An Analysis of Seoul Bike Rentals to Provide Recommendations for Implementing a Dynamic Demand-Based Pricing Model

Emma Hsu, John Hemmer, Kessupa Phopiboonsuk, and Jacqueline Chen

Introduction

In South Korea, where 85.2% of emitted CO_2 comes from transit vehicles, the rise of shared-bike rentals has been nothing short of a transportation revolution. Currently, the government is allocating more resources to help the country achieve its goal of being carbon neutral by 2050 while keeping the accessibility and convenience of transportation to its civilians at the forefront of priorities. This data science project, "An Analysis of Seoul Bike Rentals to Provide Recommendations for Implementing a Dynamic Demand-Based Pricing Model," aims to provide insights to improve bike usage given environmental factors.

Ttaruengyi, the most accessible bike system subsidized by the government, has invested more money to improve user convenience to increase the usage of bikes while mitigating the issue of carbon emission. Users of the app have increased since the system's launch in 2015 from 34,162 to 668,725 users in 2019. The implementation of bikes around the country has since led to a decrease of roughly 2,000 tons of carbon emissions. This has helped resolve the issue of traffic congestion and will continue to have a greater effect on the city as bikes become even more accessible to everyone across the country.

With our analysis insights, we want to encourage usage of bike rentals during non-peak hours through flexible rental rates, similar to other ride-sharing options, like Uber. Various rental rates would depend on factors such as demand during various hours of the day, current weather, and more. Through the findings, we hope to do so by discovering the following patterns:

1. What three variables affect the usage pattern of shared bike rentals the most in South Korea; to what degree do they affect the decision process?
2. What is a better indicator for shared bike rentals– hour or temperature?

This report plans to evaluate the various factors that affect bike-sharing in South Korea in hopes that the recommendations and analyses can be adopted by other countries or urban planners. Through a more uniform utilization rate, more cars can be kept off the streets and the air quality of the metropolis would be improved simultaneously.

Domain Knowledge

Bike riding stands as a sustainable alternative compared to other modes of transportation given its significant benefits to the environment. Unlike cars and buses, bicycles produce zero carbon dioxide during their operation and can reduce an average person’s emissions from their use of transportation by

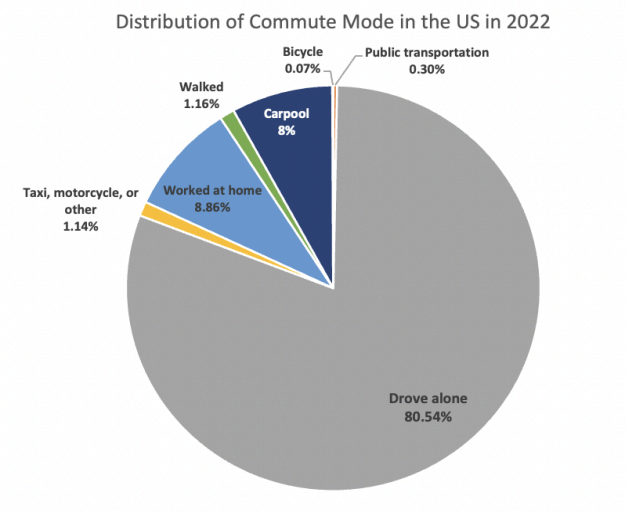
67%. This reduction would help mitigate climate change and alleviate the discharge of pollution trapped in the surrounding atmosphere. Air quality will increase along with people's livelihood, fostering a cleaner and healthier environment.

Not only do cars emit carbon dioxide, but also a significant amount of noise pollution. As roads are present in national parks and historical areas, cars will be present there. These sounds introduce environmental impacts, ultimately affecting animal ecosystems and vulnerable communities' health conditions. With bicycles, noise pollution is not an issue as it does not have an engine. Hence, they do not disturb communities nor contribute to light pollution.

According to the Bureau of Transportation, the most common mode of transit in the United States in 2022 is cars. To make matters worse, these cars contain single drivers, making them even more vulnerable to the environment. In the distribution, 0.07% use bikes as their mode of transit, which is the lowest in the distribution.

Breakdown of the commute mode to work:

80.54%	Drove alone
8.86%	Worked at home
8%	Carpool
1.16%	Walked
1.14%	Taxi, motorcycle, or other
0.30%	Public Transportation
0.07%	Bicycle



Data Sources & Acquisition

The dataset, obtained from the Machine Learning Repository at UCI, provides the number of daily rentals, the hour, the temperature, the season, if it was on a holiday, the wind speed, and more. The target variable is the rented bike count and with this information, statistical conclusions can be drawn and recommendations can be made to provide recommendations of prices for bike rentals given the variables that affect it the most.

