City University of Hong Kong

Business Data Analytics

Analysis of the Demand and Trend Factors for Shared Bicycle Rental

Group Members:

CHEN Haonan –GAN Yuhang –HUANG Xin –YANG Yichen –ZHAO Jiayi

Executive Summary

This study investigates the daily demand of a public bike-sharing system and proposes forecasting models to aid managerial choices. The main issue is finding a proper way to distribute bicycles: too many bikes will result in high costs for maintenance and redistribution, whereas the opposite situation will bring about lost demand and hence, under-utilized assets. Our goal is to find the most important factors that influence daily rentals and to choose a forecasting model that is accurate, robust, and easy for the managers to use. We make use of the publicly available "Bike Sharing Dataset," consisting of 731 days from 2011 to 2012 along with total daily rentals and explanatory variables reflecting weather, calendar, and user types. The dataset is pristine and comprehensive. We divided the dataset into 80% for training and 20% for testing, maintaining the chronological order and aiming at forecasting the total daily rental count. An exploratory analysis shows obvious seasonal patterns and very strong correlations between demand and weather: summer and fall have much higher demand than winter, and warm, sunny days produce more rentals than cold, rainy, or snowy days.

Our initial model is a multiple linear regression that considers various factors like the weather and the calendar. It has already beaten two typical time-series methods and four advanced machine learning techniques on a shared test set. The complicated models do not enhance the accuracy since they only filter the noisy small dataset. Based on scatter plots, we proceed for the linear regression improvement with the inclusion of a squared temperature term to account for a nonlinear effect. This straightforward modification leads to a decrease in test RMSE and an increase in explanatory power, thus keeping the model transparent and easy to communicate. By means of the refined model, we generate 30-day and one-year-ahead predictions. The predictions reveal a consistent seasonal cycle plus a general increase in demand. The seasonal alteration of fleet size, weather-dependent operational scheduling, and the focused marketing intended to switch occasional users to registered members are among the measures we suggest based on the findings. Even though the analysis is constrained by just two years of system-wide data and the absence of price and competition data, it still shows that a properly defined linear model can yield accurate and easy to understand demand forecasts for bike-sharing operators.

1. Business Problem

Bike sharing has been a significant mode of transportation for short distances in cities, and it has been rapidly adopted in many cities. The use of shared bicycles in urban areas not only leads to a more effective and faster transportation system but also a significant reduction of carbon emissions coming from transportation systems in cities [1], [2]. Notwithstanding, the bike sharing eco-system on a large scale is still faced with the dilemma of heavy operational costs. The bike sharing industry is inherently a cycle of users turning away when the supply is not enough and companies losing money through high operating costs when the number of bikes is too great. It is thus the case that operators make the wrong choice time and again, alternating between over and under supplying the bikes. Over assuming the demand is greater than it is will lead to the idling of a large number of bicycles and incur substantial maintenance, repair, and redistribution costs for the company, and taxes on the whole business [3].

By way of contrast, if the supply of bicycles is not enough, or they are not located according to the demand, the customer waiting for the unavailable bike when needed. This lessens the user's pleasure and at the same time, it leads to the loss of a potential usage which ultimately results in a direct loss of revenue and lower platform utilization [4]. Thus, the balancing act between supply and demand becomes vital. Simply adding more bicycles to the fleet will not necessarily improve the operations and may even result in inefficiencies instead. Knowing what factors most affect the demand for rentals and being able to predict the bicycle-sharing usage for each day with high accuracy will allow the companies to come up with better operating strategies and thus reduce wastage. [5], [6].

In order to resolve these problems, the research zeroes in on two main questions:

1. What are the determinants of the daily demand for shared bicycles rentals?

2. What model can be regarded as the most appropriate for the purpose of predicting the upcoming demand?

At the end of the analysis, the study points out the primary factors that impact the demand for rentals and ranks the models according to their capability for precise prediction of demand. Moreover, the factors influencing the demand and the model's projection of bike rental for the next year are used as the basis for suggesting strategies. The revelations are of assistance to bike-sharing companies in determining the correct locations for their fleets, making

their allocation more efficient, and thus having the right overall operational strategies.

2. Data Description and Preprocessing

2.1 Data Source and Basic Structure

The data utilized in this study is from a publicly available dataset on Kaggle labeled as "Bike Sharing Dataset." The dataset provides hourly and daily counts of rental bikes for the years 2011 and 2012 in the Capital Bikeshare system, including relating weather and seasonal data. We select the daily dataset for our project, which has 731 records, with each record depicting the rental scenario for a day.

The dataset is immaculate, having no missing values, and our scrutiny has also assured that all the variables are in their corresponding data types. As a result, there is no data substitution or replacement needed.

There are 16 variables in total that make up the dataset. Out of these, we denote cnt (total daily rental count) as our target variable. Its values lie in a range from a low of 22 to a high of 8,714, with the average daily count being about 4,504. Table 1 presents the summary statistics of cnt.

Table 1 Summary statistics of daily bike rental counts (cnt)

Count	731.000000
Mean	4504.348837
Std	1937.211452
Min	22.000000
25%	3152.000000
50%	4548.000000
75%	5956.000000
Max	8714.000000

2.2 Explanatory Variables

We take out instant from the 15 variables left as it is only an identifier and has no significant influence on the target. The other 14 variables are considered as possible factors affecting demand and can be divided into 3 groups:

1. Weather & Environmental Variables (5 variables)

• weathersit – weather status (1 = Clear, 2 = Mist/Cloudy, 3 = Light Snow/Rain, 4 = Heavy Rain/Snow/Fog)

• temp – normalized temperature (0–1)

• atemp – normalized "feels-like" temperature (0–1)

• hum – normalized humidity (0–1)

• windspeed – normalized wind speed (0–1)

It is important to note that temp, atemp, hum, and windspeed are normalized and stored as float values ranging from 0 to 1.

The reason we get rid of atemp is that it is very much correlated with temp. We want to avoid the multicollinearity issue and at the same time keep temp as a more accurate representative of the actual temperature conditions.

