Dean&Francis

ISSN 2959-6157

Demand forecasting for bike sharing based on Bayesian XGBoost

Zhuhao Huang

School of Computer Science, Zhuhai College of Science and Technology, Zhuhai, Guangdong, 519000, China
Email: 15916323280@163.com

Abstract:

A prediction model of shared bicycle rental volume based on the Bayesian-XGBoost algorithm is developed to address the issue of the imbalance between the supply and demand of shared bicycles. Correlation analysis is first used to investigate the characterization, followed by hyperparameter tweaking and assessing the model's stability and adaptability. The comparison study confirms that the Bayesian-optimized model has a greater forecast accuracy than a single model in shared bicycle rental volume prediction. The tests employ shared bicycle trip data in Washington, D.C., U.S.A.

Keywords: Bicycle-Sharing, Xgboost, Demand Forecasting, Bayesian Optimization

I. Introduction

Bicycle sharing is a mode of transportation for short-distance travel, which provides an environmentally friendly and active commuting option [1]. However, management problems arise as the scale of shared bikes and users increases. In order to solve it, accurate demand prediction models are crucial. Extreme Gradient Boosting (XGBoost) is an efficient ensemble learning method widely used in house price prediction due to its strong performance under complex data [2]. In this post, we conducted experiments using the XGBoost model with Bayesian optimization (Bayesian-XGBoost) based on real bike-sharing data. Promising experimental results were achieved. Existing studies such as Sun et al. [3], Roset Cardona [4], and Yin et al. [5] have also demonstrated the effectiveness of machine learning and XGBoost in demand prediction of shared bicycles, which provides valuable support for advancing research in this area.

II. Literature Review

A. Limitation Overview

Most early demand forecasts for bike sharing used traditional methods such as random forests and spate-temporal sequences, which can capture some correlations but also have limitations. Random forests cannot characterize spate-temporal dependencies and cope with complex nonlinear dynamics. Spate-temporal sequence models have a high demand for data stability and can produce errors in rental demand forecasting. They limit the prediction accuracy of the model when dealing with large-scale features. The latest approaches use deep learning to effectively model complex models and graph neural networks to capture spatial relationships in the bike-sharing network to improve robustness. All of these methods significantly improve the prediction accuracy.

B. Literature Review

Researchers both domestically and internationally

ISSN 2959-6157

have advanced the field of shared bicycle demand forecast and put out a number of strategies. For instance, Pan Y. et al. [6] built a prediction model based on a recurrent neural network, Xu, Chengcheng et al. [7] employed LSTM, and Xu, Jianyang et al. [8] used GCN to increase the prediction accuracy. The basis for predicting demand for shared bicycles has been established by this research.

III. Research Methods

Using regularization and a second-order Taylor expansion of the cost function, the gradient-boosting decision tree-based integrated machine learning method XGBoost avoids overfitting. In order to effectively find the best hyperparameters, this work adds Bayesian optimization to XGBoost utilizing Gaussian process regression and acquisition functions. The method is applicable in real-world circumstances and enhances model performance and accuracy.

A. XGBoost Model

XGBoost is based on the Gradient Boosting framework and improves on the GBDT (Gradient Boosting Decision Tree) algorithm, which models prediction by integrating multiple decision trees.

· XGBoost builds additive models by combining multiple decision trees.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

K is the total number of trees, the output of the K^{th} tree is $f_k(x_i)$, and the predicted value of x_i is $\hat{y}_i$.

· The objective function of XGBoost consists of a loss term and a regularization term.

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

$l(y_i, \hat{y}_i)$ tests the prediction error, while $\Omega(f_k)$ is related to the complexity of the model.

· XGBoost uses a second-order Taylor expansion to better optimize the objective function based on current predictions.

$$\begin{aligned} Obj(\theta) \quad &\approx \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] \\ &+ \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \end{aligned} \tag{3}$$

T is the number of leaf nodes, the score of the jth leaf is w_j , γ and λ are regularized hyperparameters. The first and second order derivatives of the loss function correspond to Gain and hit。

· XGBoost uses a regularization term to place a limit on the complexity of the decision tree. Its formula is:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

In this equation, T is the number of leaves, the score of the jth leaf is w_j, γ limits the number of leaves, and λ controls the score.

