JAMBURA JOURNAL OF INFORMATICS
Volume 5, No. 1 | April 2023, pp. 1-6
ISSN: 2685-4244, DOI: 10.37905/jji.v5i1.18814

Enhancing the efficiency of Jakarta's mass rapid transit system with XGBoost algorithm for passenger prediction

Muhammad Alfathan Harriz[1], Nurhaliza Vania Akbariani[2], Harlis Setiyowati[3], Handri Santoso[4]

[1,3,4] Pradita University, Tangerang, Indonesia
[2]Sekolah Tinggi Teknologi Terpadu Nurul Fikri, Jakarta, Indonesia

Article history:
Received Feb 2, 2023
Revised March 20, 2023
Published April 27, 2023

Keywords:
Jakarta
Mass rapid transit
XGBoost

ABSTRACT. This study is based on a machine learning algorithm known as XGBoost. We used the XGBoost algorithm to forecast the capacity of Jakarta's mass transit system. We used preprocessed raw data from the Jakarta Open Data website for 2020-2021 as a training medium to achieve a mean absolute percentage error of 69. However, after the model was fine-tuned, the MAPE was significantly reduced by 28.99% to 49.97. The XGBoost algorithm effectively detected patterns and trends in the data, which can be used to improve routes and plan future studies by providing valuable insights. It is possible that additional data points, such as holidays and weather conditions, will further enhance the model's accuracy in future research. As a result of implementing XGBoost, Jakarta's transportation system can optimize resource utilization and improve customer service to improve passenger satisfaction. Future studies may benefit from additional data points, such as holidays and weather conditions, to improve XGBoost's efficiency.

Corresponding Author:
Muhammad Alfathan Harriz,
Pradita University,
Tangerang, Indonesia.
Email: harrizsb@gmail.com

INTRODUCTION

Approximately 662.33 square kilometres of land covers Indonesia's capital city, Jakarta, which has an estimated population of 10.2 million (Azhar et al., 2020). Due to Jakarta's underuse of public transportation, many citizens drive instead of the public transportation system (Rachman et al., 2021). A few factors contribute to this, including the lack of public transportation, long travel times, insufficient security, discomfort (Sinaga et al., 2019), and long travel times (Utami et al., 2022). Public transportation has several advantages, including reducing the use of fossil fuels and helping to reduce the impact on the environment (Abdulrazzaq et al., 2020; Majid et al., 2020; Solihati & Indriyani, 2021). The government of DKI Jakarta issued instructions to improve air quality and reduce congestion. Encouraging citizens to use public transportation instead of private vehicles was part of the instruction (Rachman et al., 2021). To address traffic problems in Jakarta, the city's local authorities launched a mass rapid transit (MRT) service to transport travellers between the south and city center in response to the rising demand for rapid transportation (Febriani et al., 2020).

In this study, we investigated an efficient, reliable, and cost-effective machine learning algorithm– XGBoost– to predict the number of passengers in Jakarta's rapid transit system. This system is an essential part of Jakarta's transportation infrastructure, providing commuters with connections to the city center. This study utilized the total number of MRT users in the DKI Jakarta Province from 2020 to 2021, and this data is taken from the website of Jakarta open data (*Data Penumpang MRT 2021 Di*

Provinsi DKI Jakarta - Open Data Jakarta, n.d.; *Data Penumpang MRT Di Provinsi DKI Jakarta - Open Data Jakarta*, n.d.). XGBoost can pinpoint patterns and trends in the data, offering valuable insight into Jakarta's rapid transit system. XGBoost has been demonstrated to accurately predict passenger data (Ramana, 2022; Shen, 2022; Tiong et al., 2023; Zou et al., 2022), and it outperformed other models, such as KNN and LightGBM (Zhu et al., 2022). We ensure our model's precision by utilising Mean Absolute Percentage Error. The result of XGBoost regression can enhance the commuter experience, such as route optimization, capacity planning, and customer service. The main objective of this study was to evaluate the accuracy of XGBoost in improving decision-making related to passenger capacity expansion, operational efficiency, and cost management in Jakarta's transportation system. To answer this question more specifically, this study attempts to identify how XGBoost will improve decision-making related to these key areas, ultimately enhancing passenger satisfaction through improved decision-making.

