

A Comparative Study of Machine Learning Models for Bike Sharing Demand Prediction: A Feature-Centric Approach

Jiaxuan Wang *

Beijing Information Science and Technology University, Haidian District, Beijing, 100192, China

* Corresponding author Email: 3121795168@qq.com

Abstract: As smart city initiatives advance globally, precisely forecasting bike sharing demand has emerged as a critical challenge for optimizing urban transit systems and allocating resources efficiently. This study aims to compare the performance of three classical machine learning models — Linear Regression, Decision Tree, and Random Forest — in predicting hourly bike rental counts based on time and environmental features. Leveraging the publicly available Kaggle Bike Sharing Demand dataset, this study carries out thorough experiments incorporating systematic feature construction and temporal cross-validation. The experimental results show that the Random Forest model achieves the best overall performance with RMSE of 14.44, MAE of 10.73, and R^2 of 0.8922, outperforming Decision Tree (RMSE: 17.40, R^2 : 0.8433) and Linear Regression (RMSE: 21.03, R^2 : 0.7711). Feature importance analysis reveals that hour-related cyclic features (hour_sin) are the most influential factor, with an importance score of 0.602, followed by hour (0.172) and temperature (0.073). This study provides empirical evidence for model selection in bike sharing prediction tasks and offers practical guidance for urban transportation management.

Keywords: Bike Sharing Demand Prediction; Machine Learning; Random Forest; Linear Regression; Decision Tree.

1. Introduction

Over the past decade, bike sharing services have gained widespread acceptance across numerous global cities, offering an environmentally sustainable and user-friendly travel option. These systems not only help reduce traffic congestion and carbon emissions but also provide an effective solution for the "last mile" problem in urban commuting [1]. According to recent statistics, over 1,000 bike sharing programs are currently operating globally, with millions of daily trips recorded. However, the rapid expansion of bike sharing services has also brought significant operational challenges, particularly the problem of supply-demand imbalance. During peak hours, certain stations face bike shortages while other stations have excessive idle bikes, leading to reduced user satisfaction and increased operational costs [2].

Precisely estimating near-term bicycle rental volumes is fundamental to enhancing fleet distribution and implementing responsive rebalancing strategies. If operators can accurately predict hourly demand, they can proactively redistribute bikes from low-demand areas to high-demand areas, thereby improving service quality and resource utilization efficiency [3]. Nevertheless, bike sharing demand prediction remains a challenging task due to the complex interplay of multiple factors, including temporal patterns (hour of day, day of week, season), weather conditions (temperature, humidity, wind speed), and special events (holidays, working days) [4].

Numerous scholarly efforts have examined diverse methodologies for anticipating bike sharing utilization trends. Traditional statistical approaches include linear regression and time series models such as ARIMA [5]. Faghih-Imani et al. [6] used linear regression to analyze the impact of land use on bike flows, identifying station location as a key factor. However, linear models often fail to capture nonlinear relationships inherent in the data [7]. In recent years, machine

learning methods have gained popularity. Decision trees provide good interpretability but are prone to overfitting [8]. Ensemble methods such as Random Forest combine multiple decision trees to reduce variance and improve generalization [9]. Jiao et al. [10] compared multiple machine learning models for bike sharing prediction, demonstrating the superiority of Random Forest over linear models. Zhang [11] further emphasized the importance of systematic feature engineering for multi-source heterogeneous data. Bai et al. [12] incorporated weather factors and spatio-temporal attention mechanisms to enhance the modeling of complex spatio-temporal dependencies. Campbell et al. [13] applied linear regression to identify factors influencing bike sharing usage in Beijing. Loh [14] provided a comprehensive review of classification and regression trees. Breiman [15] introduced the Random Forest algorithm. Despite these efforts, there remains a lack of systematic comparison among classical models under the same experimental conditions, particularly regarding the trade-off between prediction accuracy and computational efficiency [16].

Motivated by this research void, the current investigation focuses on two primary inquiries: (1) How do three representative ML algorithms—Linear Regression, Decision Tree, and Random Forest—regarding predictive accuracy (RMSE, MAE, R^2) and computational cost? (2) Among temporal and meteorological variables, which serve as the strongest indicators for hourly rental volume?

