Hourly Rented Bike Demand Prediction in Seoul

Paulina Han(qh2251)

1. Introduction

Under the goal of lowering the carbon footprint, the bike-sharing campaign is well received all over the world, especially in metropolitan cities with large population density and well-developed bike lanes. Bike-sharing is considered an environmentally friendly way of commuting and a solution to the "last-mile" issue in the urban area. The bike-sharing system in Asian cities allows you to rent and return a bike wherever you want at a very low price, within the proper parking area.

However, after the boom of the bike-sharing business in 2017, the bike-sharing industry experienced a bust. The vicious competition between companies has caused an oversupply of bikes which is far beyond the capacity of the city. Therefore, this study aims to estimate the hourly rented bike demand based on environmental parameters. Multiple models were considered, and the final prediction model is selected based on the comparison between the overall performance within the training data.

This study focused on the bike-sharing industry in Seoul, the capital city of South Korea, from Dec 1st,2017 to Nov 30th,2018. The data is downloaded from https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand. The original data contains 8760 observations and 14 variables, including *Date*, *Rented Bike Count*, *Hour, Temperature*, *Humidity*, *Wind speed*, *Visibility*, *Dew point temperature*, *Solar Radiation*, *Rainfall*, *Snowfall*, *Seasons, Holiday*, *Functioning Day*. The *date* variable is converted into a categorical variable *Weekday* indicating the weekdays. Dummy variables are created for all categorical variables: *Seasons*, *Holidays*, *Functioning Day*, and *Weekday*. The tidied data set contains 8760 observations and 20 variables to work with. To evaluate the performance of models, 80% of the data is randomly selected as training data and 20% of the data is randomly selected as testing data with adjustment to balance the distribution of *Rented Bike Count*.

2. Exploratory Data Analysis

To examine the correlation and distribution among all continuous variables, bivariate scatter plots are displayed below the diagonal, the distribution of each variable is shown in the histogram along the diagonal, and the Spearman correlation is demonstrated above diagonal in Figure1. Figure1 demonstrated the hourly rented bike count is most correlated with temperature in the positive direction (0.51). This indicates that the demands for rented bikes increase while the temperature gets higher. Hour also has a notable large positive correlation with the hourly rented bike (0.43). This implies that the days get busier as the day goes by. Dew point temperature, solar radiation, visibility, wind speed is also somehow positively correlated with the hourly rented bike count. Humidity, rainfall, snowfall has a negative correlation with the hourly rented bike count, which indicates that people are less likely to rent a bike when humidity, rainfall, and snowfall increase.

The hourly count of the rented bike is visualized through seasons, weekdays, and holidays in Figure 2. Figure2 (a) shows that in summer the average hourly demand for rented bikes is higher throughout the day compared to other seasons, while winter has the lowest demand. It

also shows that 8:00 and around 17:30 are the peak hours in Seoul when the rented bike has the highest demand. This may be due to the work hours in Seoul, people travel to work at around 8 AM and back home around 17:30. In Figure2 (b), the non-holiday hourly rented bike count has a similar distribution with Figure2 (a). During holidays, the demand for rented bikes is higher in the later time of the day compared to non-holiday. Figure 2 (b) and Figure (c) indicate Monday to Friday have the same average hourly rented bike distribution. 8 in the morning and 17:30 in the afternoon have the highest demand for rented bikes. Weekdays tend to have higher needs for rented bikes compared to weekends.

3. Model

3.1 Model Training

The data is originally time-series data, thus each observation is somehow correlated with each other. To simplify the question, this study assumes all observations to be independent and identically distributed. The *hour* predictor is treated as a continuous variable in all models. The model considered in the study is listed below. During the model training process, the study only uses the training data set.

3.1.1 Linear Regression (LM)

The study first regresses the hourly rented bike count on all 20 predictors using a linear model:

$$Y = \beta_0 + \beta X + \varepsilon \quad (1)$$

17 predictors are selected as significant; however, the adjusted R square is around 0.5, which indicates the linear model is not a good fit for the data.