Table 2 Correlation between temp and atemp

	temp	**atemp**
temp	1.000000	0.991702
atemp	0.991702	1.000000

2. Time-related Variables (7 variables)

• yr – year (0 = 2011, 1 = 2012)

• mnth – month of the year (1–12)

• dteday – date in YYYY-MM-DD format

• season – season (1 = Spring, 2 = Summer, 3 = Fall, 4 = Winter)

• holiday – whether the day is a legal holiday (0 = No, 1 = Yes)

• workingday – whether the day is a working day (0 = No, 1 = Yes)

• weekday – day of the week (0 = Sunday, 1 = Monday, …, 6 = Saturday)

3. User-Type Variables (2 variables)

• casual – number of casual user rentals

• registered – number of registered user rentals

2.3 Exploratory Analysis

Descriptive statistics are key metrics to understand the central tendencies of the data set. Figure 1 summarizes average daily rentals by working versus non-working day, by season, by holiday versus non-holiday, and by weather situation using four subplots, providing an overview of how demand varies under different calendar and weather conditions.

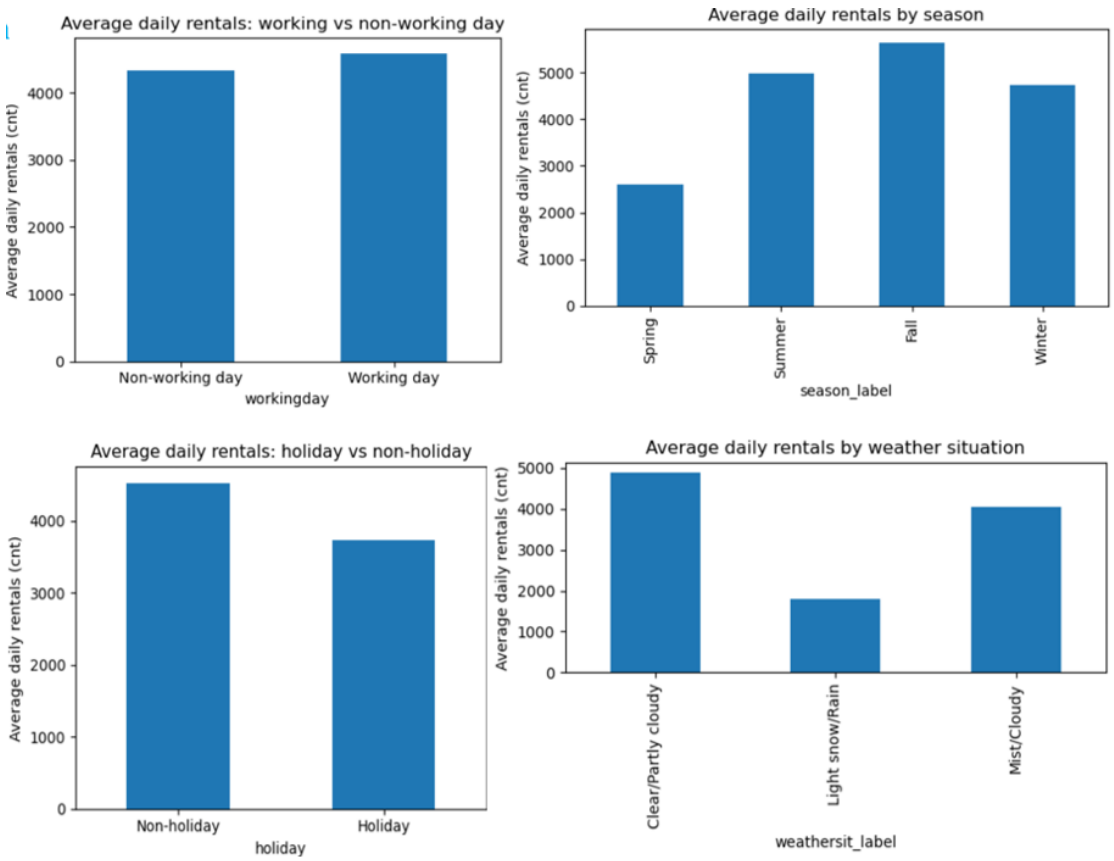


Figure 1 Average daily bike rentals under different calendar and weather conditions

Based on Figure 1, we observe that:

• Rental demand is significantly higher in summer and autumn compared with winter.

• Clear or partly cloudy weather leads to much higher rentals, while rainy or snowy conditions noticeably suppress usage.

• In contrast, the differences between working days and weekends, as well as between holidays and non-holidays, are relatively small.

These findings suggest that seasonal factors and weather & environmental conditions have a much stronger influence on rental demand than calendar-related variables such as holiday status or working days. Therefore, in the subsequent

modeling stage we prioritize season and weather-related features as key factors in explaining and predicting bike rental demand.

2.4 Train–Test Split

After completing the data preprocessing steps, we split the dataset into two parts following the chronological order:

• Training set: 80% of the data (584 days), used to build and fit the models.

• Test set: the remaining 20% (147 days), reserved to evaluate model performance and assist in selecting the most appropriate model.

This time-based split mimics a real forecasting scenario in which models are trained on past data and then applied to predict future observations.

3. Baseline Multiple Linear Regression and Time-Series Models

In this section we build the first predictive models for daily bike rental demand and compare traditional multiple linear regression with univariate time-series approaches. The goal is to understand whether using calendar and weather covariates, or instead relying purely on historical demand patterns, can better capture and forecast daily rentals.

3.1 Baseline Multiple Linear Regression

3.1.1 Variable preparation and correlation analysis

All categorical variables (season, month, weekday, weather situation, and working-day status) are transformed into dummy variables so that they can be included in a linear regression model. For example, season is expanded into four indicators representing Spring, Summer, Fall, and Winter, with one category omitted as the reference group.

To obtain a first understanding of the relationships in the data, we compute pairwise correlations between cnt and all potential predictors. The results are consistent with our descriptive analysis in the previous section:

• Registered users, casual users, and temperature display the strongest positive correlations with total daily rentals.

• Calendar factors such as the year indicator (2012 vs. 2011) and season also show substantial positive associations, capturing the strong upward trend and seasonal pattern of bike usage.

• On the other hand, the indicators for holidays and working days are just slightly correlated with the demand, which is an indication that those riders are using the bicycles a lot during both workdays and weekends, for both commuting and leisure activities.

Such results lead us to the first model specification: we give precedence to variables that are easily understood in a business context and have a strong correlation with demand, while eliminating the features that are not necessary.

3.1.2 Model specification

The baseline model is a standard multiple linear regression of the form

$cnt_t = \beta0 + \beta1 * yr_t + \beta2 * temp_t + \beta3 * hum_t + \beta4 * windspeed_t$ + (season dummies) + (month dummies) + (weather dummies) + (holiday, workingday, weekday dummies) + ε_t.

