Data 583 Project Report

Seoul Bike Sharing Demand Analysis

Yahan Cong, Jade Yu

March 2024

Contents

1 Introduction

The Seoul bike-sharing system, "Ddareungi," was launched by the Seoul Metropolitan Government in collaboration with private companies. It aims to facilitate commuting for citizens, enhance their quality of life, and alleviate congestion and air pollution in Seoul's urban transport system. "Ddareungi" stands out for its extensive accessibility, user convenience, and commitment to energy conservation, making it a vital component of the city's short-distance transportation infrastructure. As of February 2016, the total ridership of Seoul's bike sharing has reached approximately 14.9 million. For operators of bike-sharing services, maintaining a stable supply and strategically managing bicycle resources pose significant challenges. This study aims to explore the impact of weather conditions and temporal features such as holidays and weekdays on the bike rental system. Through training machine learning models, we aim to understand which factors influence the operational status of the Seoul public bike system and how the distribution of bike rental counts is characterized. Ultimately, we aspire to build a model capable of predicting the hourly rental counts of Seoul's public bikes.

2 Data Import

Our dataset includes weather conditions, date information, operation status ("Functioning Day"), and hourly bike rental counts through the Seoul bike-sharing system over 365 days, from December 2017 to November 2018, with no missing value. Our target variables are "Functioning Day" and "Rented Bike Count". The functioning day represents whether the Seoul public bike is operational on a given day. And the "Rented Bike Count" is its hourly bike rental amount. During data importing, we segmented the dates into months and weekdays and introduced a new date parameter "WorkdayOrNot", which specifies whether each day is a workday or weekend, to investigate the temporal effect on both the operational status of Seoul's public bike system and the dynamics of bike rentals.

Table 1: Parameters of Seoul public bike

Variable Name	Type	Description	Units
Date	Date	The dates on which the shared bikes were rented.	
Month	Date	The month in which the shared bikes were rented.	
Weekday	Date	The weekday in which the shared bikes were rented.	
WorkdayOrNot	binary	Whether the day is a workday or weekend	
Hour	Integer	The hour on which the shared bikes were rented.	
Seasons	Categorical	Spring, Summer, Autumn, Winter	
Holiday	Binary	Whether the rental day is a holiday or not	
Temperature	Float	Temperature in Celsius for bike rental hour	℃
Humidity	Integer	Humidity for bike rental hour	%

Table 1 – Continued

Variable Name	Type	Description	Units
Wind speed	Float	Wind speed for bike rental hour	m/s
Visibility	Integer	Distance at which objects are clearly visible under atmospheric conditions.	10m
Dew point temperature	Float	The Celsius temperature at which air becomes saturated with water vapour and begins to condense.	℃
Solar Radiation	Float	Solar radiation amount	Mj/m^2
Rainfall	Integer	Rainfall amount	mm
Snowfall	Integer	The snowfall amount	cm
Functioning Day	Binary	Whether Seoul bike sharing system is working or not	
Rented Bike Count	Integer	Count of bikes rented at each hour	

3 Exploratory Data Analysis

Following the import and initial processing of the data, we aim to utilize Exploratory Data Analysis (EDA) to explore data distribution, and the relationship between Seoul's public bike operational status, hourly bike rented amount distribution, and comprehensive temporal factors, including but not limited to holiday, workdays, and seasonal variations.

3.1 Independent Variables Distribution

At the start of EDA, we want to know the distribution characteristics of our features. The box plots presented in Figure **??** below illustrate the distribution characteristics of numerical independent variables. Most features exhibit asymmetrical distributions and skewness, indicating a deviation from normal distribution. And wind speed, solar radiation, rainfall, and snowfall possess a substantial number of outliers. Therefore, after completing the exploration data analysis, we will implement data standardization to improve the data distribution, thereby enhancing the accuracy and stability of machine-learning models.