3.1.2 Elastic net (ENET)

The elastic net model combines ridge penalty along with the lasso penalty. The elastic net coefficients are estimated by minimizing the equation below:

$$RSS + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \quad (4)$$

In R, the tuning parameter is transformed to a mixing parameter α and a penalty parameter λ. The tuning parameter α is the mixing proportion of l_2 penalty versus l_1 penalty, both α, λ is selected by using a 5-time,10-folds GCV grid search algorithm. All predictors are standardized before modeling. In our study $\alpha = 0$, $\lambda = 7.39$, which resembles ridge regression.

3.1.3 Partial least squares regression (PLS)

To reduce the dimension of our predictors included in the model, PLS is employed as a supervised dimension reduction procedure. The tuning parameter M indicates the first M components to be included in the model. M is selected by a 5-time,10-folds GCV grid search algorithm. All predictors are standardized before modeling. The study selected M as 10.

3.1.4 Generalized additive model (GAM)

The GAM model allows for flexible nonlinearities in several variables, but still retains the additive structure of linear models. The model structure of GAM is listed below:

$$g\{E(Y|X)\} = \beta_0 + f_1(X_1) + \cdots + f_p(X_p) \quad (5)$$

The tunning parameters degree of freedom for the nonlinear variables is selected by a 5-time,10-folds GCV grid search algorithm. The GAM model selected 15 predictors with 6 of them as nonlinear terms.

3.1.5 multivariate Adaptive Regression Splines (MARS)

MARS is a non-parametric regression method that automatically models the non-linearity and interactions between variables. There are 2 tuning parameters: degree of features(k) and the number of terms(m). The tuning parameters are selected by a 5-time,10-folds GCV grid search algorithm. The final model has 4 degrees of features and 30 terms.

3.2 Model Selection

To compare the performance of each model, the boxplot of RMSE is shown in Figure 3. The MARS model has the smallest average RMSE (261.65) which is almost half of other models. Given this, the study chose the MARS model to predict the hourly rented bike demand in Seoul using 11 predictors including *function_dayYes, hour, temperature, humidity, seasonSpring, seasonWinter, solar_radiation, dew_point_temperaturet, rainfall, weekdaySaturday, weekdaySunday* .

In Figure 4, the importance of predictors is listed in descending order. *Temperature*, *rainfall*, *hour*, and *functioning_dayYes* are the relatively important predictors in the MARS model. To examine the marginal effects of the predictors, partial dependence plots are shown in Figure 5. Humidity is the only predictor without any interaction term whose relationship between hourly rented bikes is shown in Figure 5. The number of hourly rented bikes remains the same while the humidity increases from 0 to 77. The number of rented bikes starts to decrease when the humidity continues to increase from 77. Other predictors involve an interaction with more than 2 predictors which is difficult to visualize.

In order to verify the performance of the MARS model, the study calculated the RMSE of the MARS model using the test data. The RMSE is approximately 257.63 which is similar to the mean RMSE in the test data. This suggests the MARS model is not overfitting the data and is a good way of predicting the hourly rented bike in Seoul.

From Figure 3, it can be shown that all linear models(elastic net, lm, pls) are not as good as non-linear models. This implies that the relationship between the predictors and the response cannot be best captured by a linear trend. Comparing the GAM model with the MARS model, the MARS model is better than the GAM model. Given MARS can include interaction terms between different predictors, the MARS model is more flexible than the GAM model which may lead to better performance. The limitation of the MARS model is that it is very difficult for people to interpret the result considering the complexity of the model.

4. Conclusion

The MARS model proposed in this study gives the rented bike industry and the government a way of estimating the needs of the customer. Both EDA and the MARS model suggest that temperature, hour, and rainfall are 3 main factors that influence the demand for rented bikes. A moderate temperature increases the demand for rented bikes. The rush hour is when customers need the bike most. People are less likely to rent a bike on rainy days. Such findings align with our daily experience. In the future study, the independency assumption could be taken care of using a more sophisticated model. To make the result more useful for policymaking, the prediction model could be constructed on the district level.