The dataset was downloaded as a CSV file with 14 columns and 8760 rows (not including the column headings). The first two columns were 'Date' and 'Rented Bike Count' and the following 12 were different variables affecting the input, such as 'Wind speed (m/s)' and 'Dew point temperature (°C)'. There were no missing values, though 295 of the inputs have zero as their rented bike count. To visualize the distribution, the file was loaded into a Colab Notebook where packages were loaded to conduct further analysis.

It is important to note that while the data was collected in 2017-2018, the focus of our analysis is on categorizing bike rental volume behavior based on factors that are not year-dependent, such as environmental conditions, the day of the week, and whether or not the day falls on a holiday. Hence, the exact volume of rentals is not the focus of this analysis as year-to-year conditions such as the economy and the popularity of the service can fluctuate.

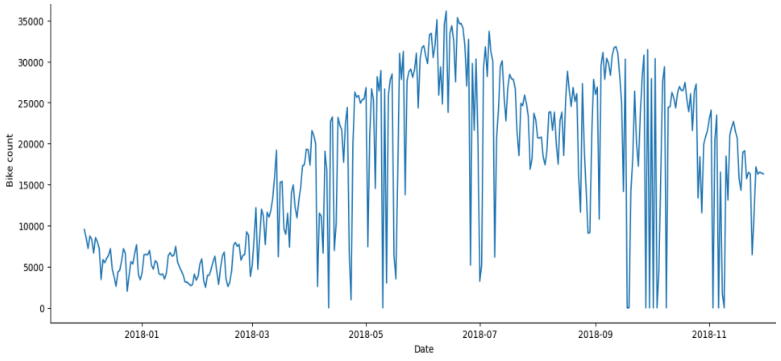
Exploratory Analysis

Univariate Analysis

To perform the analysis, the modified dataset was used to explore the correlations between different variables and bike counts. Through this investigation, correlations and patterns were analyzed and significant variables were chosen for predictive analysis.

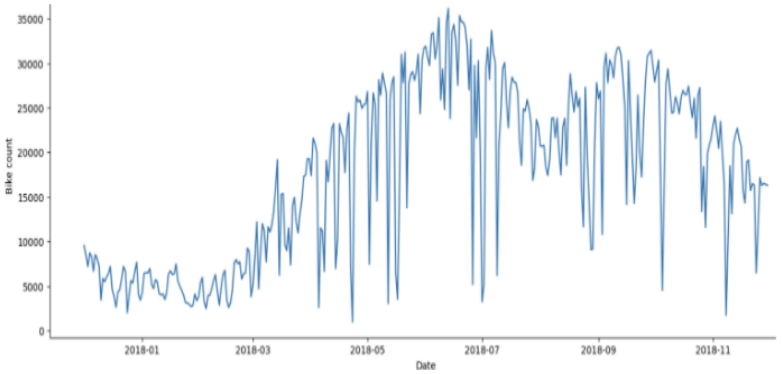
The dataset contains 8,760 records, each representing an hour over the course of one year, from December 1, 2017, to November 30, 2018. This results in an even distribution of data across months and seasons as no records are missing in the dataset. However, upon visual inspection of bikes rented by day over the course of the year, you can see that there are days where the rented bike count falls to 0. We believe that this is an error as these specific days are not holidays and do not have particularly extreme weather conditions that prevent biking. After removing the days with 0 rented bike count, there are now 8465 data points with this modified plot:

Bikes Rented by Day (with zeros)



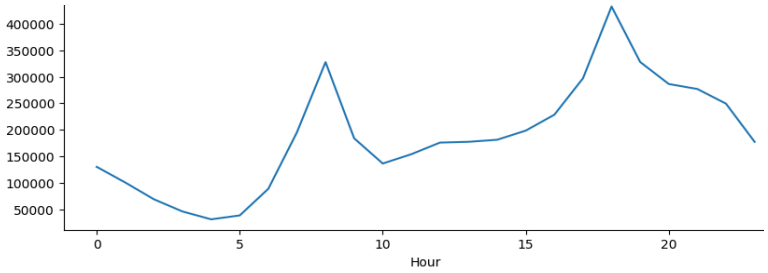
There are fewer unexplainable drops. While there are still days with alarmingly low bike rentals, we did not remove these days from our analysis because we could not be certain if they were valid or not.