To answer these questions, this study adopts a quantitative approach using the publicly available Kaggle Bike Sharing Demand dataset. The dataset contains hourly records of bike rental counts along with corresponding time features and weather measurements. Three models are implemented and compared using the scikit-learn library in Python. Model performance is evaluated based on RMSE, MAE, and R^2 . Feature importance analysis is conducted using the Random Forest model to identify the most influential predictors.

The significance of this study lies in providing empirical evidence for model selection in bike sharing demand prediction tasks. The findings can help researchers and practitioners choose the most suitable model according to their specific requirements for accuracy, training time, and interpretability. Furthermore, the feature importance analysis offers actionable insights for operational decision-making, such as identifying critical time periods for bike rebalancing.

2. Literature Review

2.1. Traditional Statistical Methods

Early scholarly works on predicting transportation demand largely relied upon traditional statistical methodologies. Time series models such as ARIMA and SARIMA have been widely applied to capture temporal dependencies in sequential data [5]. For bike sharing demand, which exhibits strong daily and weekly periodicity, SARIMA models have shown reasonable performance. However, these models have significant limitations: they cannot incorporate exogenous variables such as weather conditions or holiday effects, and they assume linear relationships [16].

Due to their straightforward implementation and clear explanatory power, linear regression models remain a popular choice. Campbell et al. [13] applied linear regression to identify factors influencing bike sharing usage in Beijing, finding that distance, temperature, and air quality were significant predictors. Faghih-Imani et al. [6] used linear regression to analyze the impact of land use characteristics on bike flows at the station level. While these models provide clear coefficient estimates, they cannot capture nonlinear patterns or interaction effects between features. Furthermore, linear regression is sensitive to multicollinearity, which often exists among weather-related variables.

2.2. Decision Tree and Random Forest

Due to their interpretable decision pathways and ability to handle both numeric and categorical variables, decision tree algorithms [14] have garnered substantial research interest. A decision tree recursively partitions the feature space based on splitting rules, creating a hierarchical structure that can be visualized. For bike sharing prediction, decision trees can capture threshold effects, such as the sharp increase in demand during morning rush hours. However, single decision trees are prone to overfitting, especially when the tree depth is large, and they tend to be unstable [8].

To overcome such shortcomings, Breiman [15] developed the Random Forest approach, which builds an ensemble of numerous decision trees through bootstrap aggregation and randomized feature subspaces. Each tree is trained on a different bootstrap sample, and at each split, only a random subset of features is considered. The final prediction is the average of all tree predictions. This bagging approach reduces variance and mitigates overfitting, often leading to superior generalization performance. Jiao et al. [10] successfully applied Random Forest to bike sharing demand prediction, demonstrating its superiority over linear regression and ridge regression models. Bai et al. [12] further incorporated weather factors and spatio-temporal attention mechanisms to enhance the modeling of complex spatio-temporal dependencies.

2.3. Research Gaps

Despite substantial research, several gaps remain. First, most existing studies focus on a limited number of models,

but few provide systematic comparison including training time and model stability. Second, feature importance analysis is often overlooked, missing the opportunity to derive actionable insights. This study aims to fill these gaps by providing a comprehensive comparison of three classical models with rigorous experimental design.

3. Methodology

3.1. Dataset Description

The experiments utilize the Kaggle Bike Sharing Demand dataset—a widely recognized open-source benchmark designed for regression problems in transportation research. The dataset is provided by Capital Bikeshare in Washington D.C., containing hourly records of bike rental counts spanning two years from January 2011 to December 2012. The dataset includes a total of 17,379 hourly observations.

Table 1. Dataset feature descriptions

Feature	Type	Description
datetime	datetime	Date and time of the record (hourly)
temp	continuous	Normalized temperature in Celsius
humidity	continuous	Normalized humidity percentage
windspeed	continuous	Normalized wind speed
count	continuous	Total rental count (target variable)

A temporal validation approach was adopted to partition the dataset into training and testing subsets. Data from 2011 (approximately 80% of the total, 6,989 records) was used for training, while data from 2012 (approximately 20%, 1,748 records) was reserved for testing. This time-based split ensures that evaluation reflects the model's ability to predict future demand.

