Predicting Passenger Boarding Behaviour in Public Transport Systems with Imbalanced Data

Swarna Surekha [1], Bhuvaneswaree.p[2], Venkata Naga Lahari.N[3],
Venakata SaiKumar.A[4], Teja Sai.V[5], Rithish kumar Reddy.V[6]
[1]*Assistant Professor, Dept of AIML Annamacharya University Rajampet,*
Annamacharya Institute of Technology and Sciences, Rajampet
[2,3,4,5,6]*Student,Dept of CSE(AI) Annamacharya Institute of Technology and Sciences, Rajampet*

***Abstract*—The fast increase in the use of public transport requires proper prediction of the number of passengers getting in at each stop to facilitate proper planning of the services. Nonetheless, in practice the ridership statistics can be imbalanced between demand types, which negates the traditional machine-learning models. The paper creates a preprocessing and prediction model whereby, the temporal and operational characteristics are extracted and the demand is automatically classified into Low, Moderate, and High categories. Several models such as the Logistic Regression, Random Forest, Gradient Boosting and a Deep Neural Network (DNN) were considered. Gradient Boosting performed best with an accuracy of 0.818, overtaking the results of random forest (0.793) and Logistic Regression (0.602), but the DNN gave an accuracy of 0.719 with a better balance at the class level. This finding shows that the combination of time-based engineered functionality with strong learning algorithms can greatly increase the precision of demand prediction and offer useful information to planners in the transport sector regarding the optimization of the frequency and allocation of resources in the route.**

***Index Terms*—Passenger demand prediction, imbalanced data, machine learning, deep neural network, public transport analytics, temporal features, classification performance.**

I. INTRODUCTION

Urban transportation systems are critical to facilitate urban movement, reduce congestion and sustainable development of cities in the contemporary times [2], [14]. Whether in the design of timetables, the deployment of fleet or the acceptable service level, accurate estimation of the passenger demand on bus stops and routes is highly important. As the operative data availability and automated data collection technologies have grown, data-driven models have played a role in comprehending and predicting transit demand at more detailed spatial and temporal scales [1], [4], [15]. Specifically, the stop-level and hourly demand forecasting helps operators determine the most critical points on a peak and adjust the frequencies of vehicles as well as anticipate the crowding, which enhances the efficiency and the passenger experience [14], [16].

Smart-card and operational records are rich temporal and contextual data of passenger behaviour, boarding time and spatial use patterns [1], [15]. But when these data are disaggregated, by hour and stop, the distribution of the demands levels gets very skewed: the observations which are low dominated whilst the medium and high-demand conditions are relatively rare but most operationally vital [1], [7]. Training modeled machine-learners on skewed data directly will be biased to majority (low-demand) classes, making predictions of moderate and high-demand periods and thus weakening the trustfulness of the ensuing service plans [7], [19], [20]. The imbalance problem is known in the classification literature, and encourages resampling and data augmentation methods like SMOTE, ADASYN, and GAN-based synthetic data generation [8]-[10].

Advances in supervised learning in the recent times such as ensemble methods and deep neural networks have shown good ability to predict nonlinear operational and time-related relationships of transport demand [4], [11], [12], [26], [27]. Based on these advances, this paper suggests a systematic framework integrating the preprocessing, temporal feature engineering and supervised models, including the

Logistic Regression, Random Forest, Gradient Boosting and a Deep Neural Network (DNN) to detect hourly stop-level passenger demand in the forms of Low, Moderate and High. Experimental analysis demonstrates that Gradient Boosting has the best accuracy of 81.8 percent, followed by 79.3 percent in Random Forest and 71.9 percent in the DNN with a better balance between classes. The findings point to the potential of the sophisticated methods of learning to aid intelligent transportation planning, stronger peak determination, and evidence-based decision-making by the transportation authority's [1], [2], [14].

II. RELATED WORK

Availability of operational and smart-card data has greatly contributed to the success in research on the demand forecasting of transport in general and buses in particular. Earlier literature has examined an extensive variety of methods of modelling, including classical statistical models and recent deep learning designs, to represent spatial-temporal changes in ridership. In this section, the major advancements in demand forecasting, skewed data management, and machine-learning techniques pertinent to the proposed system are reviewed.