Figure 1. Correlation Plot

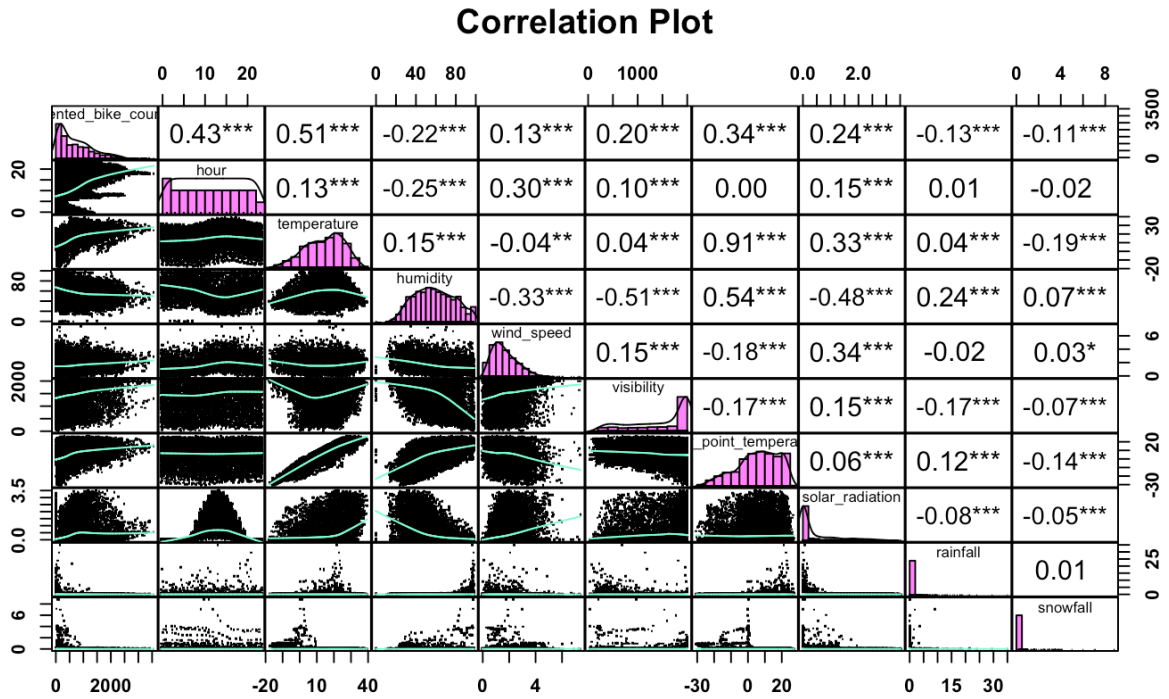

Figure 2. Discrete Variable Visualization