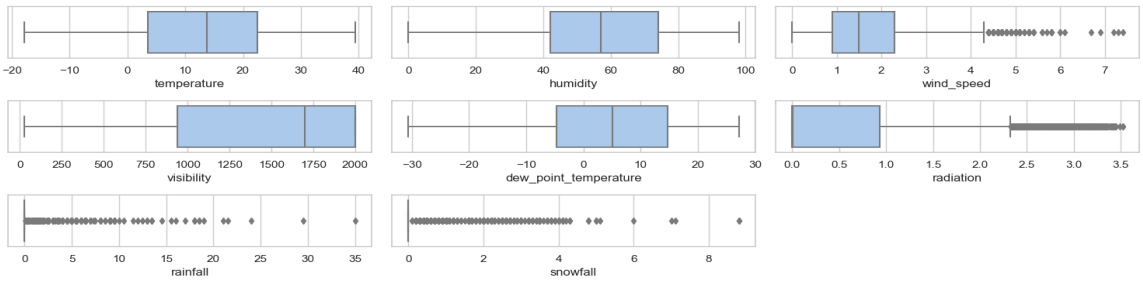


Figure 1: Distribution Characteristics of Numerical Features

3.2 Target Variables

Regarding our two target variables: the operation status (functioning day) and the rented bike count, it can be observed from Figure 2 that the number of non-functioning days is significantly lower than that of functioning days, constituting only a small fraction of the total data. Further distribution characteristics will be analyzed in the following "Seoul's Public Bike Operation Status with Temporal Factors". As for the rented bike count, which serves as our target for regression prediction, we aim to explore whether it follows a specific distribution.

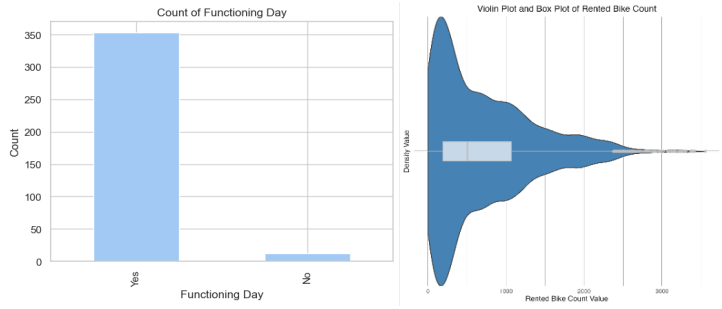


Figure 2: Rented Bike Count Distribution

3.2.1 Distribution Test of Rented Bike Count

The density plot indicates that the distribution of rented bike count is right-skewed with many outliers, potentially following a Poisson, gamma, or log-normal distribution. To verify this, we conducted Kolmogorov-Smirnov tests. Unfortunately, the p-value in the best scenario remains around 0.01, which suggests there is insufficient evidence that the rented bike count conforms to any of the specific distributions mentioned above. This implies that rented bike counts do not adhere to any particular distribution, and therefore, an unparameterized model may likely perform better in predictions.

3.2.2 time series of rented bike distribution

The rented bike counts in the dataset do not follow any specific distribution. However, given that they represent the number of bikes rented each hour, we hypothesized that there might be a temporal cyclicity. To explore this, we performed an STL seasonal decomposition, and as shown in Figure 3, there is a strong daily cycle in the rented bike data. This indicates that employing methods that analyze time series could be particularly effective when predicting these counts.

3.3 Seoul's public bike operation status with temporal factors.

In this section, we seek to explore the operational status distribution of Seoul's public bike system. Over one year, from December 2017 to November 2018, the Seoul public bike was operational for 353 days and closed for 12 days. Therefore, we focus on the specific distribution of these relatively rare non-functioning days. As shown in Figure 4, these non-operational days predominantly occur in autumn and spring, with a significant majority in autumn. The vast majority of these closure

