International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

PREDICTING HOURLY BOARDING DEMAND OF BUS PASSENGERS USING IMBALANCED RECORDS

[1]SRUSHTI G B , [2]SHRUTHI M T
[1] Student, Department of MCA, BIET, Davangere
[2] Assistant Professor,Department of MCA, BIET, Davangere

ABSTRACT

The tap-on smart-card data provides a valuable source to learn passengers' boarding behaviour and predict future travel demand. However, when examining the smart-card records (or instances) by the time of day and by boarding stops, the positive instances (i.e. boarding at a specific bus stop at a specific time) are rare compared to negative instances (not boarding at that bus stop at that time). Imbalanced data has been demonstrated to significantly reduce the accuracy of machine-learning models deployed for predicting hourly boarding numbers from a particular location. This paper addresses this data imbalance issue in the smart-card data before applying it to predict bus boarding demand. We propose the deep generative adversarial nets (DeepGAN) to generate dummy travelling instances to add to a synthetic training dataset with more balanced travelling and nontravelling instances. The synthetic dataset is then used to train a deep neural network (DNN) for predicting the travelling and nontravelling instances from a particular stop in a given time window. The results show that addressing the data imbalance issue can significantly improve the predictive model's performance and better fit ridership's actual profile. Comparing the performance of the Deep-GAN with other traditional resampling methods shows that the proposed method can produce a synthetic training dataset with a higher similarity and diversity and, thus, a stronger prediction power. The paper highlights the significance and provides practical guidance in improving the data quality and model performance on travel behaviour prediction and individual travel behaviour analysis.

Keywords :-

Predictive models, Machine learning,Data models,Training, Generative adversarial networks, Ensemble learning, Biological system modeling

1. INTRODUCTION

THE rapid progress of urbanization leads to expansion of population in the urban area, increased demand for travel and associated adverse effects in traffic congestion and air pollution [1]–[3]. Public transport has been widely recognized as a green and sustainable mode of transportation to relieve such transport problems. As a conventional public transport mode, buses have always played a dominant role in passenger transportation [4], [5]. However, unreliable travel time, bus- bunching and crowding have led to low level- of services for buses [6]–[8]. This has decreased the bus ridership in many cities, particularly with the advent of ride-hailing services in recent years [9]–[11]. To sustain and increase bus patronage, bus operators must find a way to improve its

performance and enhance its image and attraction. Advanced operation and management for bus systems can significantly improve the level- of-service and service reliability, which in turn helps increase the bus ridership [12]– [14]. This requires understanding the spatial and temporal variations in passenger demand and making necessary changes on the supply side [15]–[18]. The smart-card system is initially designed for automatic fare collection. As the system also records the boarding information, for example, who gets on buses, where and when, smart-card data has become a ready-made and valuable data source for spatio- temporal demand analysis [19], public transport planning [20]–[23], and further analysis of emission reduction for the sustainable transport [24], [25]. From the smart-card data, we can easily observe the passenger flow at bus stops and on bus lines, and from which to derive the spatial and temporal characteristics of bus trips [26], [27]. However, extracting useful information from big data automatically still poses a significant challenge. In recent years, machine learning techniques have emerged as an efficient and effective approach to analyzing large smart-card datasets. For instance, Liu et al. [28] captured key features in public transport passenger flow prediction via a decision tree model. Zuo et al. [29] built a three- stage framework with a neural network model to forecast the individual accessibility in bus systems.