Key predictors in this baseline model therefore include:

• Year indicator (2012 vs. 2011)

• Seasonal dummies and selected month dummies

• Normalized temperature (temp), humidity (hum), and wind speed (windspeed);

• Weather-situation dummies;

• Working-day and holiday dummies;

• Numbers of registered and casual users, in specifications where they are included as explanatory variables.

All coefficients are estimated by Ordinary Least Squares (OLS) using the training data.

3.1.3 Estimation results and interpretation

Overall, the baseline regression confirms our business intuition. Temperature has a strong positive and statistically significant coefficient, meaning that warmer days are associated with higher rental demand. The indicator for 2012 shows a positive coefficient relative to 2011, reflecting the general growth of the bike-sharing program. The Fall season has the largest positive seasonal effect, while Winter shows the lowest demand, consistent with earlier descriptive plots.

Weather-condition dummies indicate that clear or slightly cloudy days significantly increase rentals, while rainy or

snowy days reduce demand. Coefficients on holiday and working-day indicators remain relatively small, suggesting that usage is spread fairly evenly across the week.

These estimated effects provide interpretable and actionable insights for managers. For example, the strong positive impact of temperature and clear weather highlights the opportunity to push marketing campaigns and promotion activities during warm seasons, whereas on very poor-weather days many bikes may remain unused.

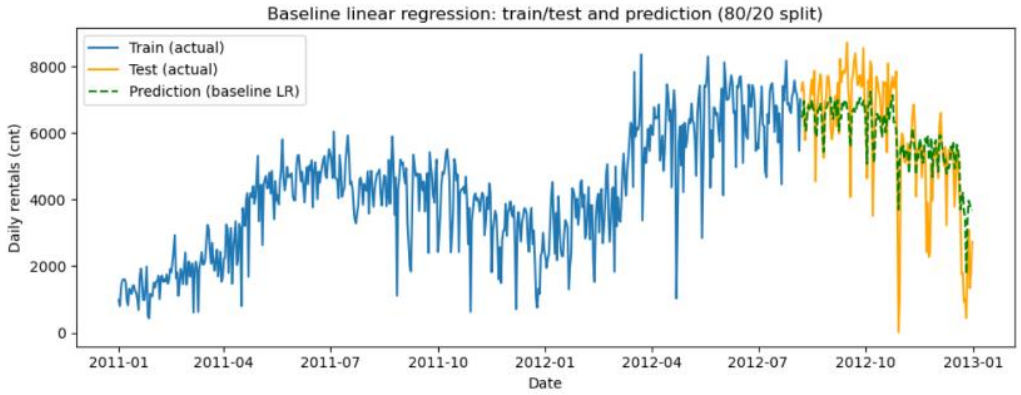


Figure 2 Baseline linear regression: train/test split and prediction versus actual daily rentals

In practical terms, the baseline model can explain about 65% of the variation in daily rentals (training $R^2 \approx 0.65$). On the common test set, its RMSE is approximately 1,103.67 bikes per day, meaning that the typical forecast error is a little above 1,100 bikes. When we overlay the fitted values on the actual demand curve, the regression tracks the overall trend and seasonal swings reasonably well but still misses some short-term peaks and troughs, particularly during special days or abrupt weather changes.

We also check the residuals over time and find no obvious remaining long-term trend; errors appear centered around zero with slightly higher variance at demand peaks. This is acceptable for a first-stage model and justifies using it as the baseline against which we compare more advanced methods in later sections.

3.2 Time-Series Forecasting Models

3.2.1 Rationale for univariate time-series models

While regression explicitly uses weather and calendar variables, another natural approach in business forecasting is to rely purely on past demand values. If daily rentals display stable patterns over time—such as trend or regular seasonality—then a univariate time-series model might achieve comparable or even better accuracy with less data

requirement on exogenous covariates.

To test this idea, we develop two standard exponential-smoothing models:

• Simple Exponential Smoothing (SES) – captures a time-varying average level but ignores trend and seasonality.

• Holt–Winters seasonal model – extends SES by including both trend and seasonal components.

All models are estimated on the same training period as the regression and evaluated on the same time-ordered test period, ensuring a fair comparison.

3.2.2 Simple Exponential Smoothing

In the SES model, the forecast for each day is a weighted average of past observations, with more weight placed on recent days. There are no explanatory variables; the model assumes that the underlying demand level evolves smoothly over time.

We estimate the smoothing parameter by minimizing the sum of squared one-step-ahead forecast errors on the training set. The resulting SES forecasts for the test period follow the general increasing trend but react slowly to sudden changes such as sharp weather shifts or holiday spikes. Consequently, SES under-predicts demand on the highest-usage days and over-predicts during low-usage periods.

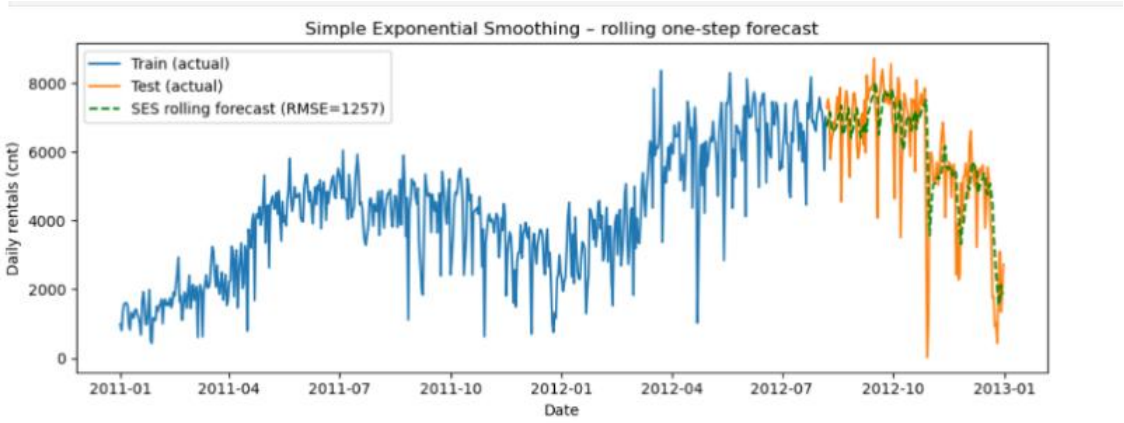


Figure 3 Simple Exponential Smoothing: rolling one-step forecasts versus actual daily rentals

Although exact numerical metrics are reported in Table 3, the key finding is that SES yields a higher RMSE than the baseline regression, indicating inferior predictive performance.