B. Bayesian Optimization

Bayesian optimization is a model-based approach for optimizing costly black-box functions through exploration and exploitation.

· Finding the optimal set of hyperparameters is at the heart of Bayesian optimization.

$$x^* = arg \max_{x \in X} E\left[f(x)\right] \tag{5}$$

E[f(x)] is the expected value predicted by the model and x represents the search space.

· The Expected Improvement (EI) function estimates how much the new result can be improved compared to the current result.

$$EI(x) = E\left[max\left(0, f(x) - f\left(x^+\right)\right)\right] \tag{6}$$

$f(x^+)$ is the current best value and the acquisition function selects a new candidate point based on the improved value.

C. Data Descriptions

Our study uses publicly available data from the University of California, Irvine (UCI) on shared bicycle trips from January 2011 through December 2012 in the Washington, DC area. The dataset contains 10,886 records and 12 features; the target variable "count" represents the total number of rentals. Table 1 lists the details of the dataset attributes. The "Datetime" feature records the hourly rental time. "Season" codes the season as 1 through 4 (spring through winter). "Holiday" and 'Working day' are binary indicators (0: no, 1: yes). "Weather" categorizes weather conditions into sunny (1) to heavy precipitation (4). "Temp" and 'Atemp' measure actual and perceived temperature in degrees Celsius. "Humidity" and 'wind speed' reflect environmental factors. "Casual" and 'Registered' indicate the number of rentals for users without and with registration. More detailed information is shown in the Table 1 below.

Table 1 Characteristics of raw data

Feature	Description
datetime	Hourly rental time record.
season	Season code (1: Spring, 2: Summer, 3: Fall, 4: Winter).
holiday	Indicates if the day is a holiday (0: No, 1: Yes).
workingday	Indicates if the day is a working day (0: No, 1: Yes).
weather	Weather condition (1: Clear, 2: Mist, 3: Light Snow/Rain, 4: Heavy Rain/Snow/Fog).
temp	Actual temperature in Celsius.
atemp	Perceived temperature in Celsius.
humidity	Relative humidity percentage.
windspeed	Wind speed
casual	Number of non-registered user rentals
registered	Number of registered user rentals
count	Number of rentals by non-registered users.

Data characteristics

In order to keep all the features in the same range, we normalize the values, whose eigenvalues are in the range [0, 1] with the following equation.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (7)$$

X is the original value, X_{norm} is the normalized value, X_{min} and X_{max} are the minimum and maximum values of the features.

D. Characteristic Importance

To analyze relationships between features, this study uses the Pearson correlation coefficient, calculated as follows.

$$\rho(x,y) = \frac{\sum_{i=1}^{n}(x_i - x?)(y_i - ?)}{\sqrt{\Sigma_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\Sigma_{i=1}^{n}\left(y_i - ?\right)^2}} \quad (8)$$

n is the number of samples, the eigenvalues of the ith sample are x_i and y_i, and $\bar{x}$ and $\bar{y}$ are their means.

The feature correlation heatmap expresses several correlation features. temp and atemp show a robust positive correlation of 0.99, revealing they are nearly identical. Total rentals correlate strongly with registered users (0.97) and moderately with casual users (0.71). A notable negative correlation of -0.70 exists between weekday and workingday, suggesting fewer casual rentals on weekdays. Additionally, temp and humidity show a weak negative correlation of -0.05. Detailed results are shown in Figure 1.