METHODS

XGBoost

Chen and Guestrin developed the XGBoost algorithm in 2016 (Paleczek et al., 2021; Pan et al., 2022), and in comparison to popular machine learning (ML) and deep learning (DL) methods, it is ten times faster. XGBoost is an ensemble approach that uses a gradient descent algorithm to progressively add new models until the performance of the existing ones cannot be further enhanced (Paleczek et al., 2021). The main aim of XGBoost is to create a single strong classifier from several weak classifiers that can enhance forecasting accuracy (Deng et al., 2022; Zhang et al., 2019). In addition, XGBoost is adept at managing sparsity, missing values, zeroes, and feature engineering results in limited memory and distributed environments. In addition, it is optimized, scalable, and capable of processing billions of examples (Deng et al., 2020). The XGBoost formula is given by

$$\hat{y} = \sum_{m=1}^{M} f_m(x) \quad (1)$$

It is estimated that f_m represents one weak learner, and M represents the number of weak learners. Using the weights of the weak learners a final prediction is generated based on the weights of the weak learners. Gradient boosting is an iterative process for minimizing the loss function by optimizing the model parameters over time. The difference between the predicted and actual values was determined using this function.

Preprocessing

Fan et al. (2021) clearly showed that pre-processing data can enhance data analysis reliability by transforming raw data into a manageable form. Prakash and Aloysius (2019) stated that this technique is critical because data are often of poor quality, complex, and have inherent limitations. It is ideal for cleaning, reducing, scaling, transforming, and partitioning data before it can be analyzed. During preprocessing, noise must be removed from the algorithm, and the area of interest must be defined such that the algorithm focuses only on that area (Ranganathan, 2021).

RESULTS AND DISCUSSION

Using Jakarta open data from 2020 and 2021, the data were compiled into a single CSV file and preprocessed for training and testing, whereas only the date and total were filtered. There were a total of 24 columns. Table 1 illustrates some of the final results.

Table 1. Compiled data

Date	Total
2020-01	2638464
2020-02	2564870
2020-03	1403638
2020-04	121578
2020-05	43544

Our next step is to visualize the data using a line chart, as shown in Figure 1, to represent it visually.

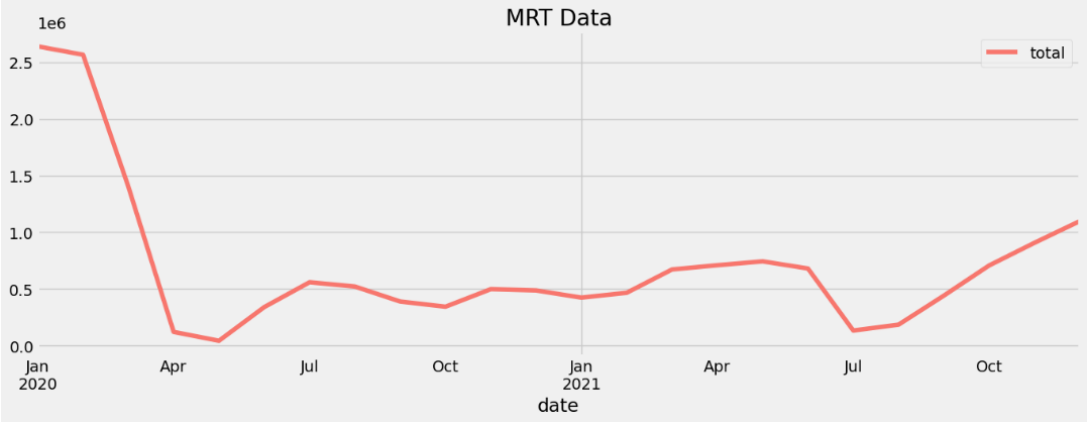

Figure 1. Display the data in a line chart

Considering the limited available data, we divided the data into 50 per cent training and 50 per cent testing. Figure 2 illustrates a line chart separating the training and testing data.

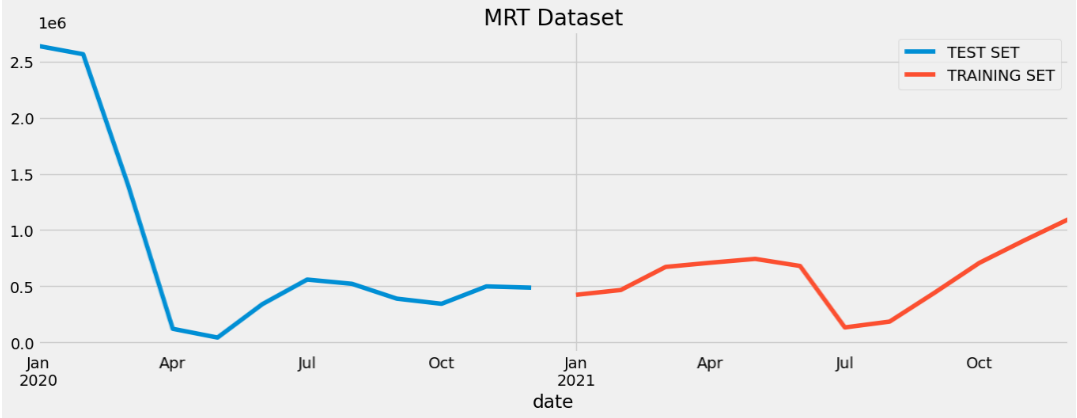

Figure 2. The red line symbolizes training, while testing is symbolized by the blue line

The XGboost library was used to predict the MRT data using a prediction algorithm with the n_estimators parameter. For the MRT dataset, we used n_estimators of 12 because we had only 24 data points and an early_stopping_round of 5. MAPE is the Mean Absolute Percentage Error used to evaluate the model. The formula used is as follows:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \quad (2)$$

where A_t is the actual value at time t and F_t is the forecasted value at time t.