3.2.3 Holt–Winters seasonal model and rolling-pattern analysis

Given the clear yearly seasonality in bike usage, we further fit a multiplicative Holt–Winters model with level, trend,

and weekly seasonal components. The idea is to exploit any regular day-of-week pattern in daily rentals (for example, weekend peaks).

Before estimating this model, we perform an additional rolling-mean visualization. Specifically, we compute a 7-day rolling average of daily rentals and plot both the raw series and the smoothed series.

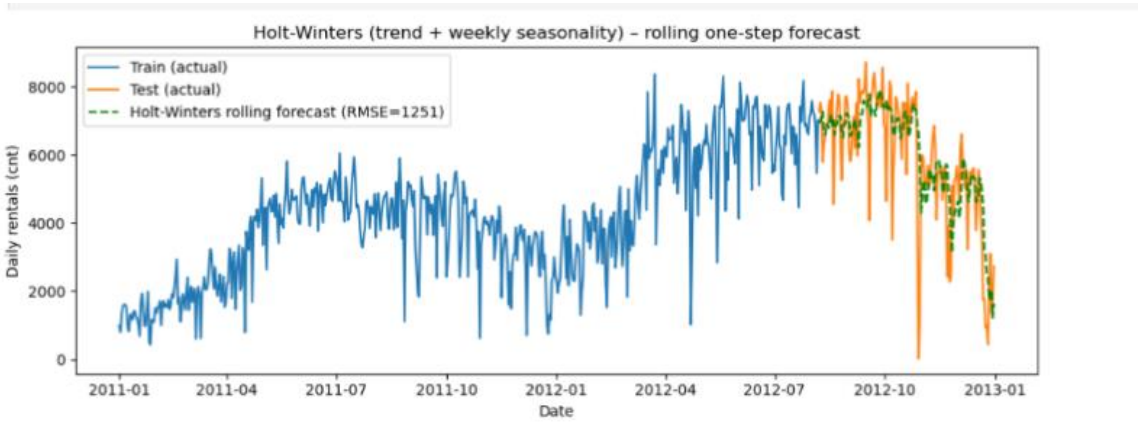


Figure 4 Holt–Winters seasonal model: rolling one-step forecasts and 7-day rolling mean

The figure shows a strong long-term seasonal cycle over the two years, with clear summer peaks and winter troughs. However, the weekly pattern is relatively weak and unstable, especially compared with the stronger seasonal and weather-driven fluctuations.

Empirically, the Holt–Winters model captures the overall trend and some medium-term fluctuations, but its test-set RMSE remains higher than that of the regression model. Forecasts tend to be too smooth, failing to respond adequately to sudden changes in demand. This is consistent with the visual evidence from the rolling-mean plot and model diagnostics.

3.3 Quantitative comparison

Using RMSE on the common 20% test set as the main metric, we summarize the performance ranking in Table 2.

Table 3 Test RMSE of baseline regression and time-series models

Model	**Baseline LR**	**SES**	**Holt-Winters**
RMSE	1103.666864294584	1257.2932408575698	1250.789887949065

The baseline linear regression achieves the lowest RMSE (about 1,103.67), while both SES and Holt–Winters produce larger errors (around 1,257.29 and 1,250.79 respectively).

In summary:

• Multiple linear regression performs best (lowest RMSE).

• Holt–Winters seasonal model has higher RMSE than regression.

• Simple Exponential Smoothing performs worst among the three.

Thus, none of the univariate time-series models outperform the baseline regression constructed from weather and calendar features. The main reasons are:

• The dataset spans only two years, which is relatively short for learning complex seasonal structures in a time-series framework.

• Daily demand is heavily influenced by exogenous factors such as weather conditions and general growth of the system, which are explicitly modeled in regression but ignored in SES and Holt–Winters.

• The weekly seasonal pattern is weak, so the additional complexity of the Holt–Winters model does not translate into better predictions.

4. Multiple Linear Regression versus Advanced Machine Learning Models

4.1 Research Background and Question

The previous section established baseline predictive models for daily rental demand using multiple linear regression and time-series methods (Simple Exponential Smoothing and Holt–Winters). The results showed that a regression model incorporating weather, seasonal, and calendar features outperformed pure time-series models, while the latter failed to deliver competitive accuracy.

To evaluate whether more sophisticated algorithms can further improve prediction performance on this dataset, we now compare the baseline multiple linear regression with several state-of-the-art machine learning models.

The fundamental inquiry for this segment of the study is:

Is there any particular model among the machine learning algorithms that can beat the baseline multiple linear regression in predicting the daily demand of shared bikes for rental?

4.2 Research Methodology

In order to conduct a valid analysis, the same data and feature setup are employed for every model.

• Feature set: The same preprocessed features as in the baseline regression (atemp taken off, 14 explanatory variables kept) are used without any change.

• Data split: Training and testing are done in the ratio of 80% and 20% respectively based on the chronological order (584 days for training and 147 days for testing).

• Training and test periods:

Training: 1 January 2011 – 6 August 2012

Test: 7 August 2012 – 31 December 2012

• Evaluation metrics: The test set's Root Mean Squared Error (RMSE) served as the main evaluation metric, followed by Mean Absolute Error (MAE) and R^2 on the training set to identify fit and overfitting as supplementary metrics.

In this standard configuration, we compare the models with the baseline OLS regression model and thus test the performance of the four advanced models—Ridge Regression, Random Forest, Gradient Boosting, and XGBoost.

4.3 Ridge Regression

• Model overview: Ridge regression imposes an L2 penalty on the OLS criterion function, resulting in the reduction of coefficients to nearly zero as a way to handle multicollinearity and variance reduction.

• Performance and characteristics: Ridge regression is a little less effective than the baseline OLS with a test RMSE that is roughly 3.24% worse.

Reasons why Ridge is not effective

• In the original dataset, temperature and "feels-like" temperature exhibit very strong correlation (correlation 0.992), however, we have already excluded atemp.

• The rest of the correlations, e.g., temp–windspeed (−0.18) and hum–windspeed (0.13), are weak.

• With 584 training samples and 14 predictors, the sample-to-feature ratio is roughly 43.7, which is more than sufficient for stable OLS estimation.

In this case, the bias brought in by Ridge is not offset by a significant drop in variance.

• Conclusion: When multicollinearity is at a moderate level and the sample size is large enough, a properly specified OLS model can be as good as or better than Ridge regression, while still being simpler and more interpretable.

4.4 Random Forest

• Model overview: Random Forest (RF) is an ensemble of decision trees trained on different bootstrap samples where at each split feature selection is done randomly to model non-linear relationships and interactions.