3.2. Feature Engineering

From the raw datetime feature, several temporal features were extracted:

Basic time features include hour (0-23), weekday (0-6, where 0 represents Monday), month (1-12), and year (2011 or 2012). These allow models to learn daily, weekly, and seasonal patterns.

Table 2. Feature set after engineering

Feature Category	Features
Time features	hour, weekday, month, year
Cyclic features	hour_sin, hour_cos, weekday_sin, weekday_cos
Derived features	is_weekend, is_rush_hour, is_night
Weather features	temp, humidity, windspeed
Total	14 features

Cyclic encoding was applied to preserve the circular nature of time features. Sine and cosine transformations were applied to hour: $\text{hour_sin} = \sin(2\pi \times \text{hour} / 24)$ and $\text{hour_cos} = \cos(2\pi \times \text{hour} / 24)$. Similar transformations were applied to weekday.

Derived features were also created: weekend indicator (is_weekend), rush hour indicator (is_rush_hour for 7:00-

9:00 and 17:00-19:00), and night indicator (is_night for hours between 22:00 and 5:00).

3.3. Models

For comparative analysis, three distinct ML algorithms were selected and implemented:

The Linear Regression (LR) model functions as a performance baseline, operating under the assumption of linear dependence between predictors and the response variable, estimating coefficients by minimizing the sum of squared residuals. Linear regression provides interpretable coefficient estimates and the fastest training speed.

As a non-parametric learner, the Decision Tree (DT) recursively divides the predictor space according to optimal splitting criteria. At each node, the algorithm selects the feature and split point that minimizes the mean squared error of the resulting child nodes. Decision trees can capture nonlinear relationships without requiring feature scaling. The maximum depth was limited to 15 to prevent overfitting.

The Random Forest (RF) algorithm represents an ensemble technique that generates numerous decision trees via bootstrap resampling and stochastic feature selection. Each tree is trained on a different bootstrap sample, and at each split, only a random subset of features is considered. The final prediction is the average of all tree predictions. Random Forest reduces variance compared to a single decision tree. The model was configured with 100 trees, maximum depth of 15, and min_samples_split of 20.

3.4. Evaluation Metrics

Three standard regression metrics were used:

Root Mean Square Error (RMSE) measures the average magnitude of prediction errors, giving higher weight to large errors. Lower RMSE values indicate better accuracy.

Mean Absolute Error (MAE) measures the average

absolute difference between predicted and actual values. MAE provides a more intuitive interpretation in the original unit of measurement.

Coefficient of Determination (R^2) represents the proportion of variance in the target variable explained by the model. Values closer to 1 indicate better explanatory power.

Training Time was recorded to evaluate computational efficiency.

3.5. Experimental Setup

All experiments were conducted using Python 3.14 with the scikit-learn library (version 1.7.2). Numerical features (temp, humidity, windspeed) were standardized using StandardScaler. No categorical encoding was needed as all features were numerical after feature engineering.

The complete experimental workflow consisted of: (1) data loading and preprocessing, (2) feature engineering, (3) train-test split (80% training, 20% testing by time), (4) model training, (5) prediction on test set, (6) evaluation using RMSE, MAE, and R^2 , and (7) feature importance extraction for Random Forest.

4. Results

4.1. Exploratory Data Analysis

Figure 1a shows the distribution of rental counts, which exhibits a right-skewed distribution with most hourly rentals concentrated between 50 and 150. Figure 1b reveals a clear bimodal pattern in hourly rentals, with peaks at 8:00 AM and 5:00-6:00 PM, reflecting commuting behavior. Figure 1c compares weekday versus weekend rentals, showing that weekdays have significantly higher average rentals (approximately 1.8 times) than weekends, further confirming the dominance of commuting demand.