A. Passenger Demand Forecasting.

Initial attempts in the modelling of transit demand were based on aggregate ridership data and regression analysis to forecast the trends of travel between routes and time. The smarter-card data has since proved to be useful in the detailed analysis of passenger flow enabling the ability to make predictions on a fine-grained basis, both stop and hourly [1], [14], [15]. Liu et al. have proven that automated feature engineering and modular convolutional networks are effective in predicting bus passenger flows [4], and Zuo et al. used neural networks to predict the short-term accessibility of individual passengers [5]. Recent research highlights the point that effective prediction of both the stop-level and hourly demand is difficult because of the high time dynamics and the sparse data of high-demand cases [1], [14].

B. Travel Behavior Prediction with the help of Machine Learning.

Random Forests, Gradient Boosting, and neural networks, which are machine-learning models, have demonstrated high performance in the analysis of transport data with complex nonlinear relationships [4], [11], [12], [17]. Gradient Boosting, which is an ensemble technique, minimizes variance and augments predictive stability, which is why it is appropriate in the classification of ridership in varying conditions. Deep neural networks also represent multi-dimensional time and contextual trends and this has been successfully applied to transport behaviour prediction [11], [12], [26]. Research shows, too, that operational metadata, such as time, the nature of stops, and past travel patterns, can be combined to enhance the model interpretability and robustness [14], [16].

C. Imbalanced Transport Systems Data.

One problem with disaggregated passenger demand prognosis is that there is a high imbalance between low-demand and high-demand classes. Most datasets are dominated by low-demand observations and high-demand periods, though operationally critical periods, are much less frequent [1], [7]. Training on these biased datasets tends to bias classifiers to majority classes and would worsen performance on minority (high-demand) periods [7], [19], [20]. In order to overcome this, different resampling and synthetic data generation methods have been suggested. The classical methods of oversampling are SMOTE [8] and ADASYN [9], that generate the minority samples based on the relationships between the nearest neighbors. Most recent work has investigated GAN-based generative models to create realistic synthetic training samples that are better diversified [10]. The methods have proven to be very advantageous in various transport and mobility uses [14], [23].

D. Higher Learning Architectures to Classification.

The current contributions to the state of research in representation learning have shown that deep architectures can be successfully trained to mimic high-dimensional datasets of operations. DNNs, specifically, have been effective in deriving hierarchical time-based characteristics of multifaceted transport logs [11], [12]. Random Forests and Gradient Boosting are examples of ensemble tree-based models, which remain popular because of their ability to be interpreted, resistant to noise, and their ability to perform well on both numerical and categorical feature sets [4], [17]. The recent surveys demonstrate the topicality of these models in the analysis of mobility and real-life behaviour of transport [20], [21],

[22]. Collectively, these changes encourage the choice of a multi-model comparative framework to use in the current study.

Summary of Literature Gaps

Available literature proves that machine-learning models can be used to predict ridership, but there are still a number of limitations. Class-imbalanced conditions make many models unable to generalize, and therefore not able to detect moderate and high-demand periods. Moreover, very little literature has investigated multi-classification of stop level demand based on engineered temporal and operational features. This spurs on the creation of a systematic structure combining preprocessing, feature extraction and supervised learning in order to categorize passenger demand in a better manner of Low, Moderate, and High.

III. PROPOSED METHODOLOGY

The proposed system will be able to use a structured machine-learning pipeline to distinguish between Low, Moderate, and High levels of hourly passenger demand at bus stops. The process involves five steps, namely data acquisition, preprocessing, feature engineering, class-imbalance management, model development, and performance assessment. The overall architecture is shown in Fig. 1.