• Performance: The performance of RF among all models is the worst, with a test RMSE that is approximately 20.33% higher than the standard OLS.

Proof of overfitting

• The RMSE of training is under 950 which is a huge improvement over OLS on the training set.

• The RMSE of test is greater than 1,325 which indicates more than 39% drop in performance from training to testing.

Root causes

• Data scale limitation: The number of training observations is only 584. When using about 300 trees, each tree is constructed from very much overlapping bootstrap samples (each observation is included about 2-3 times per tree), and this allows the trees to easily fit noise in their individual samples.

• Weak interaction signals: RF could in principle capture interactions like “autumn × clear weather,” but these signals are extremely weak in this dataset. The high independence of the trees in case of small data results in the ensemble fitting noise instead of robust patterns.

• Summary: Random Forest is not appropriate for this dataset; its model complexity that cannot be supported by the sample size available leads to and very drastic overfitting.

4.5 Gradient Boosting

• Model overview: Gradient Boosting (GB) creates decision trees one after another in a sequential manner, with every new tree being fit on the residuals of the previous trees. The process is thus aimed at gradually minimizing prediction errors and revealing the complicated patterns.

• Performance: GB's accuracy is higher than that of Random Forest but lower than that of the OLS baseline, as shown by the test RMSE, which is about 7.61% higher than that of OLS.

Overfitting indicators

• Training RMSE is below 1,000, indicating an extremely good in-sample fit.

• Test RMSE is above 1,180, producing a gap of more than 18% between training and testing errors.

Why does Gradient Boosting perform poorly on this dataset?

• Cumulative learning on a tiny dataset: The learning process of the trees occurs in a series where each tree is trained on the residuals of the prior ones. The dataset of 584 samples is small such that the residuals rapidly turn into random noise and so the trees created afterwards start learning noise instead of the signal.

• Relationships predominantly linear: The baseline OLS itself scores R^2=0.653 that indicates simple linear effects accounts for around 65.3% of the variance. The leftover 34.7% is mostly random or caused by unobserved factors, thus there is no much room for GB to expose extra structure.

• Limited gain from tuning: The data limitation is so severe that even with nice tuning such as few rounds (200 iterations) and shallow trees the GB cannot avoid overfitting without sacrificing accuracy.

• Conclusion: For the small dataset with mostly linear relationships, the powerful sequential learning mechanism of Gradient Boosting turns out to be more of a drawback than an advantage.

4.6 XGBoost

• Model overview: XGBoost is an improvement over traditional gradient boosting which introduces the settings of L1/L2 regularization over and above it, second-order (Newton-type) optimization and the use of generic learning rates.

• Performance and characteristics: XGBoost’s test RMSE is in-between Random Forest and Gradient Boosting but still significantly **worse than** OLS (about 10.85% higher RMSE).

There are certain key points that explain why XGBoost does not perform better than OLS

•Regularization constraints for small data: Regularization in XGB and complexity constraints are for large-scale problems. With only 584 training samples, the effective space for tuning parameters such as reg_lambda = 3.0 and

reg_alpha = 0.1 is limited; overfitting remains.

• Second-order optimization not needed here: Second-order Taylor expansion is beneficial for complex nonlinear problems. In this dataset, however, key feature–target relationships are largely linear (e.g., temp: r=0.627 , season r=0.406, weathersit: r=−0.297, windspeed: r=−0.235, mnth: r=0.280). First-order information already captures most of the structure.

• Adaptive learning rates chasing noise: Adaptive learning rates excel in large, complex feature spaces. On small data with mainly linear relationships, adaptivity tends to chase noise rather than genuine patterns.

• Summary: XGBoost's advanced optimization techniques cannot compensate for the fundamental mismatch between model complexity and the limited, mostly linear dataset used here.

4.7 Comprehensive Five-Model Comparison

On the shared test set, we compare the baseline OLS regression with all four advanced models using RMSE as the main metric. Figure 5 summarizes the test RMSE for the five models.

RMSE comparison on the same 20% test period (baseline features):

	Model	RMSE
0	Baseline Linear Regression	1103.666864
1	Ridge Regression	1117.250049
2	Gradient Boosting	1187.547364
3	XGBoost	1223.212778
4	Random Forest	1327.886476

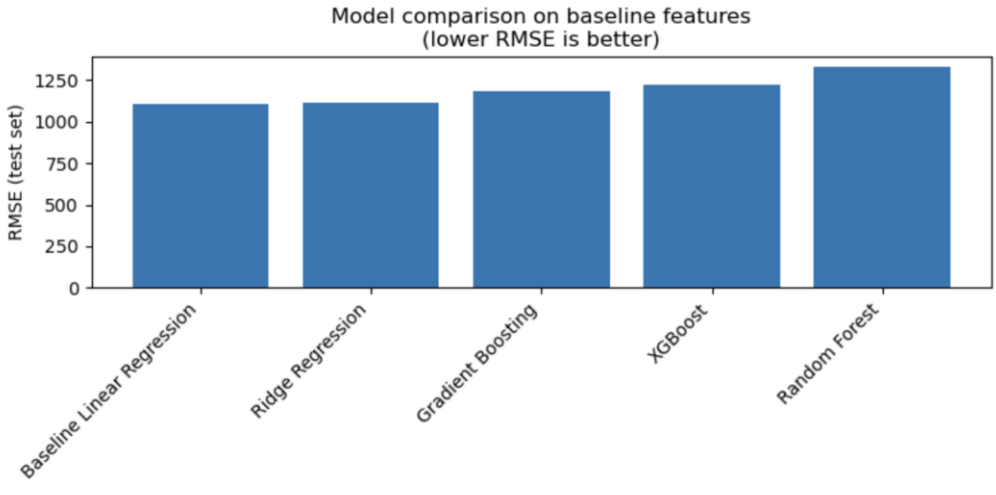


Figure 5 Test RMSE of five prediction models on the common 20% test set

From Figure 5 we observe the following performance ranking (from best to worst):

1. Baseline multiple linear regression (lowest RMSE, ≈ 1,103.67)

2. Ridge Regression (RMSE slightly higher, +3.24%)

3. Gradient Boosting (RMSE +7.61%)

4. XGBoost (RMSE +10.85%)

5. Random Forest (worst, RMSE +20.33%)

Common root causes of complex models' failures

Two structural issues jointly explain why all complex models underperform the simple OLS baseline:

1. Data-scale constraint.

• Training samples: 584 → sample-to-feature ratio ≈ 43.7

• Test samples: 147 → relatively small for robust model selection

• Total samples: 731 → small for tree ensembles and boosting methods that typically require much more data.