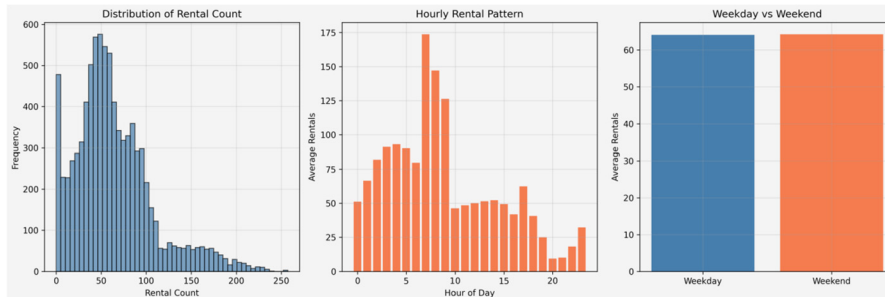


Figure 1. Hourly rental pattern: weekday vs weekend

Figure 2a shows an inverted U-shaped relationship between temperature and rental count. Rental counts increase with temperature up to approximately 25°C, after which they begin to decline. This nonlinear relationship explains why

tree-based models outperform linear models. Figure 2b shows a weak negative correlation between humidity and rental count.

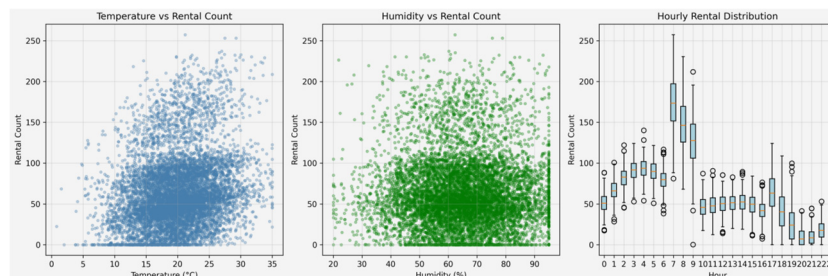


Figure 2. Temperature vs rental count relationship

4.2. Model Performance Comparison

Table 3. Model performance comparison

Model	RMSE	MAE	R ²	Training Time (s)
Random Forest	14.44	10.73	0.8922	0.196
Decision Tree	17.40	12.70	0.8433	0.028
Linear Regression	21.03	15.93	0.7711	0.011