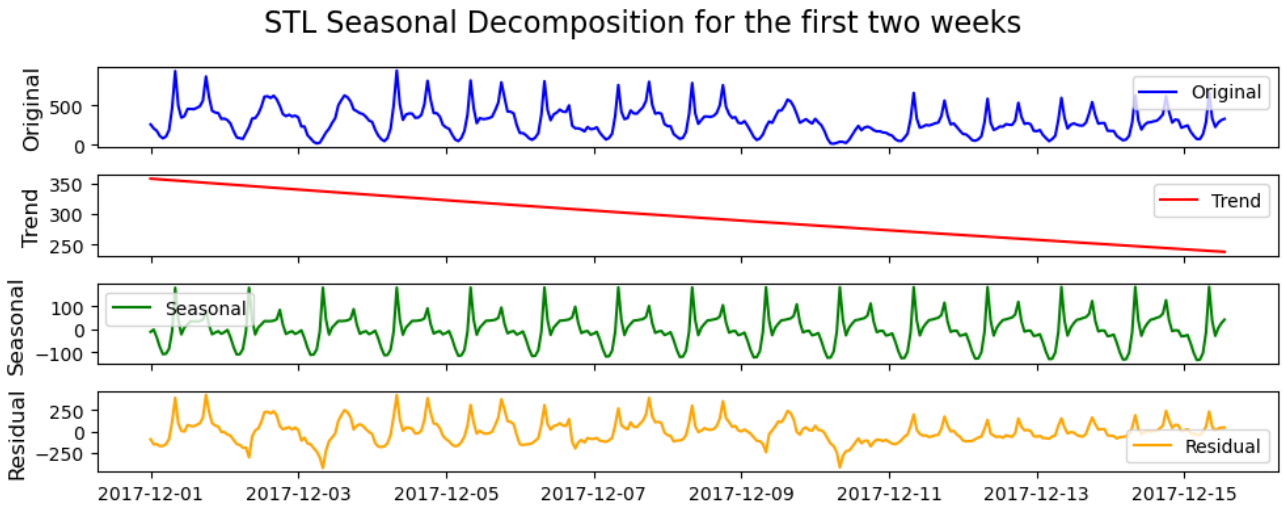


Figure 3: Rented Bike Count Distribution

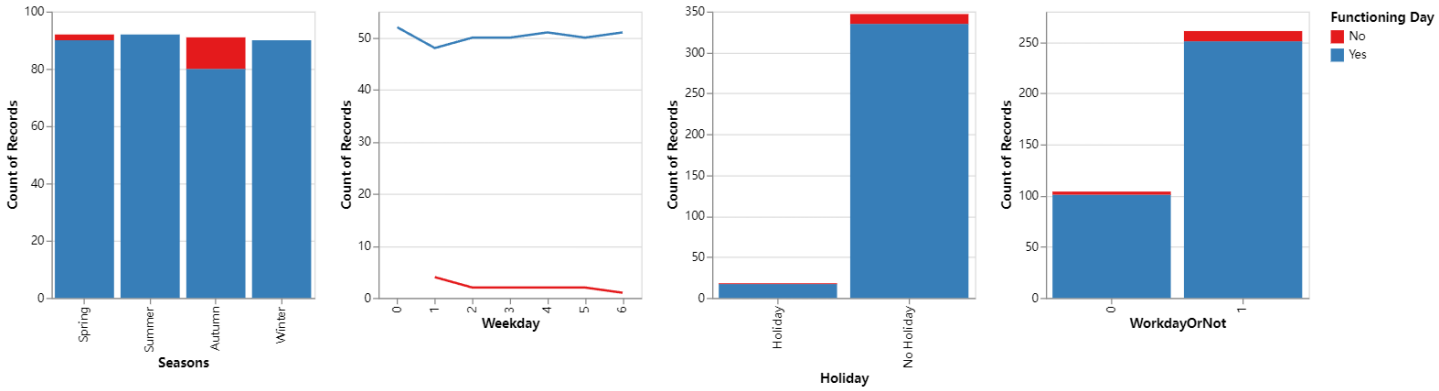


Figure 4: Seoul Bike Sharing System's Functioning Day Distribution

days are non-holidays or workdays. It is observed that, compared to workdays, there are only two closures on weekend, and only one closure during the holiday period. The concentration of Seoul public bike closures during typical working days/ non-holidays indirectly suggests that holiday rest periods likely have a minimal impact on decisions regarding the bike-sharing system's closures. The operational status of the shared bikes is more likely closely related to changes in Seoul's weather conditions.

3.4 Rented Bike Distribution with Temporal Factors

In this section, in addition to the impact of holidays and workdays, we also want to explore the impact of other time elements, such as hours, weekdays, months and seasons on hourly rental numbers of Seoul public bicycles. Through this analysis, we want to investigate whether the rental numbers exhibit a clear time series trend, thereby providing a basis for more accurate predictions and understanding of public bicycle usage patterns.