In our own recent research [30], we demonstrate that smartcard data combined with machine learning techniques can be a powerful approach for predicting the spatial and temporal patterns of bus boarding. The predictions were found to be highly accurate at an aggregated level, averaged over all travelers. However, our research has also thrown light on the data imbalance issues, when trying to predict travel behavior at the level of individual travelers and fine spatial- temporal details. For instance, the boarding of an individual smart-card holder at a specific stop during a particular time window (e.g. an hour) is a rare event: most of the records would denote negative (non- travelling, or not boarding at this bus stop during this time window) instances, and only a few are positive (travelling, boarding at this stop at this time) instances. Such data imbalance issues can significantly reduce the efficiency and accuracy of machine learning models deployed for predicting travel behavior at the level of individual travelers and fine spatial- temporal details. This motivates this current study where we propose an over-sampling method, deep generative adversarial nets (Deep-GAN) model (initially developed in the context of image generation) to address the data imbalance issue in predicting disaggregate boarding demand (i.e. individual passengers boarding behavior during each hour of the day). We show that, with the synthesized and more balanced database, the prediction accuracy improves significantly. The performance of the proposed approach, based on the Deep- GAN method, is further benchmarked against other resampling methods (including Synthetic Minority Oversampling Technique and Random Under- Sampling) and is shown to have superior performance.

The rest of the paper is organized as follows. Section II reviews the key resembling methods and their applications in transport studies. Section III describes the specific data imbalance issue in predicting the hourly boarding demand. Section IV uses a DeepGAN to provide a synthesized, more balanced training data sample and a deep neural network (DNN) to predict the individual smart-card holders' boarding actions (boarding or not boarding)

in any hour of a day. Section V applies the proposed method to a real-world case study, and the results are discussed in Section VI. Finally, Section VII summarizes the main findings and contributions of this paper and suggests future investigations.

2. LITERATURE SURVEY:

Predicting Hourly Boarding Demand of Bus Passengers Using Imbalanced Records From Smart-Cards: A Deep Learning Approach (Tang et al., 2023) This paper proposes a Deep Generative Adversarial Network (Deep-GAN) approach to address imbalanced data in smart card records. DeepGAN generates synthetic boarding instances to create a more balanced training dataset for a Deep Neural Network (DNN) predicting boarding demand at specific stops.

Multi-stage deep learning approaches to predict boarding behaviour of bus passengers (Tang et al., 2023) This study explores a multi-stage deep learning framework for predicting passenger boarding behavior. It acknowledges the data imbalance issue and suggests potential solutions like oversampling or cost-sensitive learning for future research.

HOURLY BUS PASSENGER DEMAND PREDICTION THROUGH MACHINE LEARNING ALGORITHMS
(International Journal of Information Technology and Computer Engineering, 2020)
This paper highlights the limitations of traditional machine learning algorithms in handling imbalanced datasets for bus passenger demand prediction. It emphasizes the need for techniques like SMOTE
(Synthetic Minority Over-sampling Technique) to improve model performance.

A Bus Passenger Flow Prediction Model Fused with Point-of-Interest Data Based on
Extreme Gradient Boosting (Lv et al., 2022)
While not directly addressing data imbalance, this paper introduces a passenger flow prediction model using Extreme Gradient Boosting (XGBoost). This technique might be adaptable to imbalanced data with proper parameter tuning for handling the class imbalance.

Multi-Step Subway Passenger Flow Prediction under Large Events Using
Website Data (Wang et al., 2021) This research focuses on subway passenger flow prediction but offers valuable insights for bus ridership as well. It explores a multi- step prediction approach using website data, demonstrating the importance of incorporating diverse features beyond just historical ridership data. This can potentially improve model generalizability even with imbalanced datasets.

3. MODULE DESCRIPTION :

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 Browse and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Hourly Boarding Demand Type, View Hourly Boarding Demand Type Ratio,
Download Trained Data Sets, View Hourly
Boarding Demand Type Ratio Results, View All Remote Users,