For Random Forest (300 trees), each tree is trained on heavily overlapping bootstrap samples; for Gradient Boosting (≈200 iterations) and XGBoost (≈500 iterations), many successive trees are fitted on noisy residuals. In all three cases, model complexity is high relative to the data scale, leading to strong overfitting.

1. Fundamentally linear relationships.

Feature–target correlations indicate predominantly linear patterns:

• temp: r=0.627 → strong positive linearity

• season: r=0.406 → moderate positive linearity

• weathersit: r=−0.297 → linear negative relationship

• windspeed: r=−0.235→ linear negative relationship

• mnth: r=0.280→ weak positive linearity

The baseline OLS model already explains 65.3% of the variance (R^2=0.653). The remaining 34.7% is largely non-predictable random noise or driven by unobserved factors. Complex models attempt to extract "features" from this residual component, which in practice means fitting noise and harming generalization.

Summary of findings from this comparison

• For this dataset, the simple multiple linear regression model provides the best trade-off between accuracy,

interpretability, and robustness.

• Advanced machine learning models—Random Forest, Gradient Boosting, and XGBoost—do not improve performance and instead suffer from pronounced overfitting due to small sample size and largely linear relationships.

• Ridge regression, designed to address multicollinearity, yields only a small deterioration relative to OLS, confirming that multicollinearity is not a serious issue after removing atemp.

These results justify using the baseline OLS regression as the core forecasting model for this project. In the next section, we retain this linear framework but introduce a carefully chosen nonlinear term (temperature squared) to capture the concave effect of temperature on demand and further improve prediction accuracy, while keeping the model structure transparent.

5. Nonlinear Refinement of the Baseline Linear Model

5.1 Motivation and Model Specification

The results in Section 4 show that our baseline multiple linear regression already performs better than both the time-series models and the advanced machine learning models. However, scatter plots between temperature (temp) and total daily rentals (cnt) suggest that their relationship is not perfectly linear.

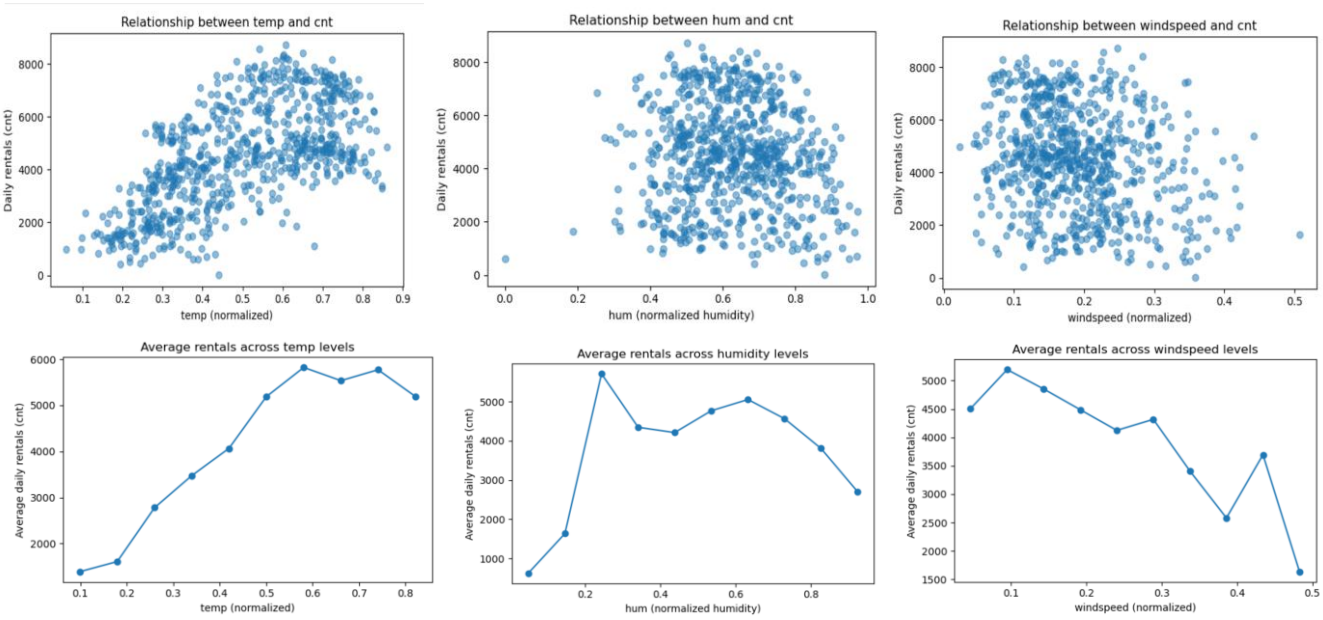


Figure 6 Temperature, humidity and windspeed versus daily rentals: scatter plots and binned averages

When temperature is very low, demand is also low. As temperature becomes warmer, demand increases rapidly. Yet

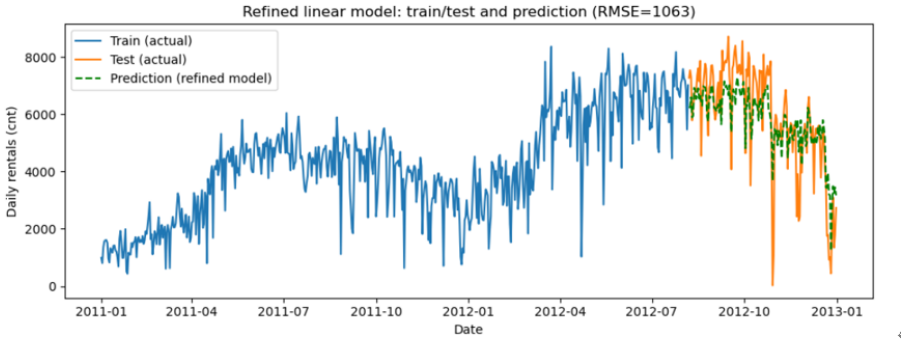
when temperature is extremely high, demand stops increasing at the same speed and may even decline slightly. This pattern looks more like a curved shape than a straight line.

To capture this nonlinear effect in a simple and interpretable way, we extend the baseline regression model by adding a squared temperature term: temp_sq=temp^2

We then re-estimate the regression model including both temp and temp_sq as explanatory variables. The model is still linear in the parameters, but we allow temp_sq to capture a nonlinear temperature effect on bike demand.