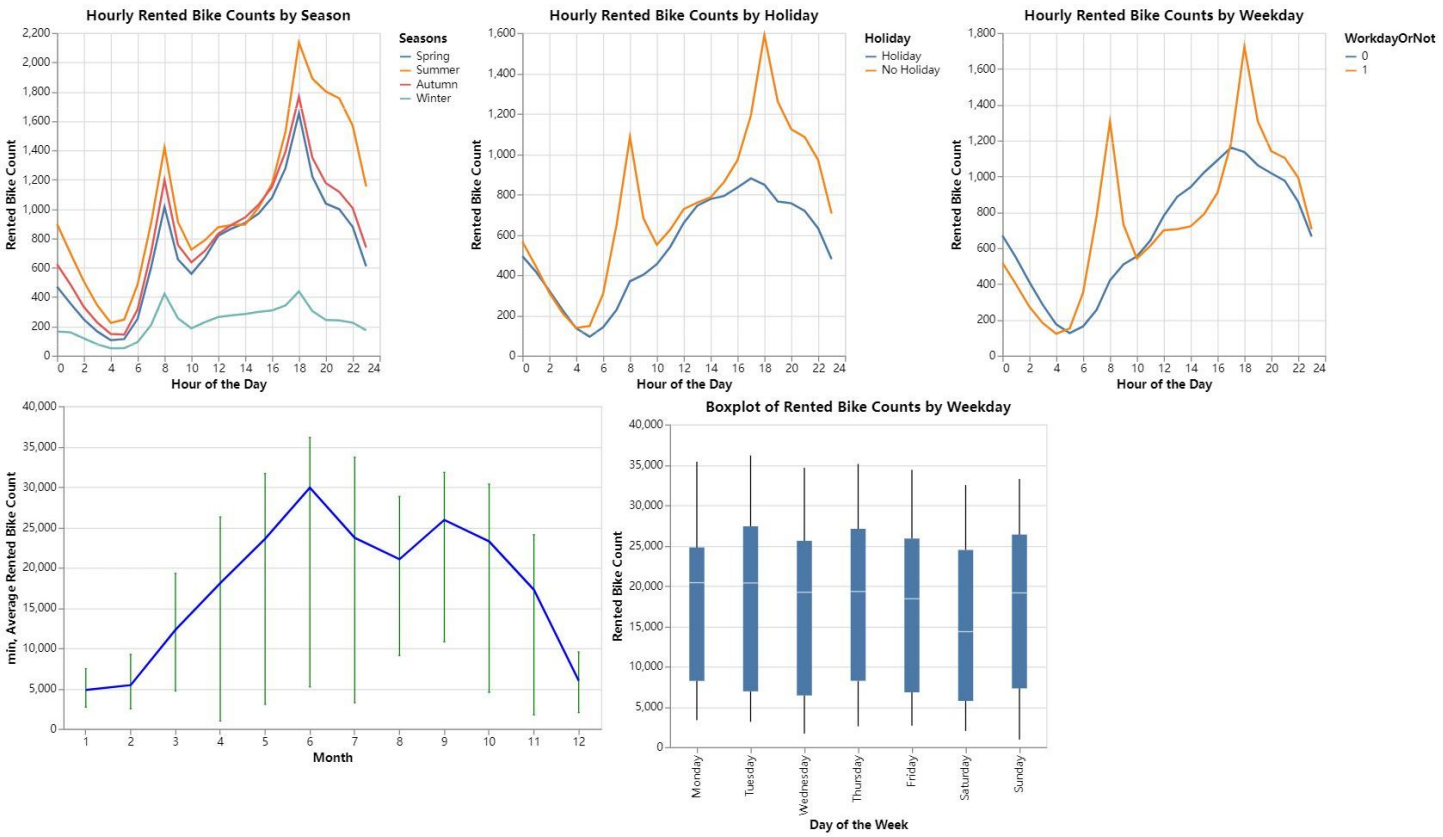


Figure 5: Distribution of Number of Rented Bikes

Figure 5 illustrates the distribution of average daily bike counts across weekdays, holidays, and seasons. It is evident from the figure that the usage of bikes reaches its peak during the summer months of May to July and hits its lowest in winter, with a generally consistent distribution across all seasons, marking two peak usage periods: 6-8 AM and 4-8 PM. The variation between peak values is more significant during the summer, while in winter, due to the overall lower demand for shared bikes, this bimodal trend appears more subdued. This may indicate that the variation in bike counts throughout the year is closely linked to commuting needs during morning and evening peak hours, as well as to seasonal temperature changes and other factors affecting bike rental numbers.

Furthermore, the average demand distribution for shared bikes also changes on holidays and non-working days. On holidays or weekends, the number of rented bikes shows a different unimodal trend compared to weekdays, with a steady increase from 6 AM, peaking around 5 PM before it begins to decline.

Additionally, it can be observed that from 10 AM to 4 PM on non-working days, there is a slight increase in the number of rented bikes compared to working days. This implies that whether it is a working day or not is a significant factor affecting bike rental numbers and distribution.

4 Machine Learning of Seoul public bike operation status

Building on the initial exploration, we now understand when Seoul public bike system experiences non-operational days. Yet, pinpointing precise conditions that halt operations proves challenging. Consequently, our next step involves utilizing classification techniques to discern the system's operational status, aiming to highlight which weather conditions are most influential in predicting the cessation of bike-sharing services.

4.1 Data Pre-processing

Given that our independent variables deviate from a normal distribution and contain numerous outliers, coupled with the fact that many machine learning models are sensitive to the range of data, we apply standardization to our independent variables. Additionally, to address the imbalance between functioning and non-functioning day categories in the original dataset—where non-functioning days are significantly less frequent than functioning days—we employ the Synthetic Minority Over-sampling Technique (SMOTE) to balance these groups. Finally, we split our dataset into training and testing sets using an 80-20 split.