It is ideal to use the Mean Absolute Percentage Error (MAPE) to assess the accuracy of forecasting techniques because it considers both the magnitude and direction of the errors. In addition, MAPE offers a clear interpretation of relative error, making it an excellent choice for tasks where sensitivity to relative variance is more important than absolute variation (Chicco et al., 2021). A MAPE of approximately 69 was used in this run. We tweaked the parameters to set n_estimators to 57, seed to 129, eta of 0.03, subsamples to 0.1, and colsample_bytes to 1. As a result of the tuning, the MAPE number is now 49.97, which means that the error rate has decreased by approximately 28.99% since the first run. Despite the MAPE of 49.97%, the prediction model did not appear very accurate. Therefore, errors and suboptimal decisions can occur based on the predictions. Therefore, when predictions inform important decisions, minimising the MAPE as much as possible is generally desirable. Despite this significant reduction in the MAPE by 28.99%, further improvements in accuracy and precision must be refined. This is shown in Fig. 3.

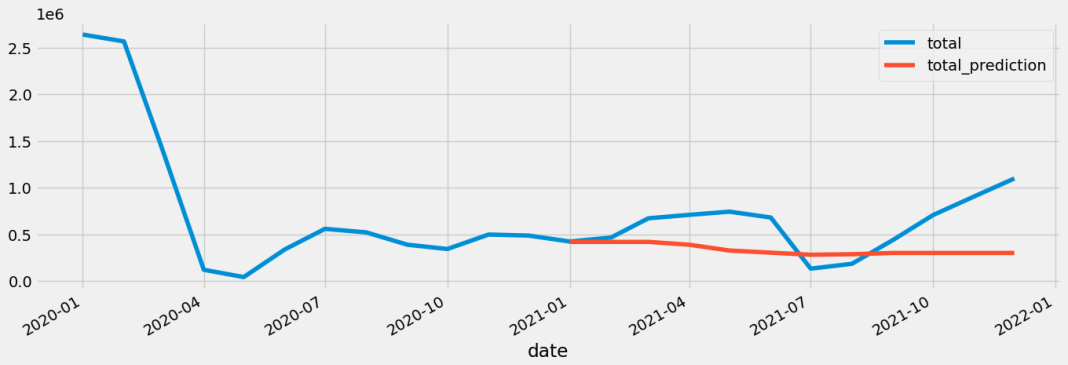

Figure 3. An illustration of the predicted value and the real value in a line chart

CONCLUSION

This study aimed to forecast passenger capacity using XGBoost, a machine-learning algorithm. This study demonstrated XGBoost's ability to uncover patterns and trends by preprocessing raw data from Mass Rapid Transit passengers in DKI Jakarta Province for 2020-2021 into structured forms, resulting in a mean absolute percentage error of 49.97. Even with limited training data, effective parameter tuning has improved predictive accuracy, despite initial concerns regarding the high MAPE rate of XGBoost. Increasing the training data is possible to increase the model's accuracy. Future studies may also use additional data points, such as holidays and weather conditions, to further enhance the results. By implementing XGBoost, Jakarta's transportation system can improve passenger satisfaction by improving planning capacity, resource utilization, and customer service.

REFERENCES

Abdulrazzaq, L. R., Abdulkareem, M. N., Mat Yazid, M. R., Borhan, M. N., & Mahdi, M. S. (2020). Traffic congestion: the shift from private cars to public transportation. *Civil Engineering Journal*, *6*(8), 1547–1554. doi: 10.28991/cej-2020-03091566

Azhar, H. N., Fatima, H. H. P., & Tamas, I. N. (2020). Preliminary study of Indonesia capital city relocation based on disaster mitigation principle with mental model approach. *E3S Web of Conferences*, *148*, 06002. doi: 10.1051/e3sconf/202014806002

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. doi: 10.7717/peerj-cs.623

Data Penumpang MRT 2021 di Provinsi DKI Jakarta—Open Data Jakarta. (n.d.). Retrieved March 20, 2023, from https://data.jakarta.go.id/dataset/data-penumpang-mrt-2021-di-provinsi-dki-jakarta

Data Penumpang MRT di Provinsi DKI Jakarta—Open Data Jakarta. (n.d.). Retrieved March 20, 2023, from https://data.jakarta.go.id/dataset/data-penumpang-mrt-di-provinsi-dki-jakarta