5.2 Results of the Refined Model

Figure 7 shows the performance of the refined regression model with the additional squared-temperature term, using the same 80/20 train–test split as before.




```
Refined linear regression with temp^2 (80/20 split):
RMSE: 1063.48312977
MAE : 793.92469945
R^2 : 0.67816398

Top 10 features by absolute coefficient (refined model):
     feature   coefficient
        temp  14500.527471
     temp_sq -11776.587660
   windspeed  -2999.133290
          yr   1943.170834
weathersit_3  -1716.307319
         hum  -1623.560806
    season_4   1166.140545
      mnth_6   1009.228748
      mnth_7    861.323565
      mnth_5    846.902418
```

Figure 7 Refined linear regression with squared temperature term

The fitted values track the overall pattern of demand very closely on both the training and test periods.

On the common test set, the RMSE decreases from about 1,103.67 for the baseline model to about 1,063.48 for the refined model, a reduction of roughly 3.6%. MAE also improves to around 793.92, and the training R^2 increases from

approximately 0.653 to about 0.678, indicating that the refined model explains more variation in daily rentals while remaining easy to interpret.

The coefficient signs on the temperature terms match our expectations:

• The coefficient on temp is positive; The coefficient on temp_sq is negative.

This means that higher temperature increases bike rentals when the day is cold or mild, but the marginal effect of additional warming becomes smaller when the day is already warm, and extremely hot days may even reduce demand. Coefficients on the other variables keep similar signs and magnitudes as in the baseline model, confirming that the refinement mainly adjusts the shape of the temperature effect rather than changing the overall structure of the model. Overall, the refined regression with temp and temp_sq provides a better fit and lower prediction error while preserving the transparency and interpretability of a linear model. We therefore adopt this refined regression as our final forecasting model.

6. Forecasting Results with the Final Model

6.1 Short-Term Forecast: 30-Day Horizon

For the short-term forecast, we focus on the 30 days immediately after the end of our test period. We construct the future input data by extending the calendar variables (date, year, month, weekday, season, working day, and holiday) and by assigning reasonable patterns for the weather-related variables based on the recent historical period. We then apply the final regression model to predict cnt for each of these 30 future days.

Figure 8 presents the resulting 30-day forecast together with the actual demand in the test period.

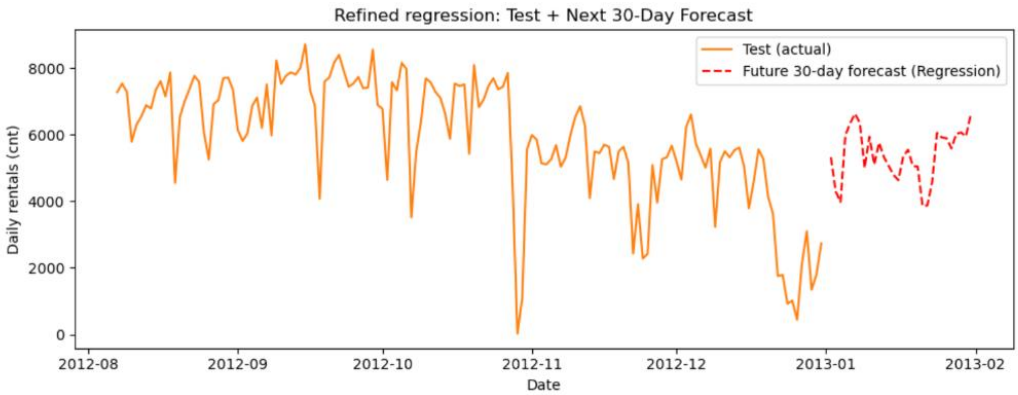


Figure 8 Thirty-day ahead forecast of daily bike rentals using the refined regression model

The short-term forecast indicates that daily demand remains relatively stable, mostly within the range of 5,000–7,000 rentals per day, with no dramatic spikes or collapses. The expected demand is a little less than the highest summer levels reported in the historical data, which is in line with the late-autumn and early-winter periods included in this forecast window. These 30-day forecasts can be used for operational planning like weekly fleet allocation, daily redistribution schedules, and staff rostering.

6.2 Long-Term Forecast: One-Year Horizon

In order to look more in-depth into longer-term trends and seasonality, we produce a forecast for one year ahead. To be precise, we take the data and add another complete year (365 days) to it beyond the original sample. The year variable is changed, and the temp_sq and other derived variables are recalculated. This method presumes that the weather and seasonal pattern in the following year will be the same as the historical one, but an upward trend is allowed to be captured by the year dummy. The very first forecast from a long perspective is displayed in Figure 9 alongside the recent historical path.

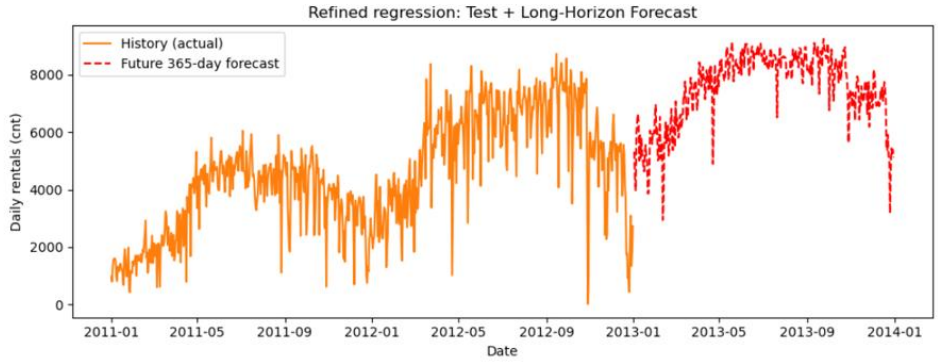


Figure 9 One-year ahead forecast of daily bike rentals using the refined regression model

The forecasted curve indicates a strong seasonal cycle which corresponds to our previous observation:

• Winter is still the period of low demand; Spring is the time of demand recovery; Late summer and fall are the main demand peak; Early winter shows a slow decline again.

To present the seasonal pattern more transparently, Table 3 illustrates the average daily rentals in 2012 alongside the model's predictions for 2013, broken down by season.