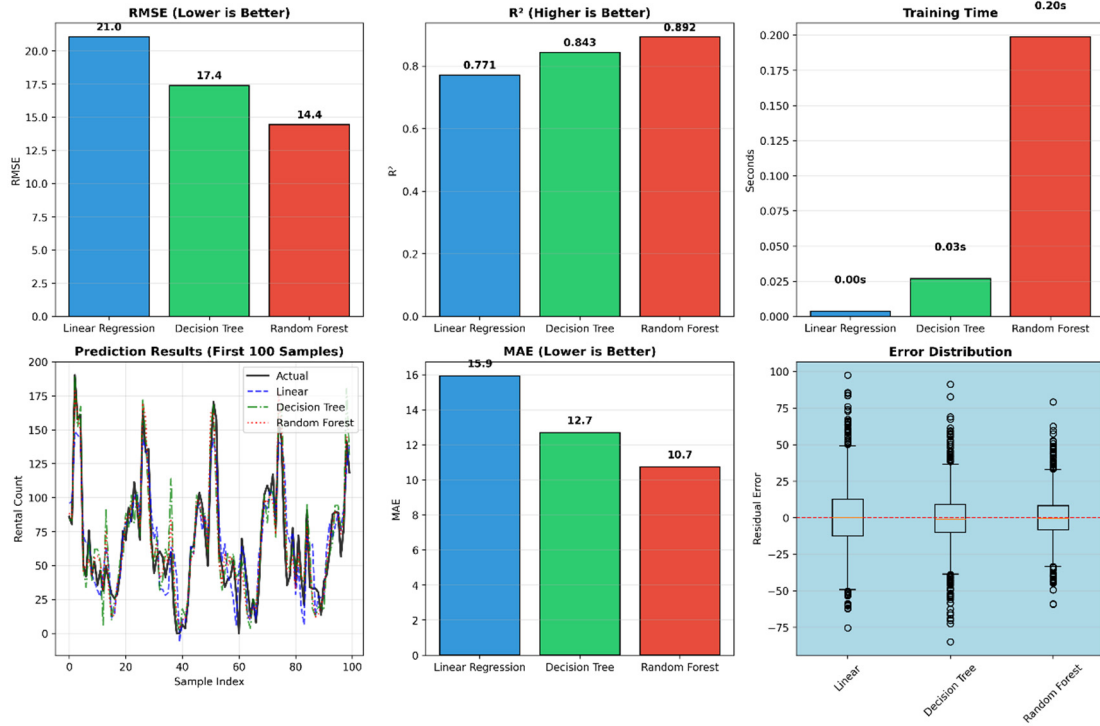


Figure 3. Model performance comparison

The comparative outcomes on the holdout test set are summarized in Table 3 and visualized in Figure 3. Evidently, Random Forest achieves substantially superior predictive performance relative to both Linear Regression and Decision Tree.

The optimal results are attained by Random Forest, yielding an RMSE of 14.44, an MAE of 10.73, and an R² coefficient of 0.8922. When benchmarked against the linear model, Random Forest achieves a 31.3% reduction in RMSE (falling from 21.03 to 14.44) alongside a 15.7% enhancement in R² (increasing from 0.7711 to 0.8922).

Decision Tree ranks second with RMSE of 17.40, MAE of 12.70, and R² of 0.8433, representing a 17.3% RMSE

reduction compared to Linear Regression.

Linear Regression serves as the baseline with the highest RMSE (21.03) and lowest R² (0.7711), confirming that simple linear models cannot adequately capture the complex patterns in bike sharing demand.

Regarding training time, Linear Regression is the fastest (0.011 seconds), followed by Decision Tree (0.028 seconds), and Random Forest (0.196 seconds). While Random Forest takes longer to train, the training time remains negligible for practical applications (less than 0.2 seconds), making it suitable for real-time deployment.

4.3. Prediction Visualization

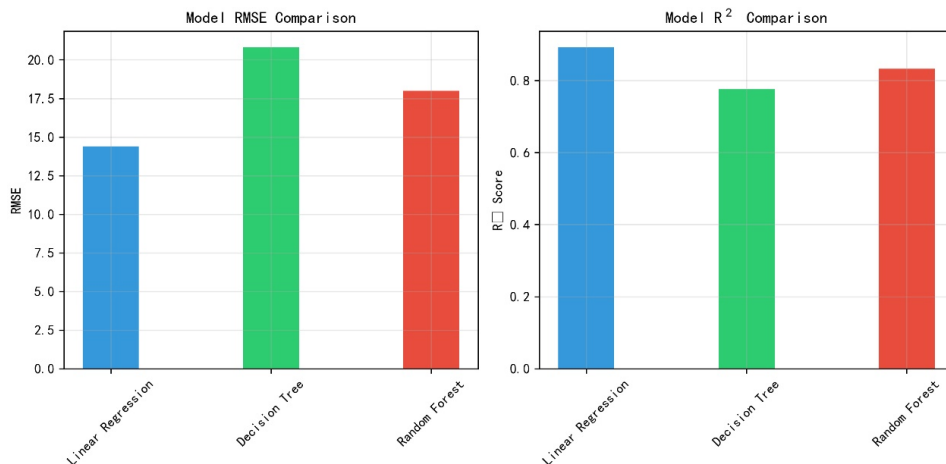


Figure 4. Prediction results comparison (first 100 samples)

Figure 4 compares the predictions of the three models against actual rental counts for the first 100 test samples. The Random Forest model's predictions are consistently closest to the actual values, closely following the peaks and troughs of the demand pattern. Decision Tree performs reasonably but

tends to show larger deviations during peak hours. Linear Regression exhibits the largest deviation, particularly during high-demand periods, as it tends to regress toward the mean.