4.2 Methodology

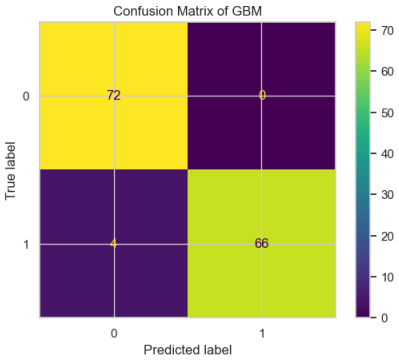
During operation status prediction, since our dataset spans only one year, we opted for gradient boosting and random forest as our classification methods because of their good performances on small datasets.

Gradient boosting is a kind of boosting tree. It begins with an initial decision tree for the first classification. Subsequently, it iteratively trains additional decision trees focusing on its misclassifications. By aggregating the predictions of these trees, the gradient-boosting model incrementally lowers its log loss and enhances its overall accuracy. After optimizing its hyperparameters with grid search, we obtained a gradient boosting model based on 100 trees with a 0.2 learning rate and a maximum tree depth of 4 for each tree.

Similar to gradient boosting, random forest is also a collection of decision trees. However, instead of continuously adding decision trees to enhance predictive accuracy, Random Forest leans towards utilizing "collective wisdom" for its predictions. It randomly divides the training data and trains decision trees on each subset and uses the majority vote of these trees as the final prediction result. We also optimized its hyperparameters using grid search. Our random forest model is composed of 250 trees with a minimum sample per leaf of 1 and a minimum sample split of 2.

4.3 Results

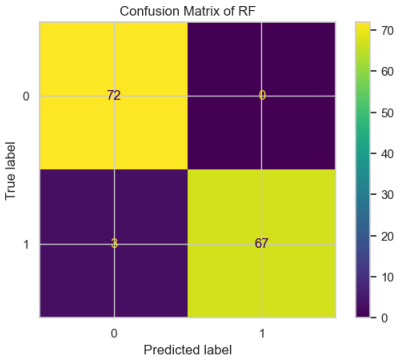
As shown in Figure 6 and Figure 7, both of our models exhibit strong performance in predicting the operation status of Seoul public bikes. The Gradient Boosting model achieves an accuracy of 0.9718 on the testing set, while the Random Forest model reaches an accuracy of 0.9789. Remarkably, both models accurately predict 100% of the non-functioning days, with the prediction accuracy for functioning days also achieving impressive rates (recall) of 0.94 and 0.96, respectively. Our models have also ranked the importance of features based on Gini Impurity for each parameter. From the bar plots associated



Classification metrics for GradientBoosting model:

	precision	recall	f1-score	support
0	0.95	1.00	0.97	72
1	1.00	0.94	0.97	70
accuracy			0.97	142
macro avg	0.97	0.97	0.97	142
weighted avg	0.97	0.97	0.97	142

Figure 6: GBM classification report



Classification metrics for Random Forest model:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	72
1	1.00	0.96	0.98	70
accuracy			0.98	142
macro avg	0.98	0.98	0.98	142
weighted avg	0.98	0.98	0.98	142

Figure 7: Random Forest classification report

with both models, it is evident that temperature, dew point temperature, and humidity are the most significant features influencing the operation status of Seoul's public bike system.

5 Hourly Rented Bike Count Prediction

In this section, we aim to perform a regression analysis to predict the number of rented bike count based on various influencing factors. Understanding and accurately forecasting bike rental demand not only facilitate efficient resource allocation for bike-sharing systems but also inform urban planning and transportation management strategies.

The dataset at our disposal comprises historical records of bike rental counts alongside corresponding environmental and temporal variables. These variables include factors such as temperature, humidity, wind speed, workday or not, day of the week, and seasonality. By leveraging machine learning techniques, we seek to uncover the intricate relationships between these factors and the observed bike rental counts.

5.1 Data Pre-processing of Rented Bike Count Prediction

As illustrated by figure 2, the response variable has outliers and right-skewness. To transform the response variable so that it has a well-formed distribution, we tried natural logarithm of one plus, square root and cubic root transformation respectively. The distribution plot are as figure 9 shows.

Therefore, we choose square root transformation now that it provides a better formed distribution.