Deng, A., Zhang, H., Wang, W., Zhang, J., Fan, D., Chen, P., & Wang, B. (2020). Developing computational model to predict protein-protein interaction sites based on the XGBoost algorithm. *International Journal of Molecular Sciences*, *21*(7), doi: 10.3390/ijms21072274

Deng, X., Ye, A., Zhong, J., Xu, D., Yang, W., Song, Z., Zhang, Z., Guo, J., Wang, T., Tian, Y., Pan, H., Zhang, Z., Wang, H., Wu, C., Shao, J., & Chen, X. (2022). Bagging–XGBoost algorithm-based extreme weather identification and short-term load forecasting model. *Energy Reports*, *8*, 8661–8674. doi: 10.1016/j.egyr.2022.06.072

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in Energy Research*, *9*, 652801. doi: 10.3389/fenrg.2021.652801

Febriani, D., Mega Olivia, C., Aniisah Sholilah, S., & Hidajat, M. (2020). Analysis of modal shift to support MRT-based urban transportation in Jakarta. *Journal of Physics: Conference Series*, *1573*(1), 012015. doi: 10.1088/1742-6596/1573/1/012015

Majid, R. A., Said, R., Mohamad, N., Abdullah, J., & Ngah, R. (2020). The impact of time attribute on mass rapid transport (MRT) ridership in Malaysia. *International Journal of Social Science Research*, *2*(4).

Paleczek, A., Grochala, D., & Rydosz, A. (2021). Artificial breath classification using XGBoost algorithm for diabetes detection. *Sensors*, *21*(12), 4187. doi: 10.3390/s21124187

Pan, S., Zheng, Z., Guo, Z., & Luo, H. (2022). An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. *Journal of Petroleum Science and Engineering*, *208*, 109520. doi: 10.1016/j.petrol.2021.109520

Prakash, T. N., & Aloysius, A. (2019). Data preprocessing in sentiment analysis using twitter data. *International Educational Applied Research Journal (IEAJR). 3 (7),* 89–92.

Rachman, F. F., Nooraeni, R., & Yuliana, L. (2021). Public opinion of transportation integrated (Jak Lingko), in DKI Jakarta, Indonesia. *Procedia Computer Science*, *179*, 696–703. doi: 10.1016/j.procs.2021.01.057

Ramana, A. V. (2022). Taxi demand prediction using ML. *International Journal for Research in Applied Science and Engineering Technology*, *10*(6), 3811–3815. doi: 10.22214/ijraset.2022.43912

Ranganathan, G. (2021). A study to find facts behind preprocessing on deep learning algorithms. *Journal of Innovative Image Processing*, *3*(1), 66–74. doi: 10.36548/jiip.2021.1.006

Shen, E. Z. (2022). *Short-time cab speed prediction model based on XGBoost* [Preprint]. In Review. doi: 10.21203/rs.3.rs-2200855/v1

Sinaga, S. M., Hamdi, M., Wasistiono, S., & Lukman, S. (2019). Model of implementing bus rapid transit (BRT) mass public transport policy in DKI Jakarta province, Indonesia. *International Journal of Science and Society*, *1*(3). doi: 10.54783/ijsoc.v1i3.51

Solihati, K. D., & Indriyani, D. (2021). Managing artificial intelligence on public transportation (case study Jakarta city, Indonesia). *IOP Conference Series: Earth and Environmental Science*, *717*(1), 012021. doi: 10.1088/1755-1315/717/1/012021

Tiong, K. Y., Ma, Z., & Palmqvist, C.-W. (2023). A review of data-driven approaches to predict train delays. *Transportation Research Part C: Emerging Technologies*, *148*, 104027. doi: 10.1016/j.trc.2023.104027

Utami, D. L., Putri, A. L., Sutomo, A. H., & Ismara, K. I. (2022). Design of ergonomic emergency car toilet seats as a solution to severe traffic in Jakarta, Indonesia. *13th International Conference on Applied Human Factors and Ergonomics* (AHFE 2022). doi: 10.54941/ahfe1002004

Zhang, R., Li, B., & Jiao, B. (2019). Application of XGboost algorithm in bearing fault diagnosis. *IOP Conference Series: Materials Science and Engineering*, *490*, 072062. doi: 10.1088/1757-899X/490/7/072062

Zhu, L., Shu, S., & Zou, L. (2022). XGBoost-based travel time prediction between bus stations and analysis of influencing factors. *Wireless Communications and Mobile Computing*, *2022*, 1–25. doi: 10.1155/2022/3504704

Zou, L., Shu, S., Lin, X., Lin, K., Zhu, J., & Li, L. (2022). Passenger flow prediction using smart card data from connected bus system based on interpretable XGBoost. *Wireless Communications and Mobile Computing*, *2022*, 1–13. doi: 10.1155/2022/5872225