Seoul Bike Sharing Demand

Jordan Larot & Amy Tan

San Jose State University

BUS2 193: Data Mining

Dr. Shaonan Tian

May 9, 2024

ABSTRACT

This study analyzes the Seoul Bike Sharing Demand dataset to identify key factors influencing bike rental demand and forecast the number of rental bikes per hour. The data, sourced from the UC Irvine Machine Learning Repository, covers bike usage from December 2017 to November 2018 and includes seasonal, temporal, and weather-related variables. Initial data preparation involved correcting data types and parsing dates, followed by feature engineering to extract predictors such as day of the week, month, and weather conditions. Using an 80-20 train-test split, three predictive models were developed and evaluated: Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost). The XGBoost model performed the best on the testing set, achieving the lowest Root Mean Squared Error (RMSE) of 135. From the XGBoost model, the key factors influencing bike sharing demand were identified: hour, humidity, solar radiation, and temperature. Based on these findings, we recommend adjusting bike availability using forecasted weather conditions, implementing dynamic pricing strategies, and creating real-time information mobile applications. These suggestions will improve operational efficiency and enhance the user experience for Seoul's bike sharing service.

1. INTRODUCTION

1.1 Background

The need for sustainable modes of transportation, growing traffic congestion, and environmental concerns have presented many challenges for urban transportation systems in recent years. Bike sharing systems have become a popular solution in many cities around the world. Seoul, the capital of South Korea, is one city that has introduced bike sharing as part of its transportation infrastructure. The Seoul Bike Sharing System, operated by Seoul Metropolitan Government, is a network of docking stations strategically located in high traffic districts throughout the city. Users can rent bicycles from one station and return them to any other station within the network, offering an affordable, convenient, and environmentally friendly transportation option to residents and visitors.

1.2 Objectives

The objectives of this project are to (1) identify key factors influencing bike sharing demand and (2) build a model to forecast the number of rental bikes per hour. Understanding the dynamics of bike sharing demand is essential for optimizing the system's operations, enhancing user experience, and planning policy decisions.

1.3 Data Source

The Seoul Bike Sharing Demand dataset, sourced from the UC Irvine Machine Learning Repository, provides historical data of bike sharing usage in Seoul from December 1, 2017 to November 30, 2018. Seasonal, temporal, and weather data are included. To see the nature of the variables, please refer to *Appendix A*.

2. METHODOLOGY

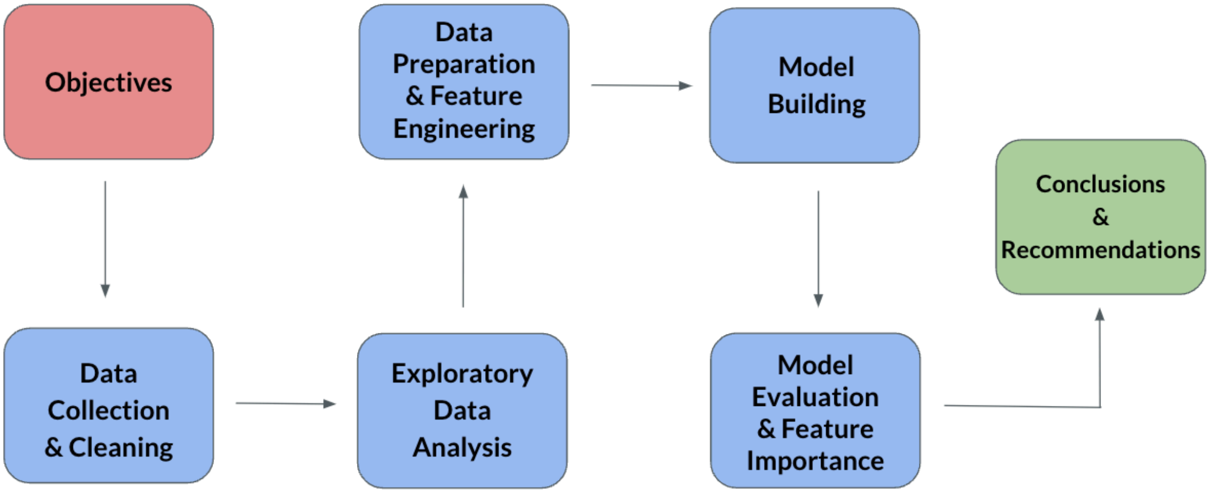
Objectives
Data Collection & Cleaning
Exploratory Data Analysis
Data Preparation & Feature Engineering
Model Building
Model Evaluation & Feature Importance
Conclusions & Recommendations

