Prediction of Monthly YouBike Rentals in Taipei City

Yu-Wen Wu

1. Introduction

Whether it's due to the rise in environmental awareness, serving as the last mile of public transportation, or other reasons, YouBike shared bicycles have gradually become a part of people's daily lives. This study aims to explore the factors that cause changes in people's behavior towards using shared bicycles. This will not only help government departments formulate responsive strategies but also provide operational planning references for operators. To achieve the aforementioned objectives, this study examines related factors including bicycle infrastructure, weather conditions, the number of COVID-19 cases, and public transportation ridership, among others.

2. Problem Statement

This study primarily focuses on examining shifts in the monthly rental frequency of YouBike shared bicycles. The questions investigated encompass:

1. How does building bicycle infrastructure encourage more people to cycle?
2. Which weather factors significantly affect the usage rate of shared bicycles?
3. What impact does COVID-19 have on the frequency of bicycle rentals?
4. How does public transportation ridership correlate with shared bicycle usage rates?
5. Does the payment plan influence the willingness to use shared bicycles?

3. Analysis Approach

3.1. Data Preprocessing

The data on monthly bike rental frequency, bicycle infrastructure, weather factors, COVID-19 confirmed cases, and public transportation ridership were obtained from the Taipei City Government, the Ministry of Health and Welfare, and the Government Open Data Platform. The data covers the period from January 2011 to March 2024, comprising a total of 159 records. 20% of these records, amounting to 32 records, were randomly selected as the test dataset, while the remaining records were used as the training dataset.

3.2. Model Building

The models were built using two methods: traditional Linear Regression and the machine learning algorithm XGBoost. A comparison and analysis of the model results were conducted. For Linear Regression, the Akaike Information Criterion (AIC) was set as the criterion, and variable selection was performed through a bidirectional stepwise approach to obtain the final model. On the other hand, XGBoost utilized Mean-Square Error (MSE) as the criterion, with a maximum depth of 3, a learning rate of 0.05, and 1000 rounds of training to derive the final model.

4. Data Description

4.1. Variable Description

YouBike monthly rental frequency is closely related to many variables. The establishment of bicycle-related facilities may enhance accessibility and user willingness to use them. Weather conditions can affect the suitability of riding shared bicycles. The changes in people's lifestyles due to the COVID-19 pandemic may also influence the use of shared bicycles. Additionally, shared bicycles naturally have a certain degree of relevance as the last mile of public transportation. Therefore, this study collected 14 variables related to bicycles, weather, COVID-19, and public transportation, as detailed below.

4.1.1. Bicycle-Related Variables

1. RentalNum(K): The number of YouBike rentals in the current month, measured per thousand people, is the primary target of this study, aimed at predicting and exploring its correlations with other variables.

2. StationNum: The number of YouBike stations. The number of sites may affect accessibility, and the more sites there are, the more likely they are to cover a greater number of needs.

3. BikeNum(K): The number of YouBike bicycles, measured in thousands of units. The more bicycles there are, the more likely they are to meet a greater number of needs.

4. LaneLength(km): The length of bike lanes in Taipei, measured in kilometers. The construction of bike lanes may increase the public's willingness to use shared bicycles.

5. FreePlan: Indicates whether a free plan offering the first 30 minutes at no cost is provided for the current month. Assigned a value of 1 if available, otherwise 0. The execution of the plan may enhance the willingness of the public to use Youbike.

4.1.2. Weather-Related Variables

1. Temperature: The average temperature of the month, with higher temperatures possibly indicating better weather conditions, suitable for cycling.

2. Humidity(%): The average relative humidity of the month. High humidity can make weather feel stifling and increase sweating, potentially discouraging cycling.

3. Sunshine(hr): The monthly sunshine hours, with sunny weather being ideal for cycling.

4. Rainfull(mm): The monthly rainfall amount. Rainy days are less suitable for cycling, affecting people's willingness to rent bicycles.

5. Raindays: The number of rainy days in the month. Similar to the previous variable, it affects the number of bike rentals due to similar reason.

4.1.3. COVID-Related Variables

1. COVIDTaiwan(K): The monthly number of confirmed cases in Taiwan. The number of cases affects government and public responses, potentially impacting bicycle rentals.

2. COVIDTaipei(K): The monthly number of confirmed cases in Taipei. It may impact the target for similar reasons as COVIDTaiwan(K).

4.1.4. Public Transportation-Related Variables

1. MRTCarryNum(K): The monthly passenger volume of MRT. The higher the passenger volume, the more likely YouBike will be used.

2. BusCarryNum(K): The monthly passenger volume of city buses. The higher the passenger volume, the more likely YouBike will be used.

