Predicting Bike-Sharing Demand: A Machine Learning Approach for Urban Mobility Analysis

Md Javed Ahmed Shanto[§], Rubina Akter[†], Dong-Seong Kim[§], and Taesoo Jun[§]
Department of IT Convergence Engineering[§], ICT Convergence Research Center[†],
Kumoh National Institute of Technology[§†], Gumi 39177, South Korea
(shantoa729[§], rubinaakter2836[†], dskim[§], taesoo.jun[§])
@kumoh.ac.kr

***Abstract*—Due to the increasing environmental pollution and the greenhouse effect, individuals are increasingly inclined towards adopting environmentally friendly modes of transportation, such as electric cars and eco-friendly machinery. The bicycle is considered one of the most environmentally friendly modes of transportation due to its low cost and minimal environmental impact. The use of bicycles is steadily increasing daily, especially in urban areas. This study seeks to investigate the level of demand for bike-sharing services in urban areas. It will explore various feature engineering methods, thoroughly analyze ensemble machine learning techniques, and employ artificial intelligence for time analysis.**

***Index Terms*—Artificial intelligence, bike-sharing, machine learning**

I. Introduction

Eliminating pollution and greenhouse gas emissions is a concern that affects cities all around the world [1]. Barth and Boriboonsomsin estimate that the conveyance of people and products accounts for around one-third of the carbon dioxide produced in the US [2]. Many cities worldwide have created bike lanes to reduce the harm that vehicle emissions cause to the environment, and they're working toward incorporating more cutting-edge technologies to promote sustainable living. As a result, bike sharing has become more prevalent over the last few years. In addition to being a short-term replacement for private vehicles or taxis, sharing bikes have an extended lifespan and application potential [3]. While customers can use shared bikes to get handy services, numerous unpleasant aspects can affect how well customers ride. One of the apparent drawbacks is the unequal distribution of bike-sharing stations throughout the service region, which leads to an inconsistency in utilization. The uneven distribution of shared bicycles indicates too many bikes at some stations, meaning supply outweighs demand.

To address this issue, many academics suggest sharing bike usage predictions for better resource allocation. For instance, to anticipate the availability of bikes at each station, Froehlich et al. suggested four straightforward prediction models, such as maximum value, historical average, historical trend, and Bayesian network [4]. In order to predict the quantity of leased and returned bikes in the future, Li et al. [5] suggested using the hierarchical model for sharing bikes. The recent advancement of deep learning [6], such as convolutional neural networks (CNN) [7], [8], can analyze the spatial correlation of multivariate bike-sharing data and predict traffic volumes. However, their proposed solution ignores client preferences and uses geographical information for clustering.

Therefore, as an exploratory investigation, this study specifically focuses on the following:

1) This study presents a new method for preprocessing and manipulating bike-sharing data. The tasks involve identifying outliers, extracting features based on time, estimating wind speed using machine learning, and handling holidays.
2) This study presents an in-depth analysis of various regression models, including ensemble techniques such as Random Forest, Gradient Boosting, and AdaBoost, as well as traditional approaches like Decision Trees and Logistic Regression.
3) This work simplifies machine learning model training time assessment utilizing the time module for realistic real-time forecasts, model performance, and processing

The paper is organized in the following manner: Section I discusses the introduction, Section II presents the methodology to predict the bike sharing demand, Section III presents the result analysis, and finally, Section IV concludes the paper.