Figure 1: Methodology for Seoul Biking Demand Dataset

2.1 Data Collection & Cleaning

It was essential to convert the variables to their correct data types. Firstly, the ‘Date’ column was not being parsed correctly; therefore, we needed to properly format it. Next, the following features were converted to a factor in R: ‘Seasons’, ‘Holiday’, and ‘Functioning Day’. Afterwards, the data was ready for exploratory data analysis.

2.2 Data Preparation & Feature Engineering

After exploratory data analysis, we needed to prepare the data for the machine learning model. First, we removed days that were non-functioning because these rows had a rented bike count of 0. Secondly, the following features were extracted from the ‘Date’ column: ‘Day of Week’, ‘Month’, and ‘Weekend’. These temporal features can significantly influence rental bike patterns; for instance, the demand of bikes may vary drastically across different days of the week and month. In addition to extracting the temporal features, weather-related predictors were created, such as ‘Is Rainy’, ‘Is Snowy’, and ‘Weather Condition’. While precipitation, snowfall, and temperature provide information about weather conditions, these new variables offer a more nuanced understanding of how different weather scenarios affect bike sharing demand. ‘Is Rainy’ and ‘Is Snowy’ are binary features; while ‘Weather Condition’ is a categorical variable with three unique possible values: Cold (Temperature < 41℉), Mild (41℉ ≤ Temperature ≤ 77℉), and Hot (Temperature < 77℉). These thresholds were established based on what is typically considered cold, hot, and mild in Seoul. Lastly, the data was split into 80% training and 20% testing, with a random seed of 42 set to ensure reproducibility.

2.3 Model Building

Using the training data, three models were developed: Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost). Both Random Forest and XGBoost are ensemble methods. Random Forest combines multiple weak learners to create a stronger overall model, while XGBoost enhances this approach through boosting–each new model focuses on correcting the errors made by the previous ones, thereby improving model accuracy. Among these models, XGBoost is the most complex, whereas Decision Tree is the simplest. Feature selection was employed to optimize model performance. Lastly, the parameters for each model were manually tuned to improve accuracy, generalization, and robustness.

2.4 Model Evaluation & Feature Importance

To compare the performance across the different models, we use the following metrics: Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Mean Squared Error calculates the difference between the predicted and actual values, squaring the errors. For easier interpretability, we use RMSE, which is measured in the same units as the target variable. Lower values for each of these metrics indicate that the model's predictions are closer to the actual values, which is desirable. Based on these metrics, once the best-performing model is identified using the testing set, we examine its feature importance using the gain method. The gain method quantifies how much each feature contributes to the model's predictive power. A higher gain indicates a more significant impact on model accuracy.

3. RESULTS

3.1 Exploratory Data Analysis

3.1.1 Descriptive Statistics

	Rented Bike Count	**Temperature (C)**	**Snowfall (cm)**
Min	0.0	-17.8	0.0
1Q	191.0	3.5	0.0
Median	504.5	13.7	0.0
Mean	704.6	12.9	0.1
3Q	1065.2	22.5	0.0
Max	3556.0	39.4	8.8

Table 1: Summary Statistics of Numeric Variables in the Seoul Biking Dataset

Based on Table 1, the range of rented bike counts is between 0 and 3556 bikes rented per hour. This suggests high variability in rental activity, with periods of no rentals to unusually high rental activity during certain periods. In addition, the temperature observed in the dataset shows a significant amount of variability, with both negative and positive values. The lowest temperature is -17.8℃, while the highest temperature is 39.4℃. This is likely due to the data representing temperatures over different seasons, capturing both cold and warm temperature extremes. Lastly, the average snowfall of 0.1 cm suggests that for the majority of the observed period, there was little to no snowfall; however, there was one instance of snowfall recorded to be 8.8 cm.