4.2. Variable Analysis

Figure 1 illustrates the variation in Youbike monthly rental frequency from January 2011 to March 2024. The Figure 2 presents the results of calculating Variance Inflation Factor (VIF) for all variables except RentalNum. It can be seen that BikeNum, LaneLength, COVIDTaiwan, COVIDTaipei, and BusCarryNum have relatively high VIF values. From the correlation heatmap in Figure 3, compared to LaneLength, BikeNum has a higher chance of collinearity with other bicycle-related variables, and thus it is filtered out. COVIDTaiwan clearly has the highest chance of collinearity with COVIDTaipei, and since COVIDTaiwan has a higher correlation with RentalNum, COVIDTaipei is filtered out. Variables with higher correlation coefficients with BusCarryNum have less direct relevance, so BusCarryNum is retained. Finally, the VIF values of the filtered variables are recalculated, and the results, as shown in Figure 4, present more reasonable.

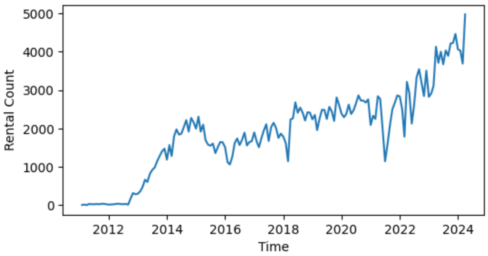
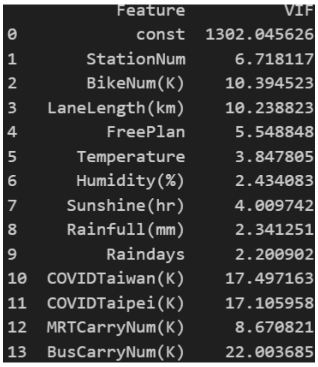


Figure 1: Scatter plot of rental frequency over time.



```
           Feature          VIF
0            const  1302.045626
1       StationNum     6.718117
2       BikeNum(K)    10.394523
3  LaneLength(km)    10.238823
4         FreePlan     5.548848
5      Temperature     3.847805
6      Humidity(%)     2.434083
7     Sunshine(hr)     4.009742
8     Rainfull(mm)     2.341251
9         Raindays     2.200902
10   COVIDTaiwan(K)   17.497163
11   COVIDTaipei(K)   17.105958
12  MRTCarryNum(K)     8.670821
13  BusCarryNum(K)    22.003685
```

Figure 2: The VIFs for all variables except RentalNum(K).

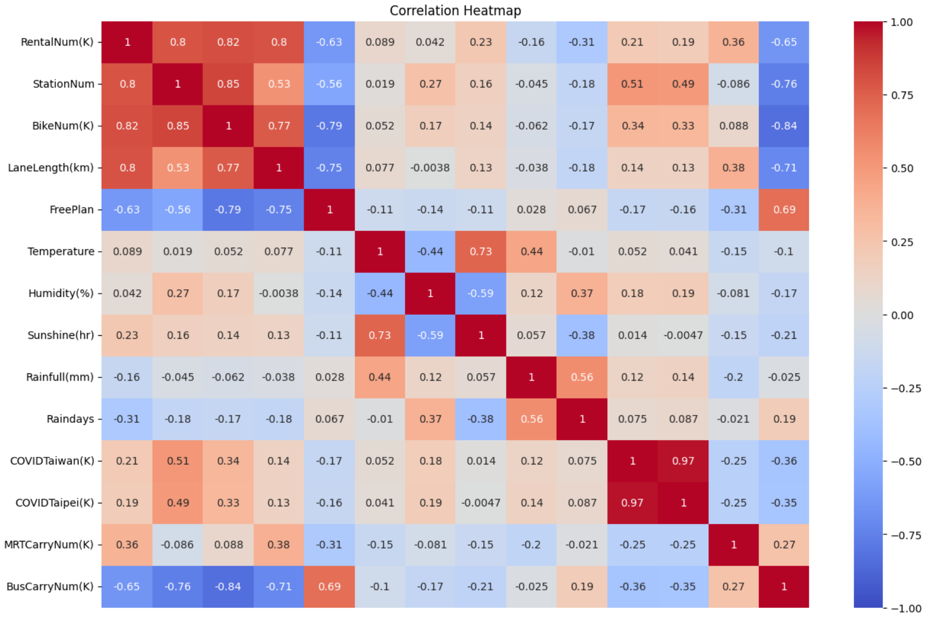
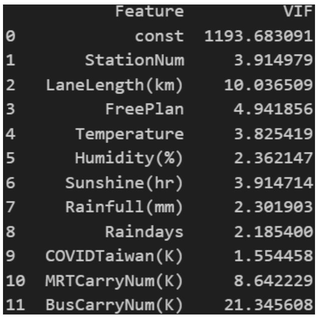


