehicle overtaking and distributed passenger boarding behaviour,” Transp. Res. B, Methodol., vol. 104, pp. 175–197, Oct. 2017.

[4] Z. Dai, X. C. Liu, Z. Chen, R. Guo, and X. Ma, “A predictive headwaybased bus-holding strategy with dynamic control point selection: A cooperative game theory approach,” Transp. Res. B, Methodol., vol. 125, pp. 29–51, Jul. 2019.

[5] M. Rahman, S. Yasmin, and N. Eluru, “A joint panel binary logit and fractional split model for converting route-level transit ridership data to stop-level boarding and alighting data,” Transp. Res. A, Policy Pract., vol. 139, pp. 1–16, Sep. 2020.

[6] X. Ma, X. Zhang, X. Li, X. Wang, and X. Zhao, “Impacts of free-floating bikesharing system on public transit ridership,” Transp. Res. D, Transp. Environ., vol. 76, pp. 100–110, Nov. 2019.

[7] N. S. Ngo, “Urban bus ridership, income, and extreme weather events,” Transp. Res. D, Transp. Environ., vol. 77, pp. 464–475, Dec. 2019.

[8] P. N. E. Chee, Y. O. Susilo, and Y. D. Wong, “Determinants of intentionto-use first-/last-mile automated bus service,” Transp. Res. A, Policy Pract., vol. 139, pp. 350–375, Sep. 2020.

[9] C. Ding, X. J. Cao, and P. Næss, “Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in oslo,” Transp. Res. A, Policy Pract., vol. 110, pp. 107–117, Apr. 2018.

[10] N. G. Polson and V. O. Sokolov, “Deep learning for short-term traffic flow prediction,” Transp. Res. C, Emerg. Technol., vol. 79, pp. 1–17, Jun. 2017.

[11] R. Jeong and L. R. Rilett, “The prediction of bus arrival time using automatic vehicle location systems data,” Texas A&M Univ., College Station, TX, USA, Tech. Rep., 2004.

[12] S. I.-J. Chien and C. M. Kuchipudi, “Dynamic travel time prediction with real-time and historic data,” J. Transp. Eng., vol. 129, no. 6, pp. 608–616, Nov. 2003.

[13] A. Shalaby and A. Farhan, “Bus travel time prediction model for dynamic operations control and passenger information systems,” presented at the 82nd Annu. Meeting Transp. Res. Board, Washington, DC, USA, Jan. 2003.

[14] B. Yu, Z.-Z. Yang, K. Chen, and B. Yu, “Hybrid model for prediction of bus arrival times at next station,” J. Adv. Transp., vol. 44, no. 3, pp. 193–204, Jul. 2010.

[15] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu, “Mining smart card data for transit riders’ travel patterns,” Transp. Res. C, Emerg. Technol., vol. 36, pp. 1–12, Nov. 2013.