3.1.2 Data Visualization

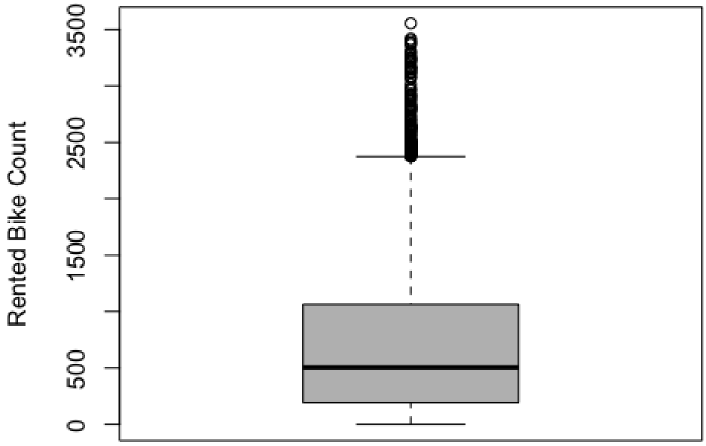

Figure 2: Boxplot of the Rented Bike Count

Based on Figure 2, the majority of the observations have rented bike counts of 504.5 or less. There are apparent outliers in the rented bike count with points extending up to 3,556 bikes rented per hour. This suggests that there is high variability in the rented bike count.

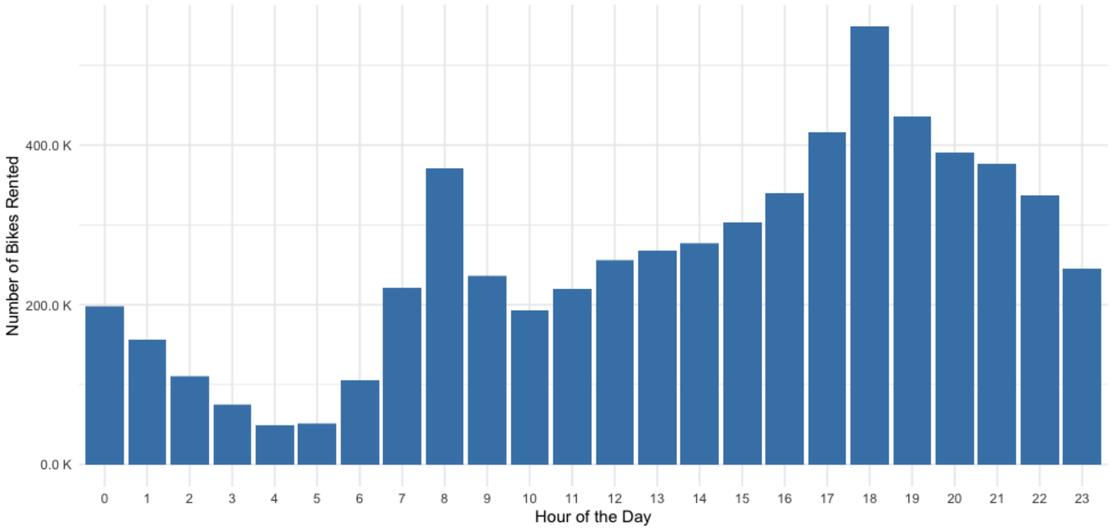

Figure 3: Total Number of Bikes Rented Per Hour

Figure 3 displays a column chart representing the aggregated number of bikes rented for each hour in a 24-hour period collected over the observation period. The data shows a bimodal distribution with two distinct peaks occurring at hours 8 and 18, which represents 8 AM and 6 PM, respectively. In general, we can expect there to be a higher number of bike rentals during these hours which align with typical commuting times. The 8 AM peak is when people are likely commuting to work or school, while the peak at 6 PM aligns with the evening rush hour as many people are returning from school or work. The highest spike occurs at 6 PM, indicating the highest demand of bike rentals throughout the day. This pattern of bike rentals suggests that the time of day may be a significant factor in predicting bike rental volumes.

3.2 Model Performance

Model	**MSE**	**RMSE**
Decision Tree	131,302	362
Random Forest	23,307	153
XGBoost	18,174	135

Table 2: Model Performance Comparison on Testing Set: MSE and RMSE

Based on Table 2, the XGBoost model performs the best on the testing set, having the lowest values for MSE and RMSE. A RMSE of 135 means that, on average, the predictions are off by 135 bikes. The random forest model performs slightly worse compared to the XGBoost model as its RMSE is 12% higher. On the other hand, the Decision Tree model performs considerably worse with a RMSE of 362.