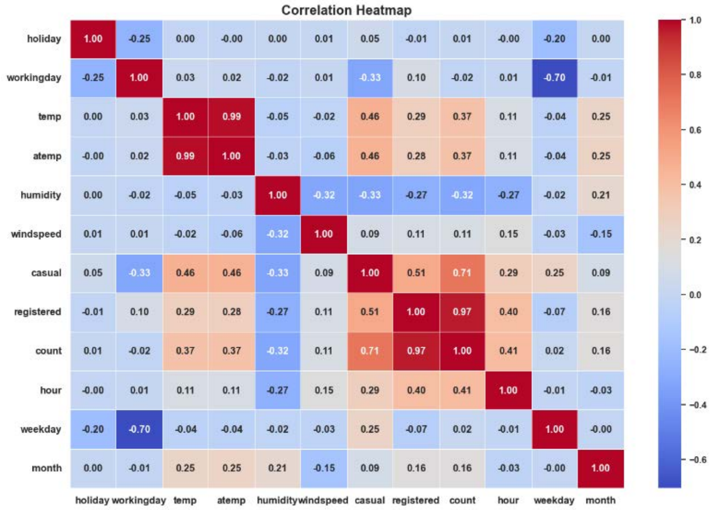

Fig 1 Feature Correlation Heatmap

ISSN 2959-6157

We analyze feature relevance, train the model, and assess the impact of each attribute on predictions, illustrated in Figure 2.

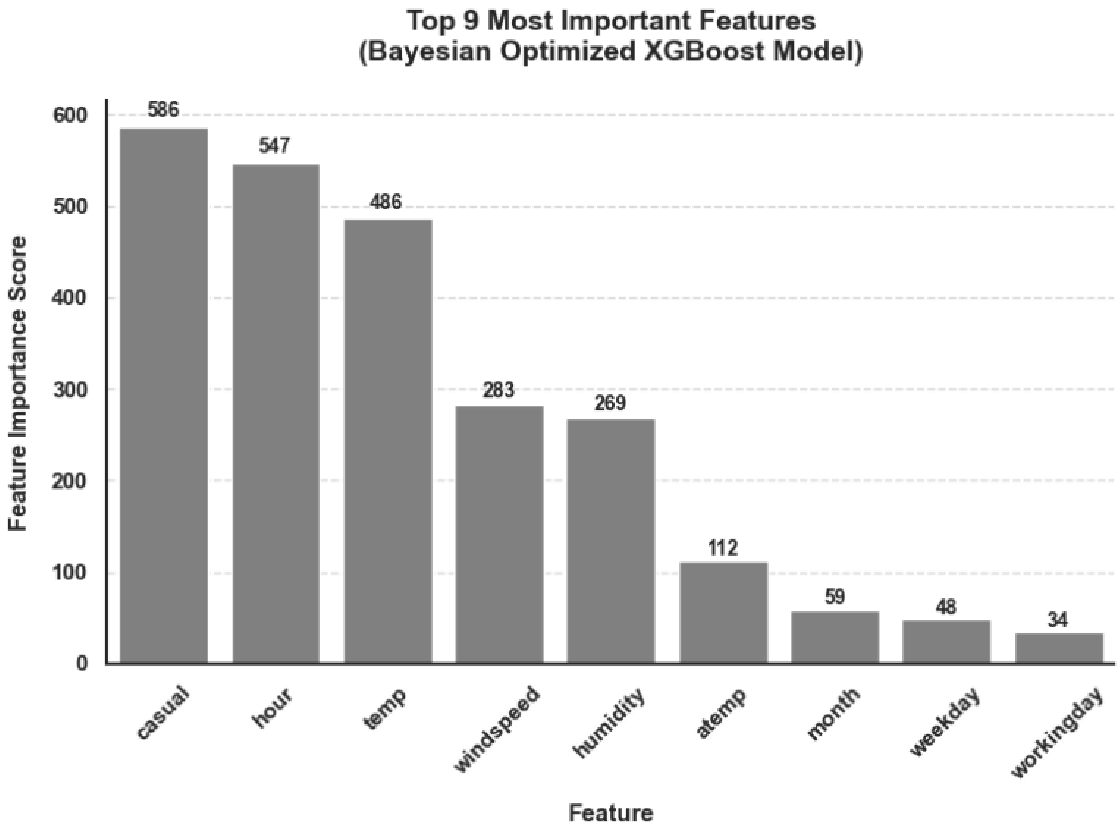

Fig 2 Feature contribution

Figure 2 shows that “casual”, “hour”, and “temp” are the most important features, with importance scores about 2.1 times higher than “windspeed” and “humidity”. The main factors include time-related variables (month, weekday) and weather conditions (atemp, windspeed, humidity). In comparison, “workingday” has little effect, indicating weak predictive value. These results support the model’s interpretability and effectiveness.