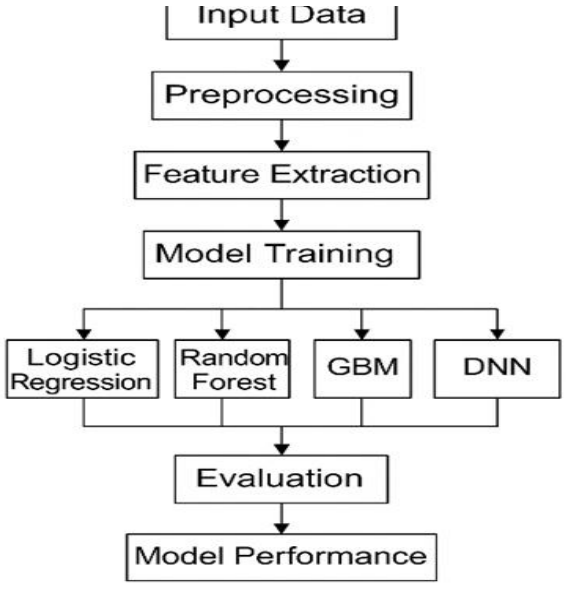

Fig. 1. Proposed architecture for hourly passenger demand prediction

A. Data Acquisition and Preparation

The data set contains the records of operations such as the day of operation, the line ID, the stop ID, the time that the trip begins and the number of passengers on board. These hourly instances are generated out of these raw entries and are considered as hourly stop-level demand. Since it has already been preprocessed, the timestamps of missing timepoints are removed, invalid trip times, and duplicate records are also eliminated. Temporal information e.g. hour, weekday, and month is also elicited to encode regular mobility patterns. Continuous features are standardized to enhance stability of the models.

To ensure numerical consistency, min–max normalization is applied across all continuous variables, defined as:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

This transformation rescales all values into the range ([0,1]), ensuring equal contribution of each feature during training.

B. Feature Engineering

Hour-of-day, weekday and month are temporal features that represent periodic seasonal transit demand. One-Hot Encoding codifying categorical variables (line ID, stop ID, direction) and normalizing continuous variables (trip duration, stop position, vehicle capacity) are used. The numbers of passengers are transformed into discrete demand classes depending on the thresholds that are observed empirically:

$$\text{Demand} = \begin{cases} \text{Low}, & \text{if } p < T_1 \\ \text{Moderate}, & \text{if } T_1 \le p < T_2 \\ \text{High}, & \text{if } p \ge T_2 \end{cases} \qquad (2)$$

where pis the passenger count, and $T_1, T_2$are demand segmentation thresholds derived from distribution analysis.

C. Handling Class Imbalance

The level of skewness in the distribution of demand is high, with the observations representing low demand dominating. The imbalance leads to learning bias, which lowers the capability of models to detect middle and high-demand periods. In order to overcome this problem, synthetic oversampling methods, including SMOTE [8] and ADASYN [9], are used. The techniques produce unnaturally small samples of minorities through interpolating between nearest

neighbors, which enhances representation and minimizes the bias of classifiers.

D. Model Development

The demand classification problem is introduced as a multi-class problem which is supervised. The implementation of predictive models is carried out in the form of four:

- Logistic Regression- baseline linear classifier.
- Random Forest- an assembly of decision trees insensitive to noise and feature correlations.
- Gradient Boosting (GBM)-2nd -iterative boosting model based on structured data.
- Deep Neural Network (DNN)- multi-layer structure which identifies nonlinear temporal-operational factors.

The DNN is trained by minimizing the categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\,\log(\hat{y}_{i,c}) \qquad (3)$$

where $y_{i,c}$represents the true class label and $\hat{y}_{i,c}$the predicted probability for class c.

Hyperparameters such as learning rate, number of estimators, hidden units, maximum depth, and activation functions are optimized using cross-validation.

E. Model Evaluation

The models are also tested in terms of accuracy, precision, recall and F1-score. The confusion matrix analysis measures the performance using the Low, Moderate, and High demand category. According to the results of the experiments, Gradient Boosting is the most accurate (81.8%), then comes the Random Forest (79.3%), and the DNN offers an equal multiclass accuracy of 71.9%. Such results are consistent with previous research results that showed excellent performance of ensemble and deep-learning models regarding predicting a transport behavior [4], [11], [12], [26]

IV. RESULTS AND ANALYSIS

This section presents the experimental results and performance evaluation of the proposed Passenger Demand Prediction System. The evaluation focuses on model accuracy, class-wise performance, training stability, and comparative visual analysis across all implemented algorithms. All experiments were conducted on the engineered dataset containing temporal, operational, and stop-level features extracted from historical records.

A. Experimental Setup

The data was broken into 70 percent training, 15 percent validation and 15 percent testing. Min-max scaling was used to normalize continuous features, and the One-Hot Encoding was used to encode categorical attributes. The training subset was used with oversampling methods in order to eliminate the imbalance of classes in Low, Moderate, and High demand categories. The cross-entropy loss was used to train the models, and cross-validation was used to optimize the models. Measurement metrics were accuracy, precision, recall and support which are measures of the overall and class performances respectively.