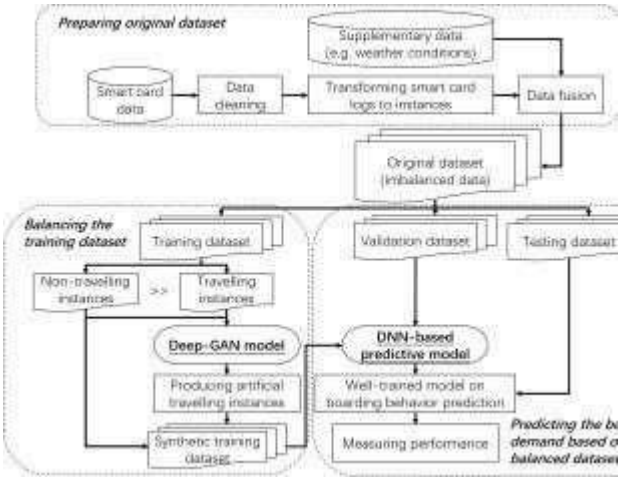
View and Authorize Users

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, Predicting Hourly Boarding Demand Type, VIEW YOUR PROFILE.

Flow chart :-



4. METHEDOLOGIES:

Data Collection and Preparation:

Gather historical data on hourly boarding demand of bus passengers.

This data might include timestamps, location, weather conditions, special events, etc.

Handle missing data and outliers appropriately.

Feature Engineering:

Extract relevant features such as time of day, day of the week, month, weather conditions, holidays, etc.

Transform categorical variables into numerical representations (one-hot encoding, label encoding, etc.).

Handling Imbalanced Data:

Imbalanced data refers to a situation where the number of examples in each class (e.g., high demand vs. low demand hours) is significantly different.

Techniques to handle imbalance include:

Resampling: Either oversampling the minority class (e.g., SMOTE - Synthetic Minority Over-sampling Technique) or undersampling the majority class.

Class Weights: Adjusting the class weights in the machine learning algorithm to penalize mistakes on the minority class more.

4. Model Selection:

Choose appropriate models that can handle imbalanced data well, such as:

Gradient Boosting Machines (GBM): XGBoost, LightGBM, etc.

Random Forests: Can handle imbalanced data to some extent.

Support Vector Machines (SVM) with balanced class weights.

Neural Networks: Especially architectures designed for handling imbalanced data (e.g., adjusting loss functions, class weights).

Training and Validation:

Split the data into training and validation sets (and possibly a test set for final evaluation).

Use cross-validation techniques (Stratified Kfold) to ensure robustness of the model.

Evaluation Metrics:

Choose appropriate metrics to evaluate the model's performance, considering the imbalance:

Precision, Recall, F1-score: Especially useful when dealing with imbalanced classes.

ROC-AUC (Receiver Operating Characteristic - Area Under Curve): Useful for binary classification tasks.

Algorithms Used

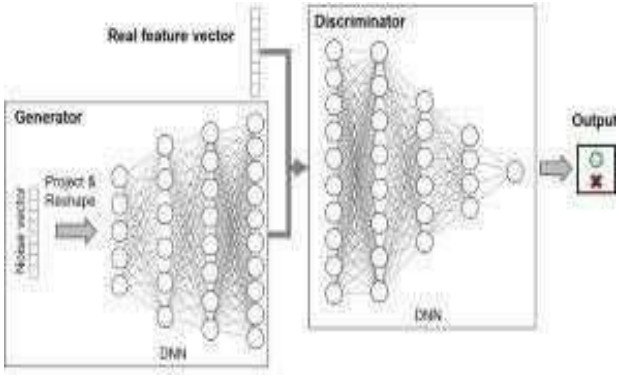
Random Forests: Ensemble learning method that can handle imbalanced data by adjusting decision thresholds.

Gradient Boosting Machines (GBM): Builds trees sequentially, focusing on instances that were misclassified in previous iterations, which can help in dealing with imbalanced data.

XGBoost (Extreme Gradient Boosting): A popular GBM implementation known for its efficiency and effectiveness in handling imbalanced data.

Support Vector Machines (SVM): Can use kernels to handle non-linear decision boundaries and class weights to adjust for imbalance.
Neural Networks: Deep learning models can be effective when tuned with appropriate loss functions (like weighted cross-entropy) and regularization techniques.

