Exploring Shared Bike Station Utilization Patterns and Availability through Machine Learning

機械学習によるシェアサイクルステーションでの利用パターンと利用可能性の分析

Department of Urban Engineering, UTokyo　03-230139 Shunta Kochi

In recent years, shared bikes have gained attention as a new mode of transportation and have been widely adopted. However, in the operation of bike-sharing systems, the nature of the system, which allows users to freely choose stations for rental and return, creates a challenge of supply-demand imbalances in the number of available bikes depending on the time of day and station. This study focuses on analyzing utilization patterns and availability at shared bike stations using machine learning. By applying K-Means clustering to the variation in availability ratio at shared bike stations in Chiba City, it was revealed that there are several distinct patterns of fluctuation in bike availability. Findings indicate that shared bikes complement trips from suburban areas to railway stations, from stations to destinations, and between city centers. Furthermore, a classification model was developed using LightGBM to predict whether stations are in a state of low or high bike availability. The model used location, time, and weather information as input features. By constructing model, the model successfully captured the availability patterns of stations and demonstrated high predictive accuracy.

1. Introduction

Shared bikes are gaining popularity as a sustainable urban transportation mode, offering benefits such as reducing car use, easing congestion, and promoting health. However, an imbalance in supply and demand among stations poses challenges, lowering user satisfaction and increasing costs. To improve system efficiency, predicting bike shortages and surpluses is crucial for effective relocation management. This study aims to analyze the temporal and spatial imbalance of bike availability, develop a predictive model, and identify key factors influencing availability through model interpretation.

2. Literature Review

Early approaches used statistical models like ARIMA and Linear Regression, which offer interpretability but struggle with non-linear relationships. [1, 2] Machine learning models, including, Random Forests, and Gradient Boosted Decision Trees, improve prediction accuracy but lack interpretability. [1, 2, 3, 4] Deep learning models like LSTM and GRU effectively capture temporal dependencies but are criticized for being "black boxes". [2, 5, 6, 7]

While past research has prioritized predictive accuracy, fewer studies have explored model interpretation and usage patterns. This study addresses these gaps by emphasizing model interpretability and analyzing bike usage patterns.

3. Data Source and Analysis Process

3.1 Overview of Data Source

This study utilizes three types of data:

1. Bike-Sharing System (BSS) data

BSS data is based on the GBFS (General Bikeshare Feed Specification) data provided by ODPT. This data was downloaded at 5-minute intervals, and a time-series database was constructed containing the number of bikes available, parking spaces available, and maximum capacity at each station for each time interval. Additionally, OD data from OpenStreet Inc, which records bike rentals and returns were used for the analysis.

2. Meteorological Data

Meteorological Data provided by the Japan Meteorological Agency from December 20, 2023, to May 8, 2024, was used. This data includes hourly records of weather variables as temperature, precipitation, humidity, etc.

3. Location Feature Data

Nine datasets were used to provide geographical and infrastructure context:

- Corporate Search Data (2023)
- Railway Data (Station: 2020, Passenger volume: 2021)
- Bus Stop Data (2010)
- Population and Household data (2020)
- Road density and length data (2010)
- Elevation and Slope data (2009)
- Population mobility, employment status, and commuting and schooling destinations (2020)
- Sectional Traffic Volume data (from December 20, 2023, to May 8, 2024)

3.2 Analysis Process

First, exploratory data analysis and preprocessing are performed to examine basic statistics and preprocess missing values, and outliers in the data. Next, the variation patterns of the station availability ratio (defined as available bikes/maximum capacity and hereinafter referred to as avratio) in BSS stations in Chiba City are extracted using the K-Means clustering method to understand usage patterns.

Following this, feature engineering is performed to construct input dataset for an availability forecasting model for stations across Chiba Prefecture.

In this study, an availability forecasting model was formulated as a classification model that predicts the class label for the avratio a (Low: $0 \leq a < 0.1$, Medium: $0.1 \leq a \leq 0.9$, High: $0.9 < a$) for each station on an hourly basis.

Dataset was divided into training and test data. Clustering and construction of availability forecasting model was performed using training dataset. Furthermore, hyperparameter tuning is conducted and the final model was evaluated by comparing predictions with the test data. Subsequently, model is interpreted and considered using Shapley Additive exPlanations (SHAP).