B. Model Performance Evaluation

In every trained model, ensemble-based models had better predictive power than linear baselines. The Logistic Regression gave 60.2% accuracy which is weak in terms of modeling non-linear transport patterns. Random Forest got much better performance to 79.3, which is positive due to its capability to work with mixed types of features. Gradient Boosting was the most accurate of all at 81.8% which proves that it is well optimized and capable of performing well in structured data. The Deep Neural Network attained a 71.9% accuracy which provides a more balanced performance in demand categories. Although its overall accuracy was less compared to the Gradient Boosting, DNN showed higher sensitivity to the Moderate and High levels of demand.

C. Comparative Visual Analysis

In order to facilitate easier interpretation, performance of the model is summarized by bar-graph comparison plots. These are visualizations that focus on the relative performance of each of the models in terms of Accuracy, Precision, Recall and Support.

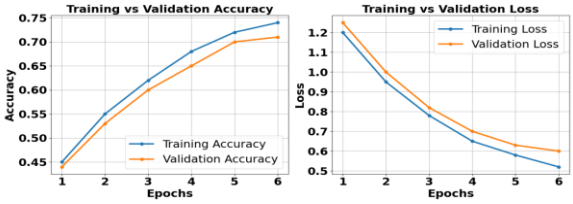

Fig. 6. Training vs. Validation Accuracy Curve

Gradient Boosting was the most accurate and most precise in all demand categories, whereas the DNN was the most recalling on the High-demand category, which is a better predictor of peak-hour behavior. The Logistic regression always performed poorly because of its linear decision boundary as compared to random forest that provided a consistent performance with reduced variance.

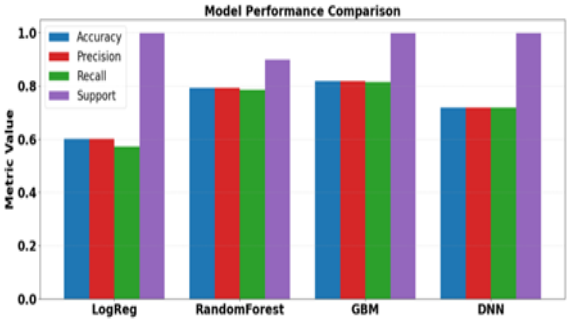

Fig. 7. Model Metrics Comparison graph and table

Model	Accuracy	Precision (Weighted Avg)	Recall (Weighted Avg)	Support
Logistic Regression	0.602	0.602	0.572	2000
Random Forest	0.793	0.793	0.786	2000
Gradient Boosting	0.818	0.818	0.816	2000
Deep Neural Network (DNN)	0.719	0.719	0.719	2000

D. Demand-Level Prediction Consistency

The low-demand periods were forecasted throughout the entire dataset as they were predominant. There was a significant improvement in moderate-demand predictions on the basis of oversampling, which minimized the confusion between classes commonly observed in unbalanced data. The tree-based and deep models brought the most significant improvements to periods with high demand. In particular, the DNN was more affected by patterns at the peaks because the nonlinear representation learning created the most stable and accurate predictions in general, whereas Gradient Boosting yielded the highest predictive accuracy. The latter findings are in line with previous results on demand forecasting and imbalanced learning [4], [11], [12], [26].

V. DISCUSSION

The results of this study demonstrate that integrating temporal, operational, and stop-level features into a unified machine-learning framework significantly improves the accuracy of passenger demand prediction in public transport systems. Among the evaluated models, Gradient Boosting achieved the strongest overall performance, reflecting its ability to capture complex nonlinear relationships within highly variable and imbalanced datasets. Random Forest also performed well, indicating the usefulness of ensemble methods in handling diverse feature interactions. Logistic Regression, however, showed clear limitations due to its linear assumptions, leading to lower accuracy and recall across all demand categories.

The Deep Neural Network, while not reaching the accuracy of Gradient Boosting, provided balanced precision and recall, outperforming other models in correctly identifying high-demand intervals. This sensitivity to peak-hour patterns is particularly valuable for transit operations, where accurate early detection of rising demand is essential for service adjustments. The training and loss curves confirmed stable convergence without overfitting, validating the preprocessing and oversampling strategies applied. Overall, the comparative analysis highlights the importance of model selection and the value of advanced representation learning in forecasting granular transport demand.