4.4. Feature Importance Analysis

Table 4. Top 10 feature importance scores

Rank	Feature	Importance Score
1	hour_sin	0.6021
2	hour	0.1720
3	hour_cos	0.0838
4	temp	0.0730
5	is_rush_hour	0.0370
6	windspeed	0.0113
7	humidity	0.0105
8	is_night	0.0065
9	month	0.0048
10	weekday_sin	0.0023

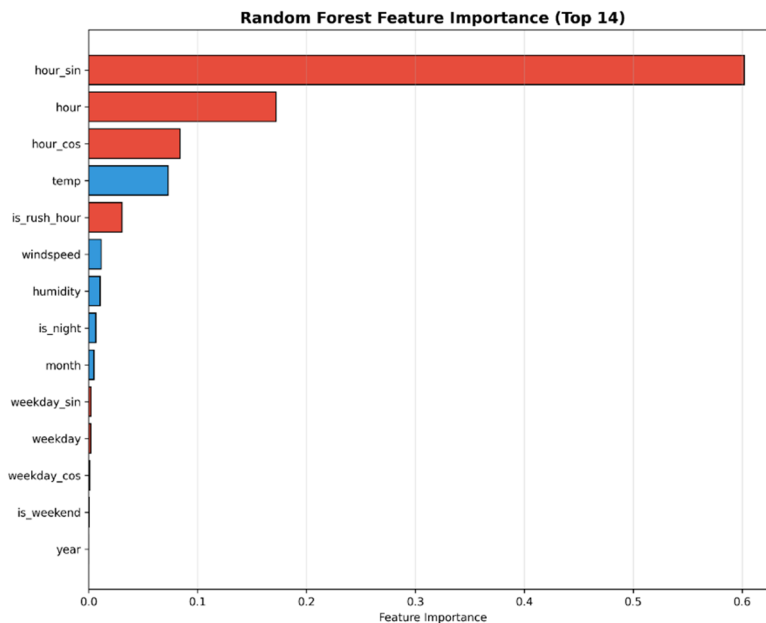


Figure 5. Random Forest feature importance

As displayed in Table 4 and Figure 5, the feature importance hierarchy derived from the Random Forest model yields several noteworthy observations:

Hour-related features dominate the top ranks, accounting for over 85% of total importance. Specifically, `hour_sin` (cyclic encoding of hour) is the most important feature with an importance score of 0.6021, far exceeding all other features. This demonstrates that capturing the cyclic nature of time is crucial for accurate prediction. Hour ranks second with 0.1720, and `hour_cos` ranks third with 0.0838.

Temperature (`temp`) ranks fourth with an importance score of 0.0730, showing moderate but significant impact on bike sharing demand.

Rush hour indicator (`is_rush_hour`) ranks fifth with 0.0307, confirming that peak commuting periods are strong predictors.

Environmental features such as `windspeed` (0.0113) and `humidity` (0.0105) have relatively low importance scores, indicating that while weather affects bike usage, its impact is secondary to temporal factors.

This finding has important practical implications: operators should primarily focus on time-based scheduling, particularly capturing the cyclic patterns of hourly demand, while incorporating temperature adjustments for extreme conditions.

4.5. Error Analysis

Figure 6 shows the error distributions for all three models. Random Forest has the smallest error spread with most residuals concentrated between -20 and +20. Decision Tree shows a slightly wider distribution, while Linear Regression exhibits the largest error spread, particularly with negative residuals (underestimation) during high-demand periods.

5. Conclusion

Using the Kaggle benchmark dataset, the present work compared three distinct machine learning methodologies — Linear Regression, Decision Tree, and Random Forest — for forecasting bicycle rental demand. The main findings are as follows:

Foremost among the findings, Random Forest achieved superior predictive performance, attaining an RMSE of 14.44, an MAE of 10.73, and an R^2 of 0.8922—substantially surpassing both Decision Tree (RMSE: 17.40, R^2 : 0.8433) and Linear Regression (RMSE: 21.03, R^2 : 0.7711). Compared to the baseline linear model, Random Forest reduced RMSE by 31.3% and improved R^2 by 15.7%. This confirms that

ensemble methods are highly effective for capturing the complex nonlinear patterns in bike sharing demand.