IV. Results of the study

This section analyzes the main factors affecting shared bicycle rentals and proposes an improved model. The comparison shows that the improved model predicts more accurately.

A. Factor Analysis

The relationship between these variables and the number of shared bicycle rentals was analyzed by examining their distributions under different conditions of temp, atemp, humidity, and windspeed, as shown in Figure 3.

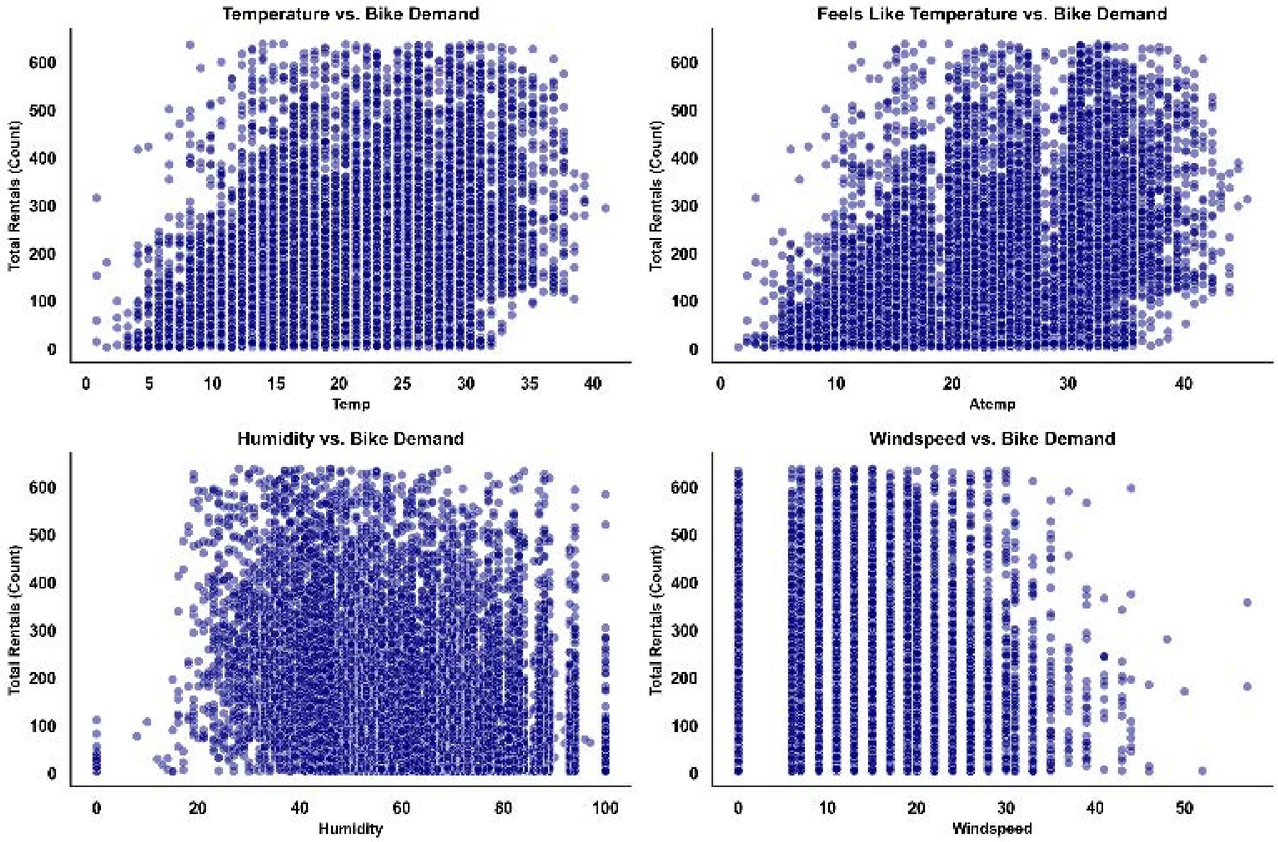

Fig 3 Continuous Features vs. Bike Rentals

Bicycle rental demand is strongly correlated with temp and atemp, peaking at 15°C-30°C. Its demand decreases sharply as humidity approaches 100%, but remains relatively stable between 20% and 80%. Windspeed has the greatest effect on demand, which is lowest at windspeed above 30. Therefore, more suitable temp and lower windspeed increase demand, while high humidity and strong winds decrease demand, which provides a useful reference for the management of bike sharing.

Next, as shown in Figure 4, we analyzed the effects of hour, season, weekday, and the number of non-registered user rentals(casual) on the bike-sharing market.