Bikes Rented by Day

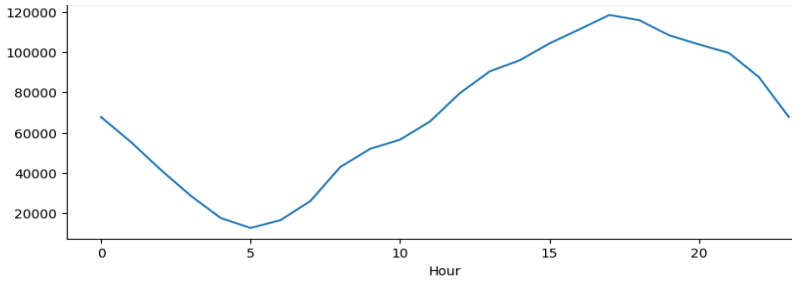


Now, we will examine the volume of bike rentals on weekdays versus weekends, which are notably different. On the weekdays, the two peaks are around 8 am and 6 pm, which is when people typically go to work and arrive back home. As for the weekends, it gradually increases throughout the day and decreases at around 4 pm.

Rented Bike Count by Hour of the Day (Weekdays)

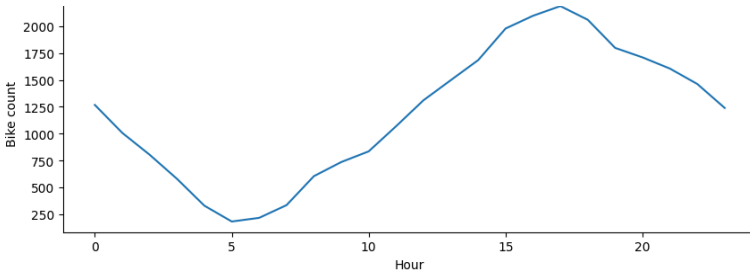


Rented Bike Count by Hour of the Day (Weekends)

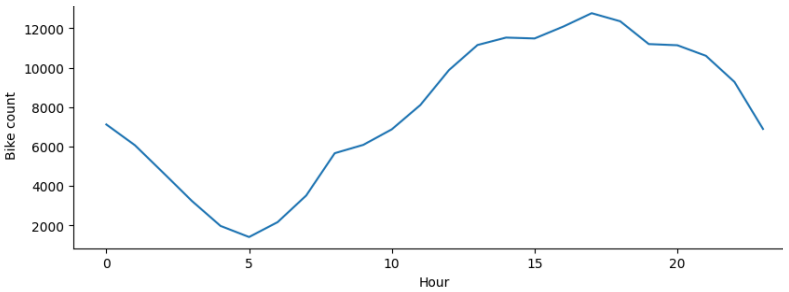


It appears that the behavior of the distribution of bike rentals on holidays resembles that of weekends, suggesting that it is important to include whether or not a day was a holiday in our models.

Rented Bike Count by Hour of the Day (Holidays on Weekends)



Rented Bike Count by Hour of the Day (Holidays on Weekdays)



Bivariate Analysis

To explore the analysis of the relationship between two variables, a correlation matrix is used. Each cell in the matrix represents the correlation between two variables, where the r values range from -1 to 1 with 1 representing a perfect positive correlation and -1 representing a perfect negative correlation. Those with a darker color imply a strong correlation, such as ‘Temp’ and ‘Dew point temp’ with an r-value of 0.914 and ‘Seasons’ and ‘Month’ and a strong, negative r-value of -0.703. Given that temperature has a higher significance than dew point in the XGBoost above, and the two features have a high correlation, it is reasonable to only select the temperature for the model.