Figure 3: Heatmap of correlation coefficients for all variables



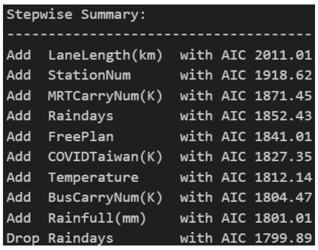
```
              Feature          VIF
0               const  1193.683091
1          StationNum     3.914979
2     LaneLength(km)    10.036509
3            FreePlan     4.941856
4         Temperature     3.825419
5         Humidity(%)     2.362147
6        Sunshine(hr)     3.914714
7        Rainfull(mm)     2.301903
8            Raindays     2.185400
9     COVIDTaiwan(K)     1.554458
10   MRTCarryNum(K)     8.642229
11   BusCarryNum(K)    21.345608
```

Figure 4: The VIFs for the filtered variables except RentalNum(K).

5. Result Analysis

5.1. Result of linear regression

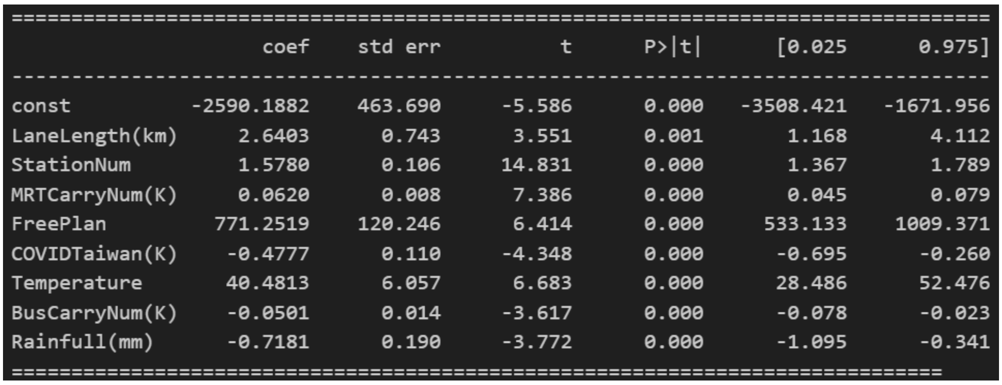
The process of stepwise selection is illustrated in Figure 5, resulting in a model with an AIC value of 1799.89. The outcomes of the model are depicted in Figure 6. The Adjusted R-squared is 0.936. The Root Mean Square Error (RMSE) obtained from predictions on the test dataset is 222.25. Judging from Figure 7, it seems to yield promising results.



```
Stepwise Summary:
------------------------------------
Add  LaneLength(km)  with AIC 2011.01
Add  StationNum      with AIC 1918.62
Add  MRTCarryNum(K)  with AIC 1871.45
Add  Raindays        with AIC 1852.43
Add  FreePlan        with AIC 1841.01
Add  COVIDTaiwan(K)  with AIC 1827.35
Add  Temperature     with AIC 1812.14
Add  BusCarryNum(K)  with AIC 1804.47
Add  Rainfull(mm)    with AIC 1801.01
Drop Raindays        with AIC 1799.89
```

Figure 5: The process of stepwise selection.

Based on the results of the linear regression model, it can be observed that among the bicycle-related variables, an increase in the number of stations, the construction of bike lanes, and the provision of the first 30 minutes for free have a positive impact on the number of rentals, which aligns with intuition. Notably, the model suggests that offering the first 30 minutes for free approximately influences 770,000 rental instances. Among the weather factors, the most significant influences are the average monthly temperature and monthly rainfall, which also align with intuition. The increase in COVID-19 cases negatively impacts bike-sharing rentals, possibly due to



```
==============================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -2590.1882    463.690     -5.586      0.000   -3508.421   -1671.956
LaneLength(km)   2.6403      0.743      3.551      0.001       1.168       4.112
StationNum       1.5780      0.106     14.831      0.000       1.367       1.789
MRTCarryNum(K)   0.0620      0.008      7.386      0.000       0.045       0.079
FreePlan       771.2519    120.246      6.414      0.000     533.133    1009.371
COVIDTaiwan(K)  -0.4777      0.110     -4.348      0.000      -0.695      -0.260
Temperature     40.4813      6.057      6.683      0.000      28.486      52.476
BusCarryNum(K)  -0.0501      0.014     -3.617      0.000      -0.078      -0.023
Rainfull(mm)    -0.7181      0.190     -3.772      0.000      -1.095      -0.341
==========================================================================
```

Figure 6: The result of final linear regression model.