ISSN 2959-6157

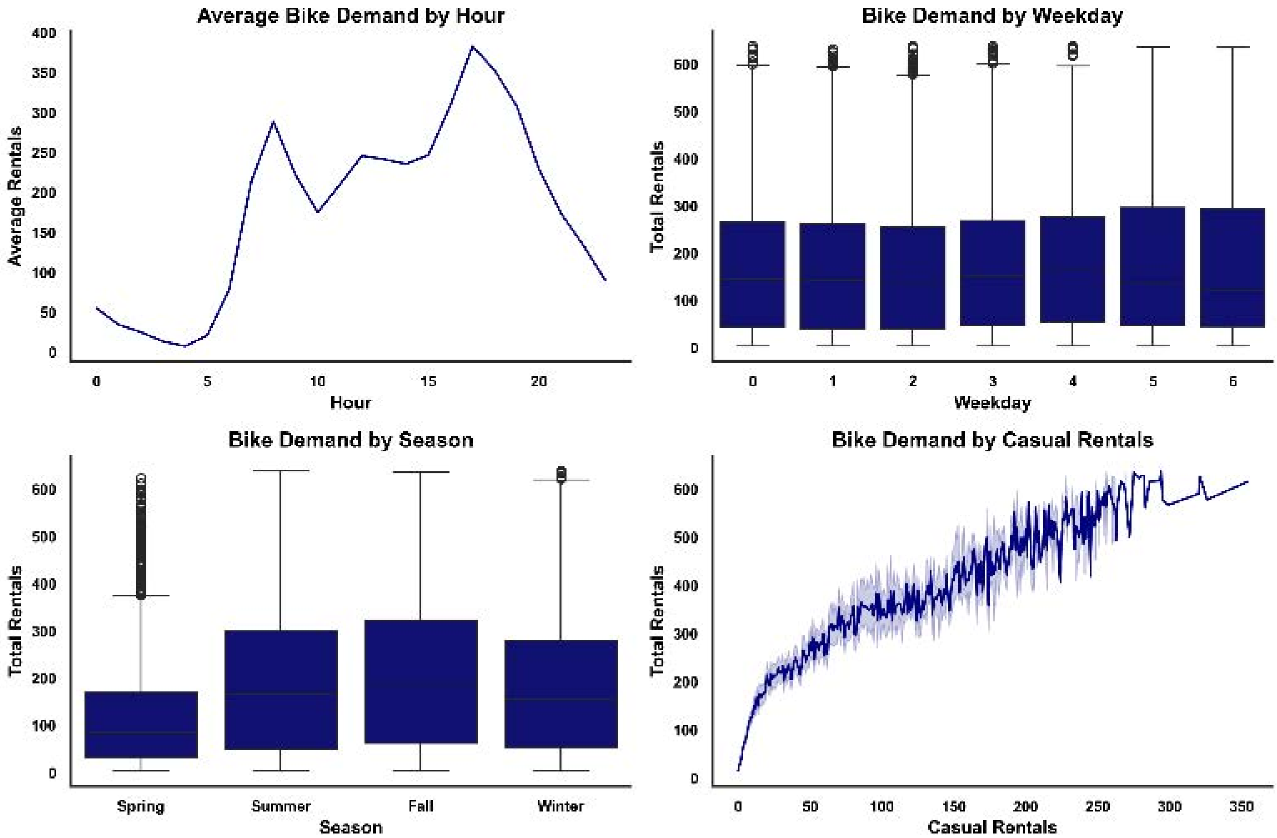

Fig 4 Discrete variables and rental distribution

As indicated in Figure 4, bike-sharing demand is influenced by hour, weekdays, seasons, and unregistered user numbers(casual). Peaks are used at 8:00 am and 5:00–6:00 pm. Demand increases on weekends and during summer and fall. Moreover, a rise in unregistered users leads to a notable growth in total rentals.

B. Model Comparisons

This study evaluates model performance using R², MAE, and RMSE, with their formulas listed below.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - ?)^2} \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - ?)^2} \tag{11}$$

In this formula, yi represents the actual observed value for the i-th data point. ŷi denotes the predicted value from the model for the same point. The variable n is the total number of data points in the dataset. The term ȳ refers to the mean of the actual values yi. These symbols compute R², MAE, and MSE for model evaluation.

The prediction results before and after model optimization are compared in Table 2 and Table 3, providing a theoretical basis for evaluating optimization effectiveness.

Table 2 Predictions before XGBoost optimization

	R_Squared	*RMSE*	*MAE*