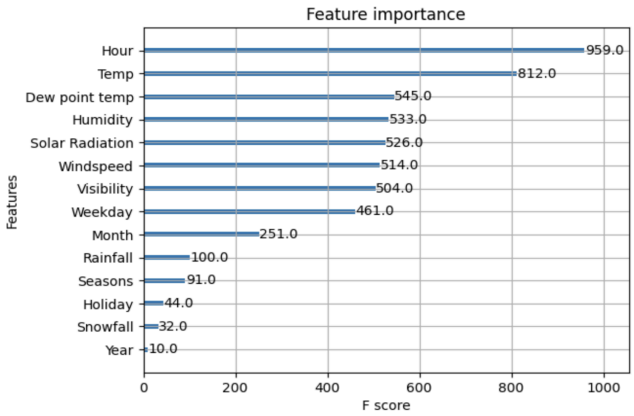
	Bike count	Hour	Temp	Humidity	Windspeed	Visibility	Dew point temp	Solar Radiation	Rainfall	Snowfall	Seasons	Holiday	Month	Year	Weekday
Bike count	1.000000	0.425256	0.562740	-0.201973	0.125022	0.212323	0.400263	0.273862	-0.128626	-0.151611	-0.313531	0.070070	0.234110	0.232004	-0.000371
Hour	0.425256	1.000000	0.122742	-0.235937	0.287780	0.103869	0.004691	0.144658	0.014345	-0.022082	-0.001420	0.000229	0.001379	0.000315	-0.000491
Temp	0.562740	0.122742	1.000000	0.166425	-0.038481	0.028262	0.914467	0.354844	0.052149	-0.217746	-0.333517	0.057977	0.137871	0.377003	0.000641
Humidity	-0.201973	-0.235937	0.166425	1.000000	-0.337352	-0.548542	0.539402	-0.457273	0.236917	0.110127	-0.120001	0.047796	0.081325	0.035188	0.031230
Windspeed	0.125022	0.287780	-0.038481	-0.337352	1.000000	0.180428	-0.177170	0.326222	-0.024931	-0.003789	0.108327	-0.031432	-0.113032	-0.003568	0.030484
Visibility	0.212323	0.103869	0.028262	-0.548542	0.180428	1.000000	-0.182586	0.153046	-0.170352	-0.122860	-0.006786	-0.022210	0.011745	0.051802	0.039295
Dew point temp	0.400263	0.004691	0.914467	0.539402	-0.177170	-0.182586	1.000000	0.098525	0.126812	-0.149760	-0.326239	0.067625	0.154209	0.334985	0.014561
Solar Radiation	0.273862	0.144658	0.354844	-0.457273	0.326222	0.153046	0.098525	1.000000	-0.074157	-0.073380	-0.080288	0.001963	0.028763	0.130141	-0.016576
Rainfall	-0.128626	0.014345	0.052149	0.236917	-0.024931	-0.170352	0.126812	-0.074157	1.000000	0.008604	-0.020083	0.013301	-0.000830	0.028228	0.017874
Snowfall	-0.151611	-0.022082	-0.217746	0.110127	-0.003789	-0.122860	-0.149760	-0.073380	0.008604	1.000000	0.142224	0.012043	-0.071683	-0.205030	0.047625
Seasons	-0.313531	-0.001420	-0.333517	-0.120001	0.108327	-0.006786	-0.326239	-0.080288	-0.020083	0.142224	1.000000	-0.057375	-0.703399	-0.410300	0.007792
Holiday	0.070070	0.000229	0.057977	0.047796	-0.031432	-0.022210	0.067625	0.001963	0.013301	0.012043	-0.057375	1.000000	-0.009607	0.117153	-0.030226
Month	0.234110	0.001379	0.137871	0.081325	-0.113032	0.011745	0.154209	0.028763	-0.000830	-0.071683	-0.703399	-0.009607	1.000000	0.309049	0.005407
Year	0.232004	0.000315	0.377003	0.035188	-0.003568	0.051802	0.334985	0.130141	0.028228	-0.205030	-0.410300	0.117153	0.309049	1.000000	0.014883
Weekday	-0.000371	-0.000491	0.000641	0.031230	0.030484	0.039295	0.014561	-0.016576	0.017874	0.047625	0.007792	-0.030226	0.005407	0.014883	1.000000

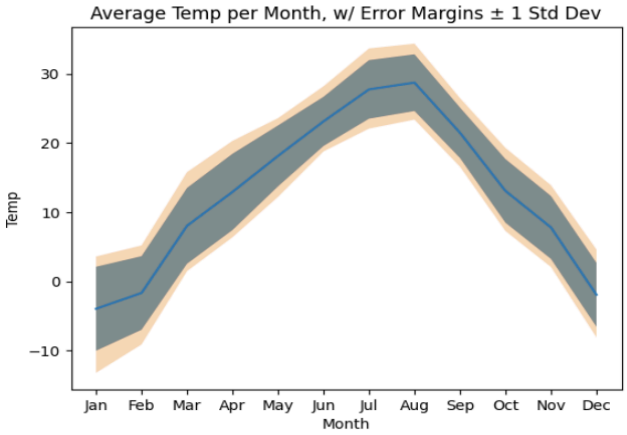
Methods and Experiment

XGBoost

XGBoost (Extreme Gradient Boosting) is used to help with the model selection step to determine important features. From XGBoost, we know that Temp and Hour are the most important numerical determinants in predicting bike count, so it is useful to visualize how the average temperature for each