Table 4 Seasonal comparison of actual 2012 rentals and model forecasts for 2013

Season	Actual_2012	Forecast_2013	Change_abs	Change_pct
Spring	3998.6	6010.0	2011.4	54.5
Summer	6295.6	8226.8	1931.2	30.7
Autumn	6924.4	8475.6	1551.2	22.6
Winter	5164.6	7246.4	2081.8	43.3

The forecasts indicate substantial growth in all four seasons, with the largest relative increases in spring and winter, suggesting that overall demand is expected to continue rising even after controlling for weather and season.

During the months with the highest demand, the model predicts that more than 8,000 bikes will be rented out daily, whereas in winter the daily demand will be significantly lower. These one-year forecasts provide useful information for long-term capacity planning, investment decisions, and strategic marketing design.

At the same time, the long-term forecasts should be interpreted with caution: they rely on the assumption that future weather and seasonal patterns resemble historical ones and do not account for major structural changes such as policy shifts or large-scale infrastructure upgrades.

7. Business Implications

The refined regression model and forecasts provide several actionable implications for a bike-sharing operator.

7.1 Seasonal Distribution and Capacity Planning

Demand shows a clear yearly cycle. Summer and early autumn are peak seasons, while winter is the trough. The operator should therefore adjust the active fleet size instead of keeping a constant number of bikes:

• Increase available bikes and maintenance capacity in summer and early autumn, including temporary parking space and more redistribution staff in popular areas.

• Reduce the number of active bikes in winter and use the slack period for deep maintenance, refurbishment, or storage to lower operating costs.

Aligning capacity with seasonal demand enables a better trade-off between service quality and cost efficiency.

7.2 Weather-Based Operational Scheduling

Temperature, weather situation, humidity, and wind speed all significantly affect daily rentals. Short-term operations should follow weather forecasts:

• On good-weather days (clear or partly cloudy, comfortable temperature), increase rebalancing frequency and monitoring of high-demand areas.

• On bad-weather days (rain, snow, high humidity, strong wind), scale down field operations and schedule non-urgent maintenance, saving labour and transport costs.

Overall, weather-based operational planning helps the company use resources more efficiently and avoid both bike shortages and oversupply on any given day.

7.3 Marketing and Customer Management

Based on the upward demand trend and significant seasonal variations, it will be possible to synchronize marketing activities with the prediction patterns. The operator will be able to:

• Unveil seasonal sales at spring beginning and summer early, so that usage of the first season would be stimulated.

• Use targeted discounts or loyalty rewards on extremely hot days when natural demand is depressed.

• Focus campaigns on converting casual riders into registered users, making demand more stable and predictable over time.

8. Limitations and Future Work

Our analysis has several limitations that define useful directions for future work.

First, the dataset covers only two years of daily observations and is aggregated at the system level. With a longer panel and station-level data, we could study long-run structural changes and spatial imbalances between different parts of the city.

Second, important business variables are missing. The data contain no information on prices, promotions, competitor actions, or major policy changes. These unobserved factors may explain part of the remaining variation in rentals and

could be incorporated in future models if such data become available.

Third, hyperparameter tuning for advanced machine learning models is intentionally simple. More systematic tuning (for example, grid search or Bayesian optimisation) might modestly improve their performance, although the small sample size and largely linear relationships suggest that our main conclusion in favour of a refined linear model is unlikely to change.

Finally, our long-term forecasts assume that future weather and seasonal patterns resemble the historical period and that market conditions remain broadly similar. If there are major policy shifts or new competitors, actual demand may deviate from the projected path.

Future work could therefore combine longer and richer datasets with more flexible spatio-temporal models to provide finer-grained recommendations for specific stations and user segments.

9. Conclusion

This project analyses daily demand in a bike-sharing system to identify key drivers and to select an appropriate forecasting model. Using two years of clean daily data with weather, calendar, and user-type variables, we first documented strong seasonality and clear weather effects: summer and autumn show much higher usage than winter, and clear, warm days substantially increase rentals.

We then built a baseline multiple linear regression model and compared it with two standard time-series models (Simple Exponential Smoothing and Holt–Winters) and four advanced machine learning models (Ridge Regression, Random Forest, Gradient Boosting, and XGBoost). On a common 20% test set, the linear regression clearly outperforms all alternatives. Complex models tend to overfit the small dataset and do not deliver better forecasts.

Guided by the observed curvature in the relationship between temperature and demand, we refined the regression by adding a squared temperature term. This straightforward supplement enhances RMSE and explanatory power at the same time keeping interpretability intact. The ultimate model accommodates a forecast period of either 30 days or a year ahead and also uncovers a constant seasonal cycle along with an overall increasing demand trend.

In general, our research demonstrates that an appropriately specified and not much changed linear regression model can yield predictions for bike-sharing operations that are accurate, strong and clear. It is very important to match the model complexity with the availability of data and the requirements of the business: here, a not very complicated model in accordance with the knowledge of the sector is more beneficial for management decision making than the application of more advanced methods.

References & Appendix：

[1] S. Cai, X. Long, L. Li, H. Liang, Q. Wang, and X. Ding, 'Determinants of intention and behavior of low carbon commuting through bicycle-sharing in China', *Journal of Cleaner Production*, vol. 212, pp. 602–609, Mar. 2019, doi: 10.1016/j.jclepro.2018.12.072.

[2] C. Bullock, F. Brereton, and S. Bailey, 'The economic contribution of public bike-share to the sustainability and efficient functioning of cities', *Sustainable Cities and Society*, vol. 28, pp. 76–87, Jan. 2017, doi: 10.1016/j.scs.2016.08.024.

[3] M. Dell'Amico, E. Hadjicostantinou, M. Iori, and S. Novellani, 'The bike sharing rebalancing problem: Mathematical formulations and benchmark instances', *Omega*, vol. 45, pp. 7–19, Jun. 2014, doi: 10.1016/j.omega.2013.12.001.

[4] D. Yu and L. Shang, 'Opportunities and Challenges Faced by Share Economy: Taking Sharing Bicycle as an Example', *DEStech Transactions on Economics, Business and Management*, vol. 0, no. icmed, 2017, doi: 10.12783/dtem/icmed2017/19328.

[5] J. C. Westland, J. Mou, and D. Yin, 'Demand cycles and market segmentation in bicycle sharing', *Information Processing & Management*, vol. 56, no. 4, pp. 1592–1604, Jul. 2019, doi: 10.1016/j.ipm.2018.09.006.

[6] C. Ma and T. Liu, 'Demand forecasting of shared bicycles based on combined deep learning models', *Physica A: Statistical Mechanics and its Applications*, vol. 635, p. 129492, Feb. 2024, doi: 10.1016/j.physa.2023.129492.