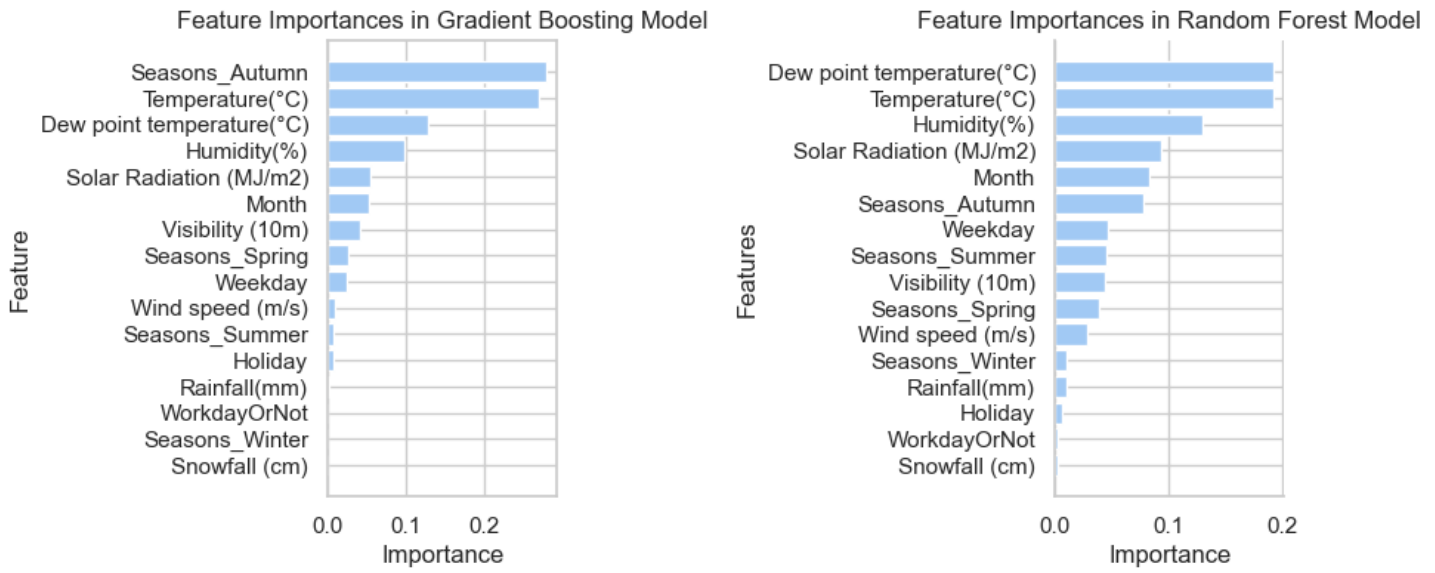


Figure 8: Feature importance of GBM and Random Forest Model

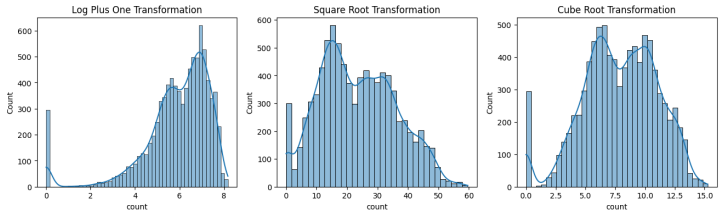


Figure 9: Distributions of Transformed Response Variable

5.2 Methodology and Result

Following the prediction of functioning days, we aim to develop machine-learning models capable of forecasting the number of shared bikes rented in Seoul. To evaluate the performance of different models, we use metrics including MSE, RMSE, MAE, R^2 and adjusted R^2 and figures giving a more direct comparison.

5.2.1 Regression Tree and Random Forest

Both decision trees and random forests are powerful machine learning algorithms used for classification and regression tasks. While decision trees are simple and interpretable, random forests offer improved accuracy and robustness by combining multiple trees trained on different subsets of data.

We first used a regression tree model to fit the data. After we set the minimum samples of every leaf as 20, minimum samples every split as 3, maximum depth as 20, the performance of the regression tree model is as table 2 and figure 10 show.