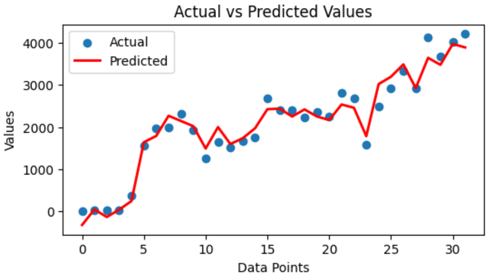


Figure 7: The performance of the final linear regression model on the test dataset.

people going out less during the pandemic or the perceived risk of transmission associated with shared bikes. The higher the MRT ridership, the higher the likelihood of renting a YouBike. With a coefficient of 0.062, it can be inferred that for every 16 MRT passengers, one person uses a YouBike. However, an increase in bus ridership has a negative impact on YouBike rentals. This might be because YouBike stations are usually around MRT stations but not necessarily near bus stops. Since buses and the MRT are competing modes of transport, an increase in bus ridership might reduce the likelihood of YouBike rentals.

5.2. *Result of XGBoost*

The RMSE of the XGBoost model on the test dataset is 180.03, as shown in Figure 8, performing better than the linear regression model but by a small margin. From Figure 9, the XGBoost model considers Rainfall, StationNum, Temperature, MRTCarryNum, and Sunshine as the most important variables influencing the number of rentals, while the others are relatively less important. This differs slightly from the results of the linear regression model but generally aligns.

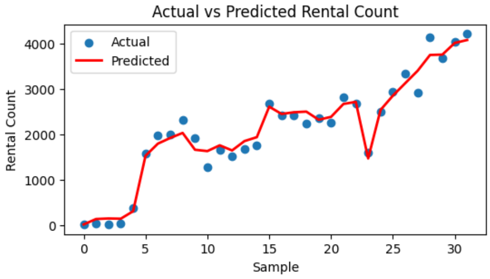


Figure 8: The performance of the final XGBoost model on the test dataset.

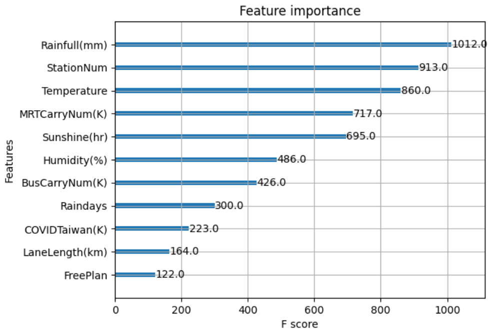


Figure 9: The feature importance of the final XGBoost model.

5.3. *Model Comparison*

Based on the results from the test dataset, although the XGBoost model shows higher accuracy, there is no significant gap. Moreover, the predictive accuracy is not as crucial in this study because the variables used to predict RentalNum are mostly results from the same month. Essentially, knowing the current month's MRTCarryNum, one would typically also know RentalNum. So, the focus lies in which variables affect RentalNum and to what extent. The advantage of the linear regression model lies in its interpretability, as one can examine the coefficients of various variables from the model results to see their impact on the prediction (RentalNum). The higher accuracy of the XGBoost model might stem from its ability to distinguish variables not just from a linear perspective but as intervals. For instance, in the linear regression model, rental frequency gradually accumulates as temperature rises, without discerning situations where the temperature is too low for biking. However, XGBoost can handle this through decision trees.

6. Conclusion

This study identifies the key factors influencing YouBike shared bicycle usage by analyzing variables related to bicycle infrastructure, weather conditions, the impact of the COVID-19 pandemic, and public transportation. The results of the model indicate that improvements in bicycle infrastructure, such as increasing the number of stations and bike lanes, as well as offering a free first 30 minutes scheme, significantly boost rental numbers. Weather conditions, particularly temperature and rainfall, also play a crucial role in influencing cycling behavior. The analysis further highlights the adverse impact of the COVID-19 pandemic on rental frequency, which may be due to reduced mobility and safety concerns during the pandemic. Additionally, there is a significant correlation between MRT ridership and YouBike rentals, suggesting that shared bicycles have been effectively integrated into the public transportation system, serving as an important connection for the last mile.

7. Appendix

7.1. Sources of Data

- Overview of Bicycle Rentals in Urban and Riverside Areas of Taipei by Month: `https://reurl.cc/9vLo3a`
- Overview of Taipei City Bus Passenger Transport by Month: `https://reurl.cc/qV4zv3`
- Overview of Taipei Metro Operations by Month: `https://reurl.cc/OML7rR`
- Taipei City Climate by Month: `https://reurl.cc/bVa4Yl`
- Statistical Table of Region, Age, and Gender - Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): `https://reurl.cc/GjLZpp`