XGBReggressor	0.928	41.599	28.404
Random Forest Classifier	0.918	44.572	28.886
DecisionTreeRegressor	0.862	57.770	40.750
Gradient Bossting	0.853	59.483	37.581
Extra Tree	0.846	60.996	38.558

Pre-optimization result.

Table 3 Predictions after XGBoost optimization

	R_Squared	RMSE	MAE
XGB Reggressor	0.944	36.758	25.238
Random Forest Classifier	0.927	41.849	26.835
Decision TreeRegressor	0.918	44.477	28.783
Gradient Bossting	0.862	57.774	40.751
Extra Tree	0.858	58.524	37.056

After-optimization result.

We compare five regression models: XGBRegressor, Random Forest Classifier, Decision Tree Regressor, Gradient Boosting, and Extra Tree. XGBRegressor has the best performance, with the optimized R^2 increasing from 0.928 to 0.944, and the RMSE and MAE decreasing to 36.758 and 25.238. Random Forest Classifier and Decision TreeRegressor are also improved, however, Gradient Boosting and Extra Tree have limited room for improvement. The post-processing model significantly improves the performance. This method is better than the pre-optimization model.

C. Model Performance Validation

To validate the Bayesian-optimized model, we plotted residual histograms and kernel density estimates. The prediction results for bike rentals are shown in Figure 5.