3.3 Feature Importance

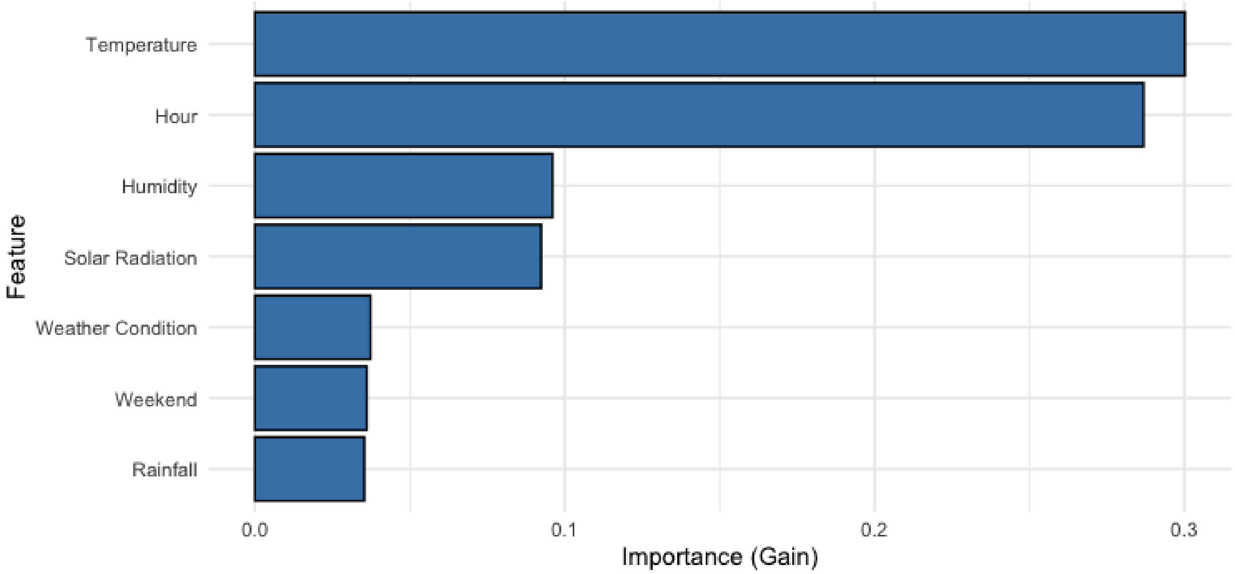

Figure 4: Top 7 Most Important Features (Gain Method) of XGBoost Model

'Temperature' is the most influential variable affecting bike sharing demand, contributing significantly more to the model's predictive power than any other feature. 'Hour' also shows a significant impact. 'Humidity' and 'Solar Radiation' are important predictors as well. ‘Weekend’ indicates a difference in rental patterns between weekdays and weekends, while 'Rainfall' shows that precipitation affects the rental bike demand.

4. DISCUSSION

4.1 Interpretation of Results

4.1.1 Model Performance

Although an RMSE of 135 might appear high, given the complexity and challenges in the Seoul Bike Sharing Demand dataset, it indicates a reasonably good performance of the XGBoost model. XGBoost and Random Forest significantly outperform the Decision Tree. This can be due to the following reasons: overfitting and model complexity. Decision Trees tend to overfit the data considerably, meaning that the model is memorizing too much of the patterns from the training set and failing to generalize well to unseen data. Additionally, the model may not be sophisticated enough, and the data may exhibit complex relationships. A simple Decision Tree is likely unable to capture these patterns. In contrast, the ensemble methods used by the XGBoost and Random Forest models are able to mitigate overfitting by averaging the outputs from multiple models, allowing it to capture more complex patterns. Therefore, XGBoost and Random Forest are better suited to accurately model bike sharing demand.

4.1.2 Feature Importance

The most important predictors tend to be weather-related features, such as humidity, solar radiation, and temperature. These factors influence the comfort and safety of cycling. For instance, extreme temperatures may deter customers from cycling, while high humidity can make the experience less enjoyable. Additionally, the time of day (hour) and weekends significantly impact cycling convenience. Bike sharing demand fluctuates with daily routines, such as commuting during rush hours and more recreational time to cycle on weekends. Together, these factors explain how and when people engage with bike sharing services.