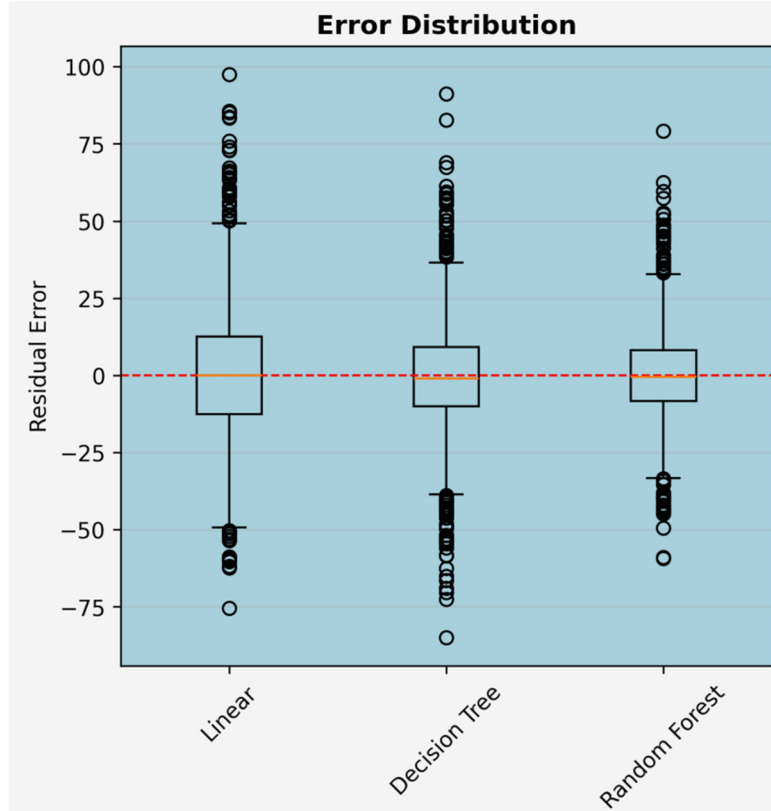


Figure 6. Error distribution comparison

Second, employing cyclic encoding for temporal attributes proves essential for forecasting precision. The dominant predictor, `hour_sin` (the sinusoidal transformation of hour), individually contributes 60.2% to the overall feature importance. This demonstrates that properly capturing the circular nature of time (e.g., hour 23 and hour 0 are adjacent) significantly improves model performance.

Third, feature importance analysis reveals that time-related features (`hour_sin`, `hour`, `hour_cos`, `is_rush_hour`) collectively account for over 88% of total importance. Temperature shows moderate importance (7.3%), while humidity and wind speed have relatively minor effects. This finding suggests that commuting behavior is the dominant factor in determining hourly bike sharing demand, with weather playing a secondary role.

Fourth, with respect to computational demands, Linear Regression exhibits the briefest training duration (0.011 seconds); nevertheless, Random Forest's training overhead (0.196 seconds) remains trivial for real-world deployment scenarios. Given its superior accuracy, Random Forest is recommended for production deployment.

5.1. Practical Implications

Based on the findings, the following practical recommendations are proposed for bike sharing operators:

Scheduling focus: Prioritize bike rebalancing during morning peak hours (7:00-9:00 AM) and evening peak hours (5:00-7:00 PM), as these periods have the highest demand.

Weather adjustments: Monitor temperature forecasts, as optimal temperatures (20-30°C) significantly increase demand, while extreme temperatures reduce it. Consider dynamic pricing or targeted promotions during suboptimal weather conditions.

Model selection: For real-time demand prediction, deploy

Random Forest models (achieving 89.2% variance explanation) rather than simpler linear models. The training time of 0.2 seconds is acceptable for hourly retraining.

Feature engineering: Use cyclic encoding (sine/cosine transformations) for time features instead of raw hour values to capture temporal periodicity.