TABLE I
Dataset Description

Features	**Description**
datetime	Date and time displayed hourly.
season	1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday	Whether the day is considered a holiday
workingday	Whether the day is neither a weekend nor holiday
weather	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp	Temperature in Celsius
atemp	”Feels like” temperature in Celsius
humidity	Relative humidity
windspeed	Wind speed
casual	Non-registered user rentals
registered	how many rentals from registered users
count	Number of total rentals

II. Methodology

Python was utilized to execute the code, and the system comprises an Intel Core i9-10980XE processor, two NVIDIA

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

GeForce RTX 3090 graphics cards, and 128GB of RAM. The dataset [9] is utilized due to its inclusion of 12 columns, which represent the necessary features for bicycle usage in any specific area. The dataset's features are displayed in the table I after undergoing preprocessing to predict the demand for bike sharing. Following the system diagram of fig. 1, outliers are identified by calculating the absolute difference between the 'count' values and their mean. The 'train' dataset undergoes a process where outliers are eliminated. The code retrieves different time-related attributes from the 'datetime' column, including 'year', 'month', 'day', 'hour', 'week', and 'weekday'. These features offer supplementary information for the training of the model. Subsequently, to perform wind speed interpolation, we employ a Random Forest (RF) classifier that has been trained on various features such as 'season', 'weather', 'temp', 'atemp', 'humidity', 'hour', and 'month' to substitute missing values in the 'windspeed' variable. The missing 'windspeed' values are replaced with the predicted values. To account for regular working days and holidays, we adjust the values of 'working day' and 'holiday' for specific dates, including Tax Day, Thanksgiving, Christmas, rainstorms, and tsunami events. These adjustments are made in both the 'train' and 'test' datasets. Next, we modify the labels 'season' and 'weather' to more specific terms such as 'season2' and 'weather2' to simplify the one-hot encoding process at a later stage.

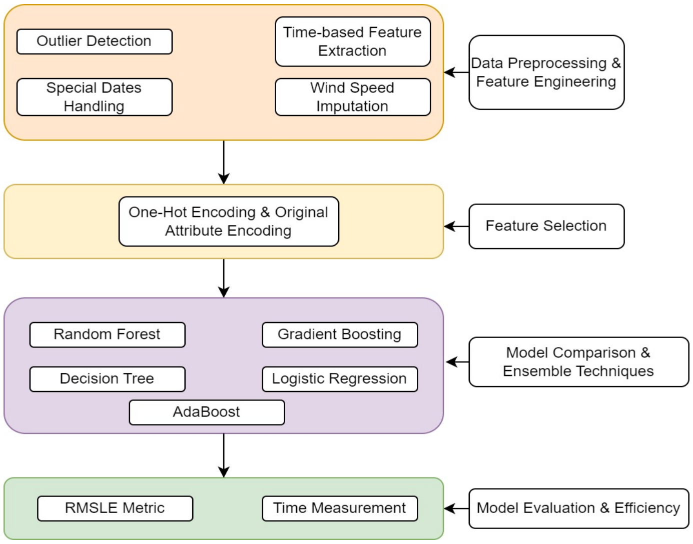
Outlier Detection
Time-based Feature Extraction
Special Dates Handling
Wind Speed Imputation
Data Preprocessing & Feature Engineering
One-Hot Encoding & Original Attribute Encoding
Feature Selection
Random Forest
Gradient Boosting
Decision Tree
Logistic Regression
AdaBoost
Model Comparison & Ensemble Techniques
RMSLE Metric
Time Measurement
Model Evaluation & Efficiency

Fig. 1. System Diagram of this study

Subsequently, we visually examine the dataset, investigating the associations among different characteristics. We examine various factors that influence the usage patterns of a cycle-sharing system. These factors include the variation in usage on an hourly basis fig. 2, the identification of peak hours fig. 2, and the variation in use monthly fig. 3, the variation in usage based on different seasons, the impact of weather conditions on usage, the influence of weekdays and working days on use, the hourly usage patterns on different weekdays, the usage patterns of registered and casual users on an hourly basis, the impact of temperature on cycle sharing demand, the influence of humidity and wind speed on usage, the analysis of correlations between different variables, the analysis of the distribution of usage data, and the application of log transformation to the data. Next, we utilized one-hot encoding to represent categorical variables. Feature selection involves the creation of distinct sets of feature columns using fig. 4 for different regression models. Next, we apply log transformations to the target variables 'casual', 'registered', and 'count' to enhance the modeling process.