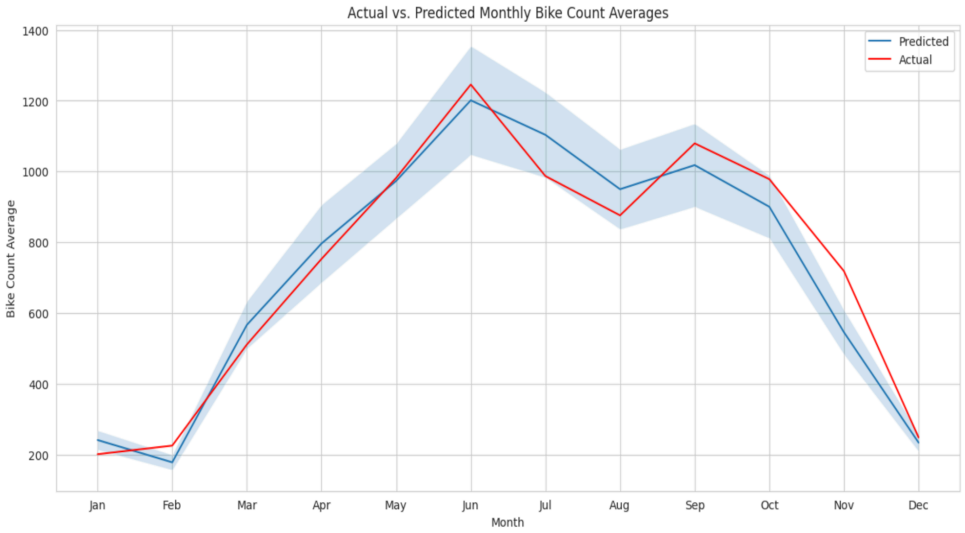
month changes. The blue margin indicates temperatures that are within 1 SD of the average. The yellow margin shows temperatures between the 10th and 90th percentile.





Random Forest

Due to its robustness to overfitting, ability to handle nonlinearity, and ability to adjust hyperparameters a Random Forest Regression Model was created to predict Bike Count based on the various explanatory variables. Random Forest can not only predict a continuous numerical outcome but can identify variable importance as well, thus verifying later variable selection from XGBoost. For this model specifically, dummy variables were created for all the categorical variables: season, holiday, month, and weekday. Each season creates a new variable (Seasons_winter, Seasons_summer, etc.), and if it is summer, Seasons_summer gets a 1, if not it gets a 0. The R-squared value was 0.86, meaning the model explains about 86% of the variance in Bike count. This will vary slightly for each run as the Random Forest changes each time, even with the same training and validation datasets. Since Random Forest can handle multicollinearity, all variables except for date were used. Including more variables increased the r-squared values on the test dataset, which is likely due to inherent feature randomness as well as the independence of individual tree creation.



For the selection of the hyperparameters, a grid search was used, which evaluates the model's performance with all of the possible parameter combinations and selects the best-performing set. In this case, it was 'max_depth': 150,