5. RESULT:



6. CONCLUSION:

The motivation of this study was because we have faced the challenge of imbalanced data when we used the real world bus smart-card data to prediction the boarding behavior of passengers at a time window. In this research, we proposed a Deep-GAN to over-sample the travelling instances and to re-balance the rate of travelling and non- travelling instances in the smart-card dataset in order to improve a DNN based prediction model of individual boarding behavior. The performance of DeepGAN was evaluated by applying the models on real-world smart-card data collected from seven bus lines in the city of Changsha, China. Comparing the different imbalance ratios in the training dataset, we found out that in general, the performance of the model improves with more imbalanced data and the most significant improvement comes at a 1:5 ratio between positive and negative instances. From the perspective of prediction accuracy of the hourly distribution of bus ridership, the high rate of imbalance will cause misleading load profiles and the absolutely balanced data may over predict the ridership during peak hours. Comparison of different resembling methods reveals that both over-sampling and under-sampling benefits the performance of the model. Deep- GAN has the best recall score and its precision scores best among the over- sampling methods. Although the performance of the predictive model trained by the Deep-GAN-data is not significantly beyond other resembling

methods, the Deep- GAN also presented a powerful ability to improve the quality of training dataset and the performance of predictive models, especially when the under-sampling is not suitable for the data.

The contributions of this study are:

- The data imbalance issue in the public transport system has received little attention, and this study is the first to focus on this issue and propose a deep learning approach, Deep- GAN, to solve it.

- This study compared the differences in similarity and diversity between the real and synthetic travelling instanced generated from Deep-GAN and other over-sampling methods. It also compared different resembling methods for the improvement of data quality by evaluating the performance of the next travel behavior prediction model. This is the first validation and evaluation of the performance of different data resembling methods based on real data in the public transport system.

- This paper innovatively modeled individual boarding behavior, which is uncommon in other travel demand prediction tasks. Compared to the popular aggregated prediction, this individual-based model is able to provide more details on the passengers' behavior, and the results will benefit the analysis of the similarities and heterogeneities.

As technology and computing power develop, predicting models will become more and more refined. In the field of demand prediction of the public transport systems, the target will gradually evolve from the bus network and bus lines to individual travel behavior. This advancement can greatly benefit public transport planning and management, such as the digital twin of the public transport system. It is foreseeable that future prediction work in public transport systems will also encounter the challenge of imbalanced data. Our research proposes a Deep-GAN model to address the data imbalance issue in travel behavior prediction. The validation via real world data illustrated that the Deep-GAN showed a better ability to deal with the data imbalance issue and benefits the predictive models compared to other resembling methods. This research provides valuable experience for more researchers and managers in dealing with similar data imbalance issues, especially in public transport.

7. REFERENCES:

[1] X. Guo, J. Wu, H. Sun, R. Liu, and Z. Gao, "Timetable coordination of first trains in urban railway network: A case study of beijing," Applied Mathematical Modelling, vol. 40, no. 17, pp. 8048–8066, 2016.

[2] W. Wu, P. Li, R. Liu, W. Jin, B. Yao, Y. Xie, and C. Ma, "Predicting peak load of bus routes with supply optimization and scaled shepard interpolation: A newsvendor model," Transportation Research Part E: Logistics and Transportation Review, vol. 142, p. 102041, 2020.

[3] N. Beˇsinovi´c, L. De Donato, F. Flammini, R. M. Goverde, Z. Lin, R. Liu, S. Marrone, R. Nardone, T. Tang, and V. Vittorini, "Artificial intelligence in railway transport: Taxonomy, regulations and applications," IEEE Transactions on Intelligent Transportation Systems, 2021.

[4] S. C. Kwan and J. H. Hashim, "A review on co-benefits of mass public transportation in climate change mitigation," Sustainable Cities and

Society, vol. 22, pp. 11–18, 2016.