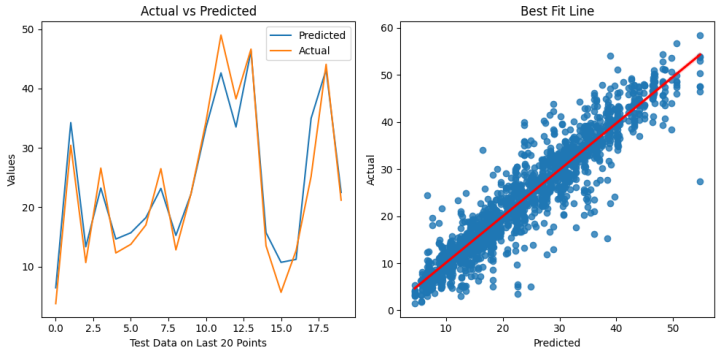


Figure 10: Decision Tree Result

	MSE	RMSE	MAE	R^2	Adj. R^2
Train	13.395	3.66	2.467	0.905	0.905
Test	16.915	4.113	2.836	0.878	0.877

Table 2: Decision Tree Performance

We then used grid search cross validation to tune the parameter of the random forest model, including "n_estimators", "max_depth", "min_samples_leaf", and "min_samples_split". The best parameter of random forest has 600 trees with max depth 10 and number of leaves 10. The performance of the random forest model is as table 3 and figure 10 show.

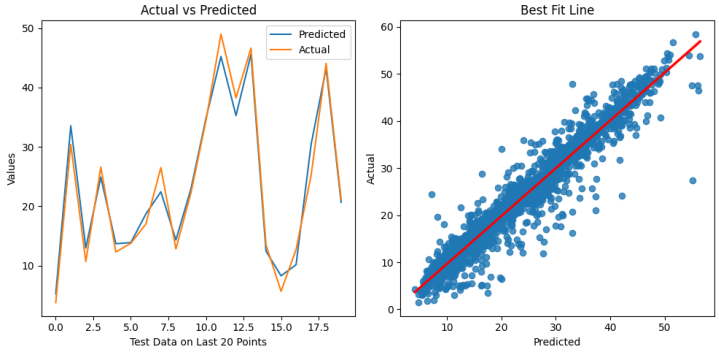


Figure 11: Random Forest Result

	MSE	RMSE	MAE	R^2	Adj. R^2
Train	6.257	2.501	1.615	0.956	0.956
Test	10.848	3.294	2.2	0.922	0.921

Table 3: Random Forest Performance

5.2.2 LightGBM

As we introduced in the classification part, lightGBM is a specific implementation of the gradient boosting framework. Compared with gradient boosting, lightGBM uses a leaf-wise growth strategy instead of level-wise (depth-first) growth used

in traditional gradient boosting implementations. This means it grows the tree leaf-wise, choosing the leaf with the maximum delta loss to grow at each step. This strategy can lead to a better fit for the data but might also result in overfitting if not properly regularized.

Different ways of parameter tuning of lightGBM have different effects. To achieve better accuracy, it is advised to use large "max_bin", small "learning_rate", large "num_iteration", large "num_leaves" and bigger traning data. Since we aim to get prediction result, we adopt this way of parameter tuning. With learning rate 0.05, max_bin 160, num_leaves 30, we get the best model. The performance of the lightGBM model is as table 4 and figure 12 show.

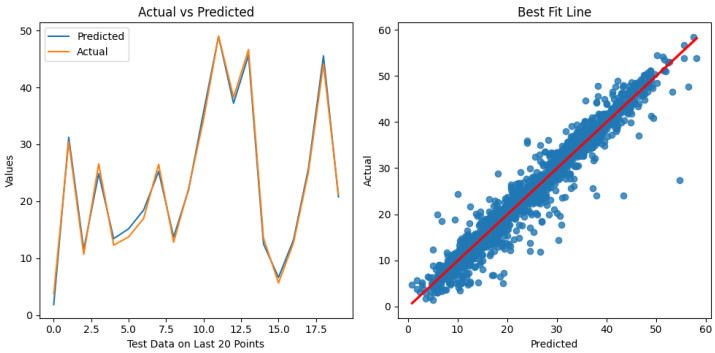


Figure 12: LightGBM Result

	MSE	RMSE	MAE	R^2	Adj. R^2
Train	7.277	2.698	1.855	0.948	0.948
Test	9.558	3.092	2.153	0.931	0.931

Table 4: LightGBM Performance

5.2.3 Long Short Term Memory Networks

In our previous EDA, we observed that although the hourly rented bike count does not follow a specific distribution, it exhibits daily seasonal characteristics. Consequently, we opted to employ Long Short-Term Memory Networks (LSTMs) to embrace the daily cycle and boost prediction accuracy. As a variant of Recurrent Neural Network, LSTMs excel in capturing long-term dependencies within data, which is essential for our objective to reflect the daily variations in bike rental numbers. To enhance the LSTM model's ability to recognize temporal patterns, we included both the identified trend and seasonal components in our training dataset after performing a seasonal decomposition on the square-root transformed rented bike counts. Additionally, we utilized a random search for hyperparameter optimization of the neural network. To maintain the data's temporal integrity, we implemented a time-sliding window method for evaluation, predicting the bike-sharing rental count for the next hour based on the data from the preceding 24 hours.