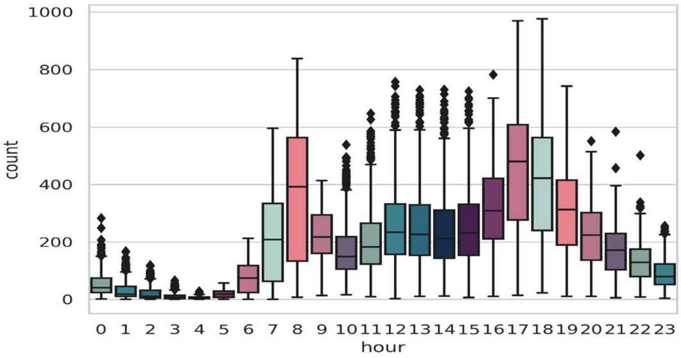

Fig. 2. Analysis of bike at different hours

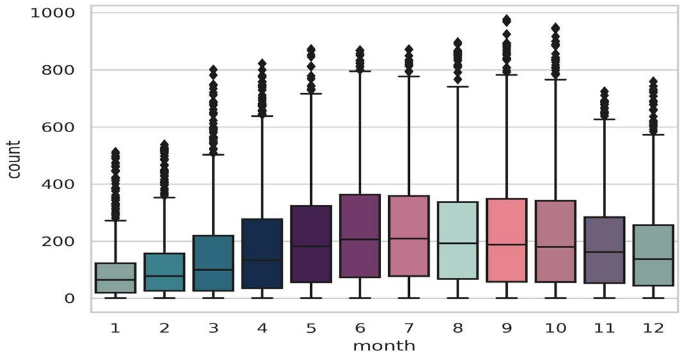

Fig. 3. Analysis of bike in different months

III. Experimental Results and Discussion

TABLE II
MODELS EVALUATION RESULTS

Model	RMSLE
LGBM	0.0859
Random Forest	0.0839
Decision Tree	0.2371
Gradient Boosting	0.0835
AdaBoost	0.1574

The measurement of evaluation (RMSLE), which stands for Root Mean Squared Logarithmic Error, is used to measure the performance of models. Afterward, we train and assess different regression models, such as LightGBM, Random

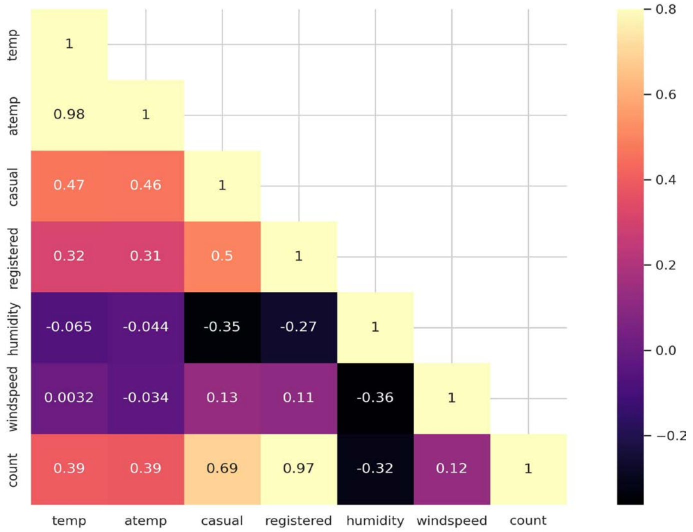

Fig. 4. Correaltion analysis of the features

Forest, Decision Tree, Gradient Boosting, Logistic Regression, and AdaBoost. Table II displays the RMSLE results for each model. The findings suggest that the models exhibit different levels of performance. The Random Forest and Gradient Boosting models demonstrated their effectiveness in accurately predicting bike rental counts by achieving the lowest RMSLE scores. On the other hand, the Decision Tree model exhibited a relatively higher error level, suggesting the possibility of overfitting. The AdaBoost model demonstrated a higher error rate, indicating the potential for further enhancements.