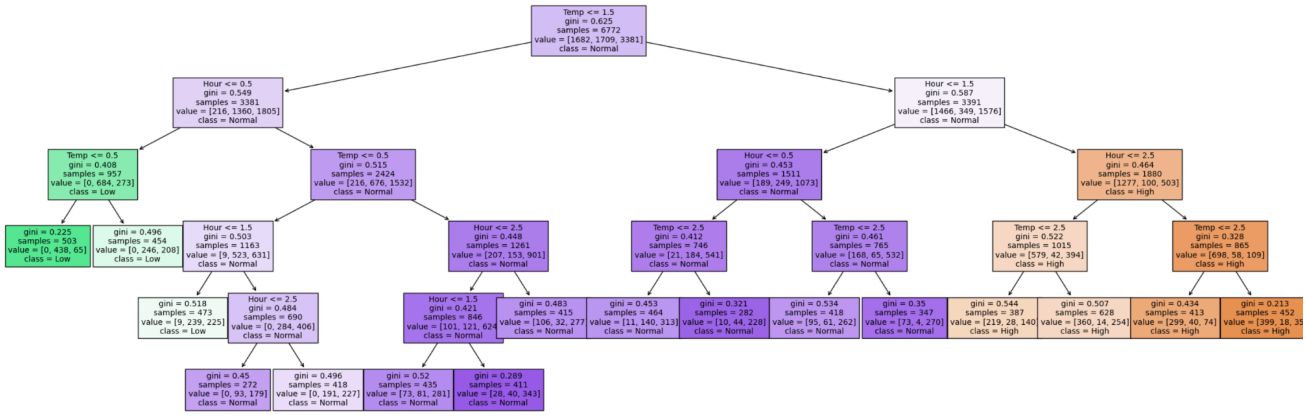
'max_features': 6, 'min_samples_leaf': 2, 'min_samples_split' : 12, 'n_estimators': 200. For visualization, monthly bike count averages were taken from the actual data and the predicted data and displayed on the same graph with the predicted data having a confidence interval around it. The Random Forest Model does a good job of capturing the seasonal behavior as well as local maxima and minima, but it is missing something in the Winter months.

Decision Tree

After determining that Hour and Temperature had the greatest influence on Bike count, a decision tree model was developed to further explore this relationship. The dataset was loaded into a pandas DataFrame, and the relevant columns ('Hour', 'Temp', 'Bike count') were extracted.

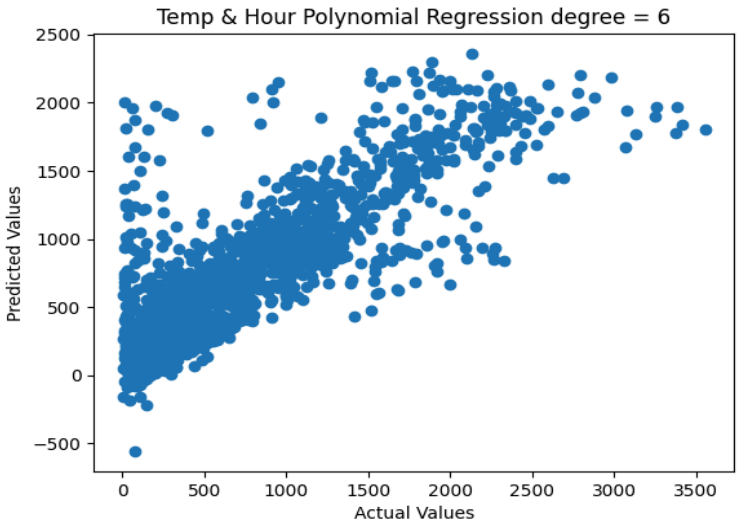
A binning strategy through pandas was employed to discretize the 'Bike count' variable. The bins were defined based on threshold values, taken from the 1st and 3rd quartile ranges, with labels assigned as 'Low', Normal, and 'High'. The goal was to transform the target variable into a categorical format suitable for a decision tree classification task. The KBinsDiscretizer from scikit-learn was used to create bins for the 'Hour' and 'Temp' features. The number of bins was set to 4, and the strategy chosen was 'uniform'. This strategy ensures an equal width for each bin, simplifying the interpretation of the bins and accommodating a range of values within each bin. Then, the dataset was split into training and testing sets using the train_test_split function from scikit-learn. The testing set comprised 20% of the data. A DecisionTreeClassifier was chosen as the modeling algorithm due to its ability to capture non-linear relationships between features and the target variable of ‘Bike count’. The Gini impurity in the decision tree shows the frequency of a random datapoint to be incorrectly classified. The model was trained on the training set using the fit method and the model was evaluated based on the accuracy score on the test set. The decision tree structure was visualized using the plot_tree function. The resulting plot provides insights into how the decision tree makes predictions based on the binned features. The decision tree rules, as extracted from the model, can be found in the printed output of the tree_rules variable. These rules describe the conditions at each node of the tree. The figure on the left shows all the rules pointing out low bike rental counts. The decision tree accuracy was 70.7%. ‘Normal’ indicates that the bike count given the Temp and Hour is fine and need not be changed. As a result, the focus is mainly on the green-colored boxes in the decision tree where the bike count is low. The goal is to pinpoint the cause and attempt to create solutions where the class is labeled ‘Low’.

```
|--- Temp <= 1.50
|   |--- Hour <= 0.50
|   |   |--- Temp <= 0.50
|   |   |   |--- class: Low
|   |   |--- Temp >  0.50
|   |   |   |--- class: Low
|   |--- Hour >  0.50
|   |   |--- Temp <= 0.50
|   |   |   |--- Hour <= 1.50
|   |   |   |   |--- class: Low
|   |   |   |--- Hour >  1.50
|   |   |   |   |--- Hour <= 2.50
|   |   |   |   |   |--- class: Low
|   |   |   |   |--- Hour >  2.50
|   |   |   |   |   |--- class: Low
```