[5] Y. Wang, W. Zhang, T. Tang, D. Wang, and Z. Liu, "Bus od matrix reconstruction based on clustering wi-fi probe data,"

Transportmetrica B: Transport Dynamics, pp. 1–16, 2021, doi:

10.1080/21680566.2021.1956388.

[6] S. J. Berrebi, K. E. Watkins, and J. A. Laval, "A real-time bus dispatching policy to minimize passenger wait on a high frequency route," Transportation Research Part B: Methodological, vol. 81, pp.

377–389, 2015.

[7] A. Fonzone, J.-D. Schm¨ocker, and R. Liu, "A model of bus bunching under reliability-based passenger arrival patterns," Transportation Research

Part C: Emerging Technologies, vol. 59, pp. 164–182, 2015.

[8] J. D. Schm¨ocker, W. Sun, A. Fonzone, and R. Liu, "Bus bunching along a corridor served by two lines," Transportation Research Part

B: Methodological, vol. 93, pp. 300–317, 2016.

[9] D. Chen, Q. Shao, Z. Liu, W. Yu, and C. L. P. Chen, "Ridesourcing behavior analysis and prediction: A network perspective," IEEE Transactions on Intelligent Transportation Systems, pp. 1– 10, 2020.

[10] E. Nelson and N. Sadowsky, "Estimating the impact of ride-hailing app company entry on public transportation use in major us urban areas,"

The B.E. Journal of Economic Analysis & Policy, vol. 19, no. 1, p.

20180151, 2019.

[11] Z. Chen, K. Liu, J. Wang, and T. Yamamoto, "H-convlstm-based bagging

learning approach for ride-hailing demand prediction considering imbalance problems and sparse uncertainty," Transportation Research

Part C: Emerging Technologies, vol. 140, p. 103709, 2022.

[12] R. Liu and S. Sinha, "Modelling urban bus service and passenger reliability," 2007.

[13] J. A. Sorratini, R. Liu, and S. Sinha, "Assessing bus transport teliability using micro-simulation," Transportation Planning and Technology, vol. 31, no. 3, pp. 303–324, 2008.

[14] Y. Wang, W. Zhang, T. Tang, D. Wang, and Z. Liu, "Bus od matrix reconstruction based on clustering wi-fi probe data," Transportmetrica B: Transport Dynamics, pp. 1–16, 2021.

[15] Y. Hollander and R. Liu, "Estimation of the distribution of travel times

by repeated simulation," Transportation Research Part C: Emerging

Technologies, vol. 16, no. 2, pp. 212–231, 2008.

[16] W. Wu, R. Liu, and W. Jin, "Modelling bus bunching and holding control

with vehicle overtaking and distributed passenger boarding behaviour,"

Transportation Research Part B: Methodological, vol. 104, pp. 175–197, 2017.

[17] W. Wu, R. Liu, W. Jin, and C. Ma,
"Stochastic bus schedulecoordination

considering demandassignment

and

rerouting of passengers,"

Transportation Research Part B: Methodological, vol. 121, pp. 275–303, 2019.

[18] W. Wu, R. Liu, and W. Jin, "Designing robust schedule coordination

scheme for transit networks with safety control margins," Transportation

Research Part B: Methodological, vol. 93, pp. 495–519, 2016.

[19] S. Zhong and D. J. Sun, A Spatio- temporal Distribution Model for Determining

Origin–Destination Demand from Multisource Data. Springer,

Singapore, 2022, pp. 33–52.

[20] M. Bordagaray, L. dell'Olio, A.
Fonzone, and Ibeas, "Capturing the conditions that introduce systematic variation in bike-sharing travel behavior using data mining techniques,"
Transportation Research Part

C: Emerging Technologies, vol. 71, pp. 231– 248, 2016.

[21] B. Chidlovskii, "Mining smart card data for travellers' mini activities,"