4.2 Recommendations

4.2.1 Weather Forecasts

City transportation planners could adjust bike availability based on weather forecasts, particularly temperature, solar radiation, and humidity. These three weather features were found to significantly impact bike rental demand and should be primary considerations in operational planning. This will ensure that there are sufficient bikes available during optimal weather conditions, which would increase user satisfaction and maximize rental opportunities. Additionally, reducing the number of bikes during adverse weather conditions, such as heavy rain, could help manage resources more efficiently and contribute to safer and more organized streets. This minimizes the likelihood of bikes being left unused and exposed to weather-related damage and helps prevent congestion on sidewalks and bike racks. Overall, this approach aims at aligning bike availability to user demand patterns influenced by weather conditions.

4.2.2 Dynamic Pricing Strategies

Dynamic pricing strategies could be implemented using the real time data of the current demand and supply of bikes at different locations. For example, offering discounts during times when bike rentals are less frequent will help to incentivize more people to choose this eco-friendly mode of transportation. This would also increase the utilization of bikes that would otherwise remain idle and promote environmental sustainability by broadening the user base for bike sharing. Moreover, higher prices can be set at stations where demand is high but bikes are scarce to help manage demand.

4.2.3 Real-Time Information Apps

To improve user experience, we suggest implementing real-time information in mobile applications used for bike rentals. Helpful features to include are bike availability, weather updates, and route planning. Bike availability would show the current number of bikes and docking spaces available at each station, which would help users plan their trips more efficiently by allowing them to see where they can pick up and drop off bikes. Real-time weather forecasts would aid users in making decisions about when to rent a bike and whether it is safe to do so.

Route planning would incorporate GPS navigation to suggest the best routes for biking, considering factors like bike lanes, traffic, and weather changes. These features would help make the bike sharing experience safer and more enjoyable.

5. CONCLUSION

This project analyzes the Seoul Bike Sharing demand dataset using Decision Tree, Random Forest, and XGBoost to predict bike sharing demand. XGBoost outperformed the other models, yielding the lowest values in MSE and RMSE. Through XGBoost, we identified key predictors impacting bike rental demand, including time of day, temperature, solar radiation, and humidity. These factors are crucial in understanding demand fluctuations, which assist in making informed decisions regarding resource distribution and operational planning. Given the performance of the XGBoost model, we recommend using it to forecast the number of rented bikes per hour for Seoul's bike sharing system. Utilizing this model can enhance service efficiency by enabling more precise adjustments in bike allocations during peak hours and varying weather conditions, thus optimizing the overall user experience and operational efficiency of Seoul's bike sharing system.

REFERENCES

Seoul Bike Sharing Demand. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5F62R.

APPENDIX A

Variable; ***data type***	**Definition**
Rented Bike Count; *int*	The number of bikes rented in the system at each hour of day. This is the response variable.
Date; *date*	The day the bike was rented from the system; formatted as day-month-year
Hour; *int*	Each number represents a specific hour of the day; formatted in 24-hour clock format, ranging numerically from 0 to 23.
Temperature; *numeric*	The average outdoor temperature recorded at each hour in Seoul. This value is measured in degrees Celsius.
Humidity; *int*	The measure of the amount of water vapor present in the air at each hour, represented as a percentage.
Wind Speed*; numeric*	The speed at which the wind is blowing at each hour, measured in meters per second (m/s).
Visibility (10m); *int*	The distance at which objects can be clearly seen and identified at each hour, measured in meters.
Dew Point Temperature; *numeric*	The temperature at which air becomes saturated with moisture, causing water vapor to condense into dew.
Solar Radiation (MJ/m2); *numeric*	The measure of the intensity of the sun's rays reaching the Earth's surface; recorded in megajoules per square meter.
Rainfall (mm); *int*	The amount of precipitation that has fallen during a specified time period and is reported in millimeters.

Snowfall (cm); *int*	The quantity of snow that has accumulated over a specific time period and is reported in centimeters.
Seasons; *categorical*	Categorized into the following four periods: Spring, Summer, Autumn, and Winter.
Holiday; *categorical*	Indicates whether a particular day is a national holiday or not.
Functioning Day; *boolean*	Indicates whether the bike rental system is operational on a given day.