Our final model is a Sequential model with 200 neurons per layer and a dropout rate of around 0.3. Below is its performance on the testing set. (Since we use time-

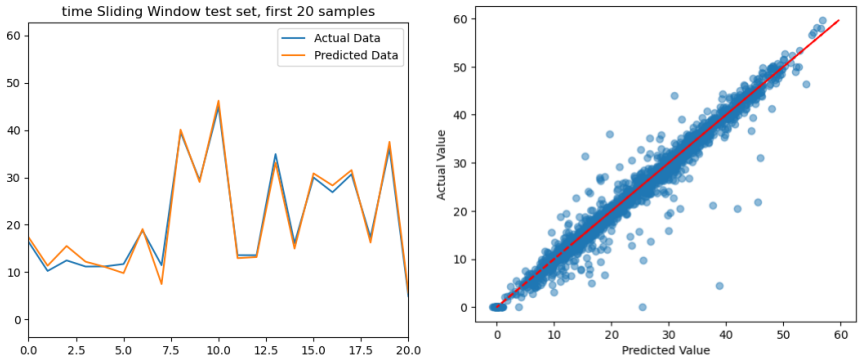


Figure 13: LSTM Result

	MSE	RMSE	MAE	R^2	Adj. R^2
Train	2.04	1.43	1.01	0.948	0.987
Test	6.22	2.49	1.45	0.931	0.958

Table 5: LSTM Result

As we can see from above, the LSTM network, a deep learning architecture renowned for its sequential modeling capabilities, emerged as the standout performer in our analysis. Leveraging its ability to capture temporal and seasonal dependencies and long-range dependencies inherent in time series data, LSTM achieved the lowest test MSE and RMSE despite that its parameter tuning process is very time consuming . This suggests that the intricate temporal and seasonal dynamics of bike rental patterns are effectively captured by LSTM, enabling more precise predictions compared to traditional machine learning approaches.

While decision tree, random forest, and lightGBM provide viable options, particularly for interpretable models and computational efficiency, LSTM stands out as the optimal choice in terms of predictive accuracy. However, the selection of the most suitable methodology should consider the specific requirements and constraints of the application, balancing between accuracy, interpretability, and computational resources.

6 Conclusion

After conducting KS test, we find out that there is no simple distribution fit to the rented bike count, therefore we decided to use non-parametric methods to fit the distribution of the rented bike count.

Our classification analysis has demonstrated robust performance, achieving high accuracy and precision in predicting the functioning or non-functioning days. Through meticulous examination, it is evident that temperature, dew point, and humidity emerge as pivotal features in our classification task. These variables have exhibited significant importance in determining the outcome, showcasing their indispensable role in influencing the target classification.

Besides, our regression analysis has achieved exceptional accuracy and precision in forecasting the number of rented bike

counts. Through comprehensive experimentation and evaluation, lightGBM has emerged as the paramount model, showcasing superior long-term predictive performance compared to alternative methodologies. Its ability to capture intricate patterns and nonlinear relationships within the data has rendered it indispensable in accurately predicting bike rental counts. However, we find that every one of our regression models is unable to predict precisely the extreme values of all the data. We suspect that it has something to do with the square root transformation of the response variable. After this transformation, the general performances of models are improved, but this flaw is brought about unfortunately. We hope to find better models or better transformation methods in the future so that extreme values can be better fit to and predicted.

While the current models show promising results, there are notable flaws that should be addressed to enhance its reliability and generalization capabilities. One major concern is the potential for overfitting, where the model performs exceptionally well on the training data but fails to generalize to unseen data. To mitigate this issue, regularization techniques such as dropout or L2 regularization could be applied in the future during the training process to prevent the model from becoming overly complex and fitting noise in the data. Additionally, the dataset used for training the model covers only the period of one year, leading to insufficient coverage of the underlying data distribution. To address this, collecting more observations across a wider range of period and ensuring a balanced representation of different classes or categories within the data could help improve the model's robustness and accuracy. By incorporating these adjustments, the models could achieve better performance and reliability when it comes to classification or regression.

Appendix

All of our code, data and supporting materials can be found on https://github.com/Isawsomethingb4/Seoul-Bike-Sharing-Demand-Analysis.

Data: SeoulBikeData.csv

Supporting files: my_dir folder. It contains the optimized hyperparameters for the LSTM model. Please put this folder in your current directory.