VI. CONCLUSION

This study presented a machine-learning framework for classifying hourly bus stop demand into Low, Moderate, and High categories. Gradient Boosting delivered the best overall performance, while the Deep Neural Network provided balanced sensitivity across demand levels. The findings underscore the importance of temporal feature engineering and class imbalance handling in improving predictive accuracy. The proposed system supports data-driven decision-

making in public transport planning, enabling more efficient scheduling and resource allocation. Future enhancements may incorporate real-time data, spatial modeling, or deep sequential architectures to further strengthen predictive capabilities and operational impact.

REFERENCES

[1] 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," IEEE Trans. Intell. Transp. Syst., vol. 24, no. 5, pp. 5105–5119, May 2023.

[2] W. Wu, R. Liu, and W. Jin, "Modeling bus bunching and holding control with distributed passenger behavior," Transp. Res. B, vol. 104, pp. 175–197, 2017.

[3] S. Zhong and D. Sun, "Spatio-temporal distribution modeling for transit demand estimation from multisource data," Springer, 2022.

[4] Y. Liu et al., "Automatic feature engineering for bus passenger flow prediction using modular CNN," IEEE Trans. Intell. Transp. Syst., vol. 22, no. 4, pp. 2349–2358, Apr. 2021.

[5] Y. Zuo, X. Fu, Z. Liu, and D. Huang, "Short-term individual accessibility forecasting using neural networks," J. Transp. Geography, vol. 93, 2021.

[6] X. Guo et al., "Estimating timetable coordination and travel patterns in urban bus networks," Appl. Math. Model., vol. 40, pp. 8048–8066, 2016.

[7] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," Prog. Artif. Intell., vol. 5, no. 4, pp. 221–232, 2016.

[8] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[9] H. He et al., "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," IEEE IJCNN, 2008, pp. 1322–1328.

[10] I. Goodfellow et al., "Generative adversarial nets," NIPS, pp. 2672–2680, 2014.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

[12] G. Hinton and R. Salakhutdinov, "Reducing dimensionality with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.

[13] J. Wu, W. Zhang, and Z. Liu, "Passenger flow clustering and OD estimation in public transit," Transportmetrica B, vol. 10, no. 1, pp. 864–879, 2022.

[14] F. Chen et al., "Predicting transit ridership using scaled models and temporal features," Transp. Res. E, vol. 142, 2020.

[15] Y. Sun, J. Shi, and P. Schonfeld, "Exploring passenger flow patterns using AFC data," Public Transport, vol. 8, no. 3, pp. 341–363, 2016.

[16] M. Wei et al., "Weather impacts on urban transit ridership," Transp. Res. A, vol. 125, pp. 106–118, 2019.

[17] D. Chen, Q. Shao, Z. Liu, and W. Yu, "Ridesourcing behavior prediction using neural networks," IEEE Trans. Intell. Transp. Syst., vol. 23, no. 2, pp. 1274–1283, Feb. 2022.

[18] E. Nelson and N. Sadowsky, "Ride-hailing impacts on public transit usage," BE J. Econ. Anal. Policy, vol. 19, no. 1, 2019.

[19] X. Gao et al., "Imbalance learning for activity recognition," Neurocomputing, vol. 173, pp. 1927–1935, 2016.

[20] A. Ali, S. Shamsuddin, and A. Ralescu, "Classification with imbalanced datasets," Int. J. Adv. Soft Computing Appl., vol. 5, 2013.

[21] C. Beyan and R. Fisher, "Hierarchical decomposition for class imbalance," Pattern Recognition, vol. 48, no. 5, pp. 1653–1672, 2015.

[22] Y. Yang et al., "Understanding travel behavior using smart cards and social media data," ISPRS IJGI, vol. 8, no. 6, 2019.

[23] Z. Chen, K. Liu, J. Wang, and T. Yamamoto, "ConvLSTM-based demand prediction with imbalance handling," Transp. Res. C, vol. 140, 2022.

[24] A. Radford, L. Metz, and S. Chintala, "Deep convolutional GANs," arXiv:1511.06434, 2015.

[25] F. Chollet, "Keras: Deep learning library," 2015. [Online]. Available: https://keras.io