TABLE III
ANALYSIS OF RANDOM FOREST AND GRADIENT BOOSTING

Model	Fit Degree	Time (seconds)
Random Forest		
Model 1	0.9709	20.7781
Model 2	0.9805	21.2073
Model 3	0.9814	20.9171
Gradient Boosting		
Model 1	0.9199	1.2729
Model 2	0.9692	1.2321
Model 3	0.9692	1.2507

We further analyzed the performance of three customized models by tuning the hyperparameters of random forest and gradient boosting, as these algorithms have shown to produce the best results. The Random Forest models, specifically Model 1, Model 2, and Model 3, exhibited high levels of accuracy, with fit degrees ranging from around 97% to 98% (table III). This suggests that these models effectively captured the underlying patterns present in the bike rental data. Surprisingly, these models achieved high accuracy while maintaining consistent training times of approximately 20 seconds. In contrast, the Gradient Boosting models demonstrated marginally lower levels of accuracy, varying between approximately 92% and 97%. Nevertheless, the models demonstrated exceptional computational efficiency, as each achieved a satisfactory fit within an average time of approximately 1.25 seconds. These findings highlight the balance between the precision of a model and the amount of computational resources required.

IV. Conclusion

This study examines the demand for bike-sharing services in urban areas as an eco-friendly alternative. It also examines different preprocessing methods, feature engineering methods, and visual analysis techniques in depth. It also thoroughly analyzes ensemble machine-learning techniques and utilizes artificial intelligence for time analysis. In order to gain insights into the demand for bike sharing, this study can be analyzed using deep learning techniques. Additionally, explainable AI can provide a more comprehensive understanding of the subject for future research. To implement this in real life, real-time data processing and prediction can be a practical concern.pace-3mm

Acknowledgment

This work was supported by Priority Research Centers Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science & Technology (MEST) (2018R1A6A1A03024003), MSIT (Ministry of Science & ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2023-2020-0-01612) supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation), and the project for Industry-University-Research Institute platform cooperation R&D funded Korea Ministry of SMEs and Startups (S3311338).

References

[1] W. B. Blogs. (2023) Cutting global carbon emissions: Where do cities stand? 16/07/23. [Online]. Available: https://blogs.worldbank.org/sustainablecities/cutting-global-carbon-emissions-where-do-cities-stand

[2] M. Barth and K. Boriboonsomsin, "Traffic congestion and greenhouse gases," *Access Magazine*, vol. 1, no. 35, pp. 2–9, 2009.

[3] T. Xu, G. Han, X. Qi, J. Du, C. Lin, and L. Shu, "A hybrid machine learning model for demand prediction of edge-computing-based bike-sharing system using Internet of Things," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7345–7356, 2020.

[4] J. Froehlich, J. Neumann, N. Oliver *et al.*, "Sensing and predicting the pulse of the city through shared bicycling," in *IJCAI*, vol. 9, no. Jul, 2009, pp. 1420–1426.

[5] Y. Li, Y. Zheng, H. Zhang, and L. Chen, "Traffic prediction in a bike-sharing system," in *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*, 2015, pp. 1–10.

[6] M. J. A. Shanto, E. A. Tuli, R. Akter, D.-S. Kim, and T. Jun, "Federated learning empowered spam message detection for multilingual short message service (sms)."

[7] R. Akter, M. Golam, V.-S. Doan, J.-M. Lee, and D.-S. Kim, "IoMT-Net: Blockchain-Integrated Unauthorized UAV Localization Using Lightweight Convolution Neural Network for Internet of Military Things," *IEEE Internet of Things Journal*, vol. 10, no. 8, pp. 6634–6651, 2022.

[8] M. Golam, R. Akter, J.-M. Lee, and D.-S. Kim, "A long short-term memory-based solar irradiance prediction scheme using meteorological data," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[9] Kaggle, "Bike sharing demand dataset," https://www.kaggle.com/competitions/bike-sharing-demand/data, accessed: 03/07/2023.