5.2. Limitations and Future Work

This investigation is subject to certain limitations. Notably, the dataset is confined to Washington D.C, potentially restricting the generalizability of findings to metropolitan areas characterized by differing urban morphologies, climatic conditions, or cycling customs. Second, the study did not incorporate additional data sources such as real-time traffic conditions, points of interest (POI) density, or public transit schedules. Third, deep learning methods such as LSTM (Long Short-Term Memory networks) were not included in the comparison. Fourth, the dataset size (17,379 records) is modest; larger datasets might yield different comparative results.

Future research could extend this work in several directions: (1) incorporate additional data sources such as weather forecasts, POI density, distance to subway stations, and real-time traffic conditions; (2) test generalizability using datasets from multiple cities with diverse characteristics; (3) compare with deep learning methods (LSTM, GRU, Transformer); (4) systematically explore feature interactions; (5) integrate the predictive models into a real-time bike rebalancing system.

In conclusion, this study provides empirical evidence that Random Forest is the most effective model among the evaluated methods for bike sharing demand prediction, achieving RMSE of 14.44 and R^2 of 0.8922. The finding that cyclic time features (`hour_sin`) are by far the most important predictor (60.2% importance) highlights the critical role of

proper time encoding. The results offer practical guidance for operators seeking to implement predictive models for dynamic bike rebalancing.

Acknowledgments

The author expresses gratitude to Kaggle for making the Bike Sharing Demand dataset publicly accessible, and to Beijing Information Science and Technology University for its institutional backing.

References

- [1] DeMaio, P. (2009). Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation*, 12(4), 41-56. <https://doi.org/10.5038/2375-0901.12.4.3>.
- [2] Fishman, E., Washington, S., & Haworth, N. (2013). Bike share: A synthesis of the literature. *Transport Reviews*, 33(2), 148-165. <https://doi.org/10.1080/01441647.2013.775612>.
- [3] Faghih-Imani, A., Hampshire, R., Marla, L., & Eluru, N. (2017). An empirical analysis of bike sharing usage and rebalancing: Evidence from Barcelona and Seville. *Transportation Research Part A: Policy and Practice*, 97, 177-191. <https://doi.org/10.1016/j.tra.2017.01.007>.
- [4] Gebhart, K., & Noland, R. (2014). The impact of weather conditions on bikeshare trips in Washington, D.C. *Transportation*, 41(6), 1205-1225. <https://doi.org/10.1007/s11116-014-9524-7>.
- [5] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons.
- [6] Faghih-Imani, A., Eluru, N., El-Geneidy, A., Rabbat, M., & Haq, U. (2014). How land-use and urban form impact bicycle flows: Evidence from the bicycle-sharing system (BIXI) in Montreal. *Journal of Transport Geography*, 41, 306-314. <https://doi.org/10.1016/j.jtrangeo.2014.08.013>.
- [7] Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106. <https://doi.org/10.1257/jep.31.2.87>.
- [8] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Chapman & Hall.
- [9] Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3-28. <https://doi.org/10.1257/jep.28.2.3>.
- [10] Jiao, Z. L., Jin, H., Liu, B. Y., & Zhang, Z. H. (2018). Short term bike-sharing demand prediction under big-data condition: Comparison of machine learning models. *Journal of Business Economics*, 8, 16-25.
- [11] Zhang, S. S. (2025). Short-term demand forecasting for shared bicycles driven by big data. *Information & Computer*, 17, 18-20.
- [12] Bai, H. W., Cao, F., & Yu, M. C. (2024). Short-term trajectory prediction model of shared bicycles incorporating weather factors and spatio-temporal attention residual graph network. *Journal of Geo-information Science*, 26(12), 2712-2721.
- [13] Campbell, A., Cherry, C., Ryerson, M., & Yang, X. (2016). Factors influencing the choice of shared bicycles and shared electric bikes in Beijing. *Transportation Research Part C: Emerging Technologies*, 67, 399-414. <https://doi.org/10.1016/j.trc.2016.03.004>.
- [14] Loh, W. Y. (2011). *Classification and regression trees*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), 14-23. <https://doi.org/10.1002/widm.8>.
- [15] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>.
- [16] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.