4. Clustering

A two-level clustering, first-level with the avratio at 3:00 AM, second-level with the avratio across the entire study period is conducted and the BSS stations were split into four clusters.

Figure 1 shows the average avratio changes for each cluster, along with the hourly average number of rentals and returns for stations within each cluster. Also, Figure 2 is the geographical distribution of stations in each cluster.

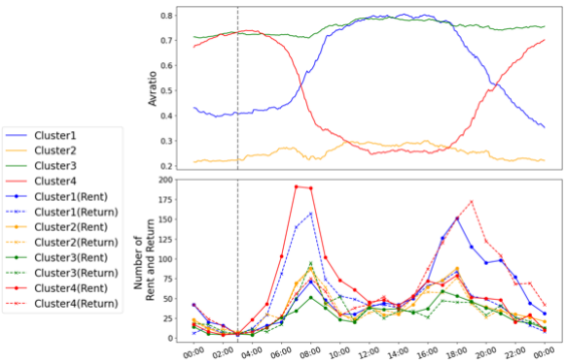

Figure 1: Average avratio variation and rentals and returns for stations within each cluster

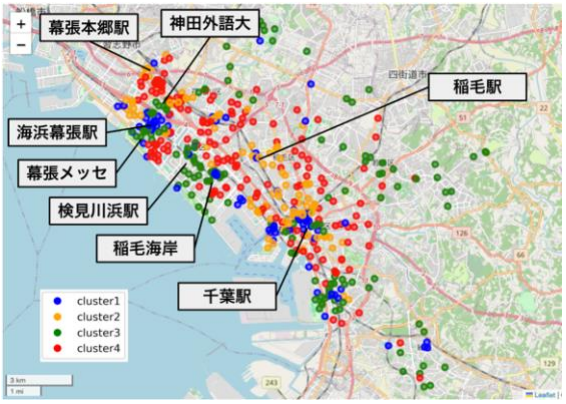

Figure 2: Geographical distribution of stations in each cluster

Cluster 1 stations showed low avratio during the early morning hours, which increased during the day. In contrast, the avratio variation pattern for cluster 4 stations was opposite to that of cluster 1. The rental

and return volumes also reflected these trends.

Cluster 1 and 4 showed complementary spatial distributions: cluster 1 stations were concentrated near railway stations, while cluster 4 stations were located slightly farther from railway stations. This spatial relationship suggests that shared bikes complement public transportation by facilitating trips from suburban areas to stations and then to destinations via rail.

5. Feature Engineering

Feature engineering was conducted to create input features for the availability forecasting model.

For location feature, 21 were created: "Facility Agglomeration (commercial, office-related, medical, educational)", "Distance to the Nearest Public Transportation", "Population Density", "Road Density", "Maximum Slope", "Average Slope", "Hourly Traffic Volume", "Distance to the Nearest Shared Bike Station", … etc.

For time feature, 14 were created: "Morning Flag", "Daytime Flag", "Evening Flag", "Late-night/Early Morning Flag", "Weekend Flag", "Month", "Day", "Day of Week", "Day(sin)", "Day(cos)", "Week(sin)", "Week(cos)", "Year(sin)", "Year(cos)".

For weather feature, 13 were created: "Ground Pressure", "Rainfall", "Temperature", "Dew Point Temperature", "Vapor Pressure", "Humidity", "Wind Speed", "Wind Direction", "Sunshine Hours", "Snowfall", "Snow Depth", "Weather", "Visibility".

6. Results of Modeling

In this study, we utilize LightGBM, a library for Gradient Boosting Decision Tree (GBDT) [7]. Also, F1-score and its average for all classes, macro-F1-score was employed as the primary evaluation metric. Additionally, a confusion matrix was employed for visualizing the results.

Table 1: F1-scores and macro-F1-score

Data	Class	F1-score
Train	Low	0.94
	Medium	0.88
	High	0.93
	Macro	0.92
Test	Low	0.67
	Medium	0.75
	High	0.73
	Macro	0.72

Table 1 and Figure 3 summarize the prediction results on the test dataset using constructed model. A model with consistent accuracy was constructed for both train and test datasets. However, the difference in scores between the train and test data is significant, indicating that the model is overfitting.

7. Discussion

Next, the predictions for each class were analyzed using SHAP value. SHAP provides insights into the contribution of individual features to specific predictions.