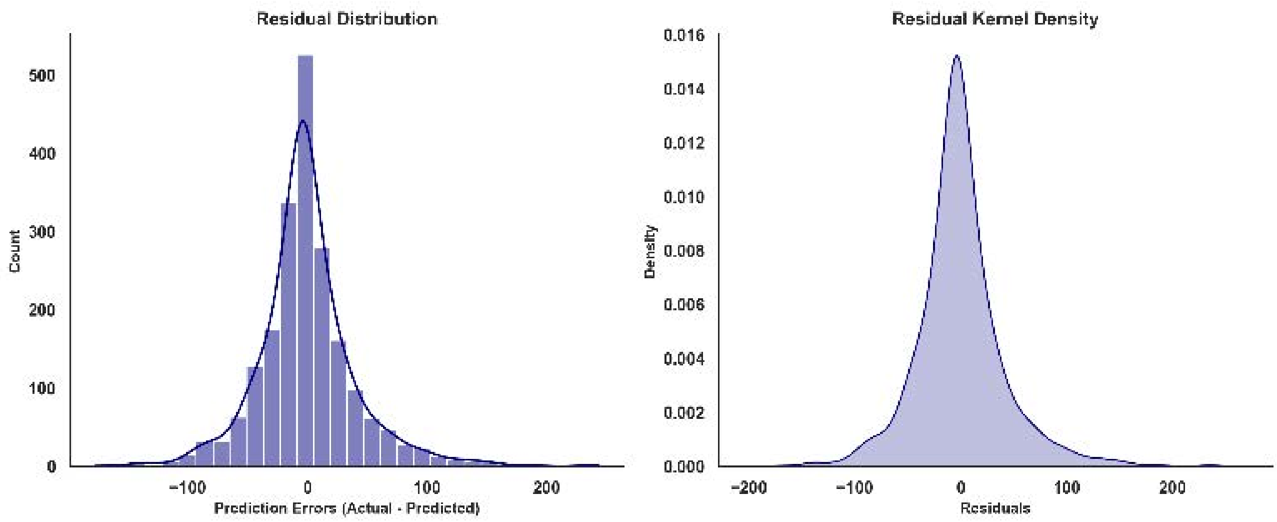

Fig 5 Prediction Error Distribution and Density

The residual analysis shows that most errors are concentrated around 0, and the symmetrical single-peak curve indicates high accuracy. However, the long tails on both sides of the model indicate that the errors are in an extreme state.

Conclusion

In this experiment, we use a Bayesian-optimized XGBoost model with UCI public data to predict bike-sharing rentals. Key factors such as temperature, wind speed, weather, time of day, weekday, and season significantly impact rental volume. After preprocessing and characterizing the data, the significant factors were selected and modelled. The results show that the Bayesian optimized model outperforms the other models by increasing R^2 to 0.944 and decreasing RMSE and MAE to 36.758 and 25.238. However, this model has problems when facing other influencing factors. In future studies, the accuracy can be improved by improving optimizing outlier handling.

References

[1] Jensen, P., Rouquier, J.-B., Ovtracht, N., Robardet, C. (2010) Characterizing the speed and paths of shared bicycle use in Lyon. Transportation Research Part D: Transport and Environment, 15: 522–524.

[2] Peng, Z., Huang, Q., & Han, Y. (2019). Model research on forecast of second-hand house price in Chengdu based on XGBoost algorithm. Proceedings of the 2019 IEEE International

ISSN 2959-6157

Conference on Artificial Intelligence and Information Technology (ICAIT), 168–172.
[3] Sun, Yibo, Ma, J., Zhang, R., Lyu, Q. (2025). Demand Analysis and Prediction of Shared Bicycle Based on Machine Learning. In Proceedings of the 4th Asia-Pacific Artificial Intelligence and Big Data Forum (AIBDF ‘24), Association for Computing Machinery, New York, NY, USA, 1014–1021.
[4] Roset Cardona, J. (2024). BiciMAD: Bike-sharing system dataset creation and preliminary hourly demand prediction using machine and deep learning approaches. Universitat Politècnica de Catalunya.
[5] Yin, Ganmin, Fu, C., Ren, S., Yan, X., Qi, J., Bao, Y., Huang, Z. (2025). Traffic prediction and road space optimization for the integration of dockless bike-sharing and subway. Sustainable Cities and Society, 121: 106162.
[6] Pan, Y., Zheng, R. C., Zhang, J., Yao, X. (2019) Predicting bike sharing demand using recurrent neural networks. Procedia Computer Science, 147: 562–566.
[7] Xu, Chengcheng, Ji, Junyi, Liu, Pan. (2018) The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. Transportation Research Part C: Emerging Technologies, 95: 47–60.
[8] Xu, Jianyang, Qian, Yi, Zhang, Song, Wu, Cunchuan. (2023) Demand Prediction of Shared Bicycles Based on Graph Convolutional Network-Gated Recurrent Unit-Attention Mechanism. Mathematics, 11: 4994.