Multinomial Polynomial Regression

From the results of XGBoost, Temperature and Hour were demonstrated to be features indicative of bike rental volume. Additionally, from exploratory analysis, it was found that the day of the week and whether or not the day was a holiday had a noticeable impact on bike rental volume. A multinomial polynomial regression model is the most fitting regression model for the number of features chosen to include and the variability in the data—a linear regression model could not capture the interactions of the features. The features for the day of the week and whether or not the record was for a holiday were one-hot encoded to allow these categorical variables to be features in a regression model. The test set was 20% of the data. After training the model, a degree=6 polynomial regression model using features temperature, hour, holiday, and day of the week yielded an R^2 value of 0.686, meaning that 68.6% of the variability of bike rentals in the records in the test set can be explained by the regression model. Since overfitting is more likely to happen as the degree of the polynomial increases, the degree hyperparameter was tuned with caution, and degree=6 was selected as it yielded the highest R^2 on the test set. An R^2 value of 0.686 indicates that this model is reliable in predicting rentals. Using this regression model, to the right is a graph comparing the actual and predicted bike rental volume values using the randomly generated test set. There are 848 summed terms in the regression model as the model is of degree 6 and there are 10 input variables—temperature, hour, holiday, 7 for each day of the week from the encoder.



Observation and Conclusion

Through the result of the XGBoost, we can answer our first research question–what three variables affect the usage pattern of shared bike rentals the most? The top three are hour, temperature, and dew point temperature. The degree to which they affect the process can also be seen in the same model, with hour coming in first and temperature second, making temperature a better indicator for the number of shared bike rentals. The Random Forest feature selection outputs Temperature, Dew Point Temperature, and Windspeed as the top variables in feature selection. Similar to XGBoost, Random Forest provides more importance to unique values, which lowers the priority of hours. Utilizing all variables, the Random Forest predicts bike count with around an 85% accuracy showing when there would be lower demand. To limit the complexity of a decision tree classifier and create a more interpretable model, only temperature and hour were chosen as features to include. Yielding an accuracy of ~70%, the decision tree can be used to determine at what times of the day, and at what temperature bike rentals fall short of a certain threshold count of 214 (25th percentile). To further analyze their relationship, a multinomial polynomial regression model was created. When degree=6, it yielded an R^2 value of 0.686, indicating the reliability of the model. Using these three models, business leaders for the rental company can predict and gain insight into the bike rental volume and adjust prices based on the following recommended business model.

To encourage the usage of bike rentals during predicted low-demand hours based on our three models, a promotion can be implemented where renting one bike can result in the second bike being half off. This would encourage greater usage of bikes as 2 people are needed for the campaign, in addition to promoting a safety measure as the hours start before dawn. Through this implementation, we hope to contribute to the mitigation of pollution by supporting the usage of bicycles as a mode of transportation to work, while benefiting the company by utilizing all the hours of the service.

References

Hallisey, Karen. “How Riding A Bike Benefits the Environment” UCLA. May 11, 2022. [Link]

Jo, Hanghun, Seong-A Kim, and Heungsoon Kim. “Forecasting the Reduction in Urban Air Pollution by Expansion of Market Shares of Eco-Friendly Vehicles: A Focus on Seoul, Korea” International Journal of Environmental Research and Public Health 19, no. 22: 15314. 2022. [Link]

Kim, Sujin. “How Seoul Eased Traffic Congestion and Reduced Pollution through Bike Sharing” Development Asia. May 30, 2023. [Link]

UCI Machine Learning Repository. “Seoul Bike Sharing Demand.” February 29, 2020. [Link]

United States Department of Transportation. “Commute Mode” Bureau of Transportation Statistics. 2022. [Link]