For both the Low and High classes, the top five features in terms of SHAP values were the

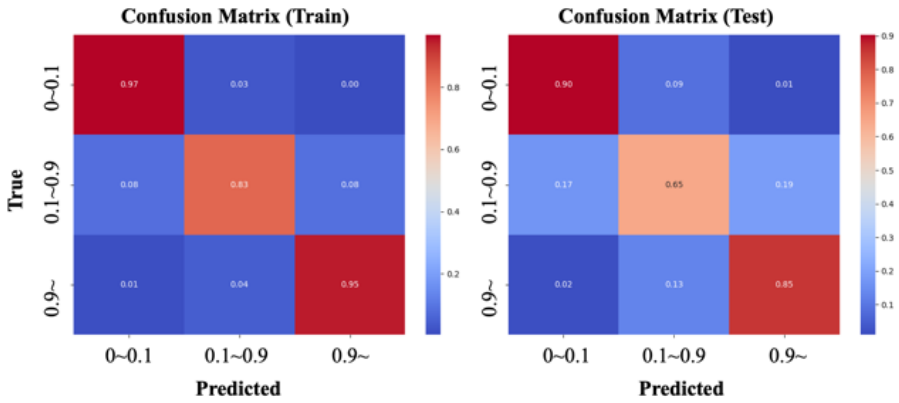

Figure 3: Confusion Matrix

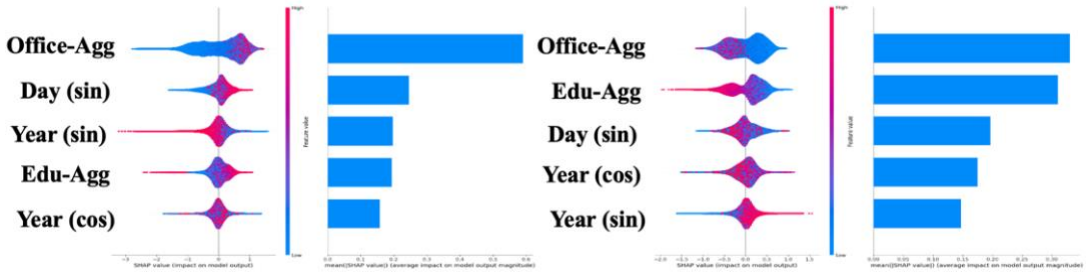

Figure 4: SHAP values

same, agglomeration of office and educational facilities, and time features. These results indicate that the availability of shared bikes at stations is significantly influenced by the physical environment, such as surrounding facilities, as well as temporal conditions that correspond to the location's characteristics.

8. Conclusion

Based on clustering analysis and model construction, it can be inferred that the imbalance between supply and demand occurs primarily at stations where bikes are used for commuting to work and school during peak hours, as well as at stations where the availability of bikes is consistently high or low.

The limitations and future directions of this study include addressing the issue of overfitting in the availability forecasting model, increasing the variety of facility categories, and incorporating residential density as a feature. Additionally, it is necessary to analyze the relationship between bike availability and actual supply-demand dynamics. By analyzing the extent of latent demand alongside availability, it would be possible to identify the true conditions under which supply-demand imbalances occur.

References

1) Y. Feng and S. Wang. A forecast for bicycle rental demand based on random forests and multiple linear regression. 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), pages101–105, 2017.

2) M. M. Isalm, M. E. Biswas, M. Shahzamal, M. D. Haque, and M. S. Hossain. An effective data driven approach to predict bike rental demand. In 2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI), pages 1–5, 2023.

3) Sathishkumar V E, J. Park, and Y. Cho. Using data mining techniques for bike sharing demand prediction in metropolitan city. Computer Communications, 153:353–366, 2020.

4) A. A. Ramesh, S. P. Nagisetti, N. Sridhar, K. Avery, and D. Bein. Station-level demand prediction for bike-sharing system. In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pages 0916–0921, 2021.

5) S. Choi and M. Han. The empirical evaluation of models predicting bike sharing demand. In 2020 International Conference on Information and Communication Technology Convergence (ICTC), pages 1560–1562, 2020.

6) Y. Pan, R. C. Zheng, J. Zhang, and X. Yao. Predicting bike sharing demand using recurrent neural networks. Procedia Computer Science, 147:562–566, 2019. 2018 International Conference on Identification, Information and Knowledge in the Internet of Things.

7) G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.