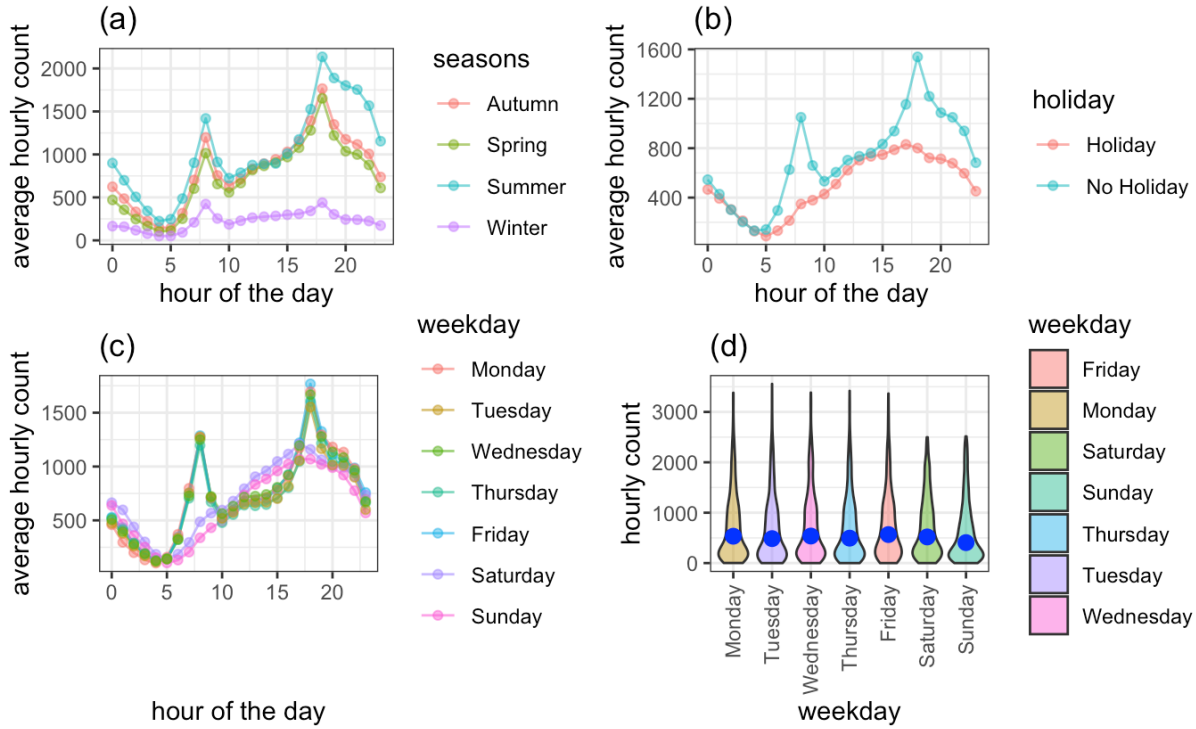

Figure 3. Model Comparison

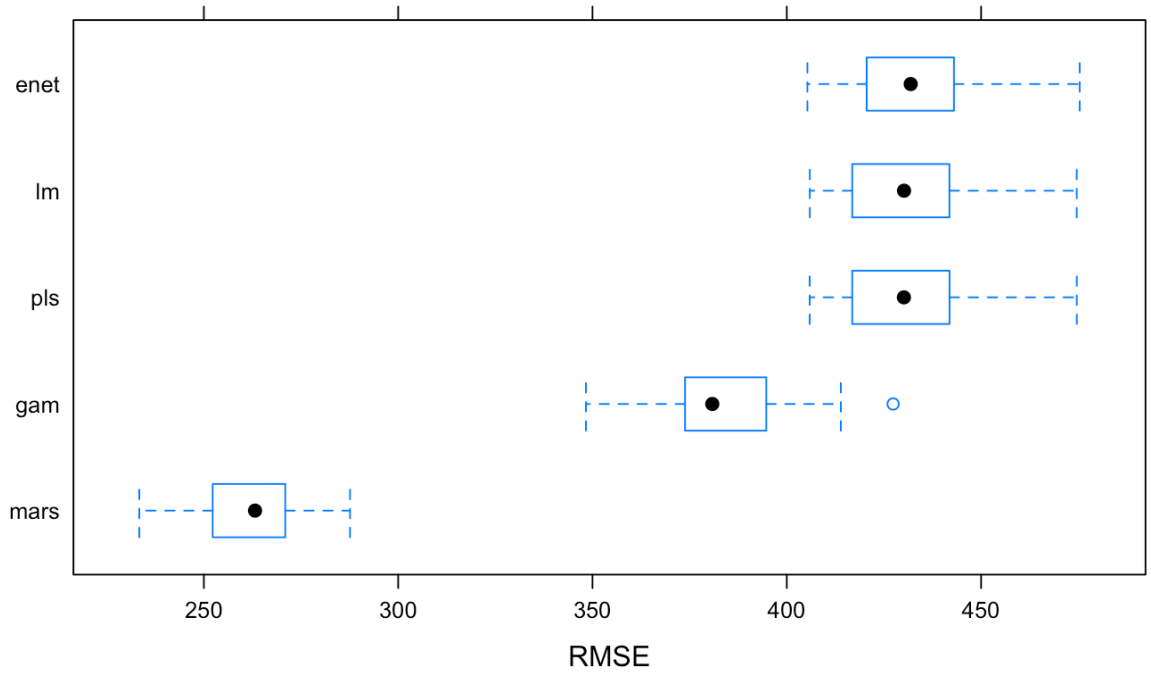

Figure 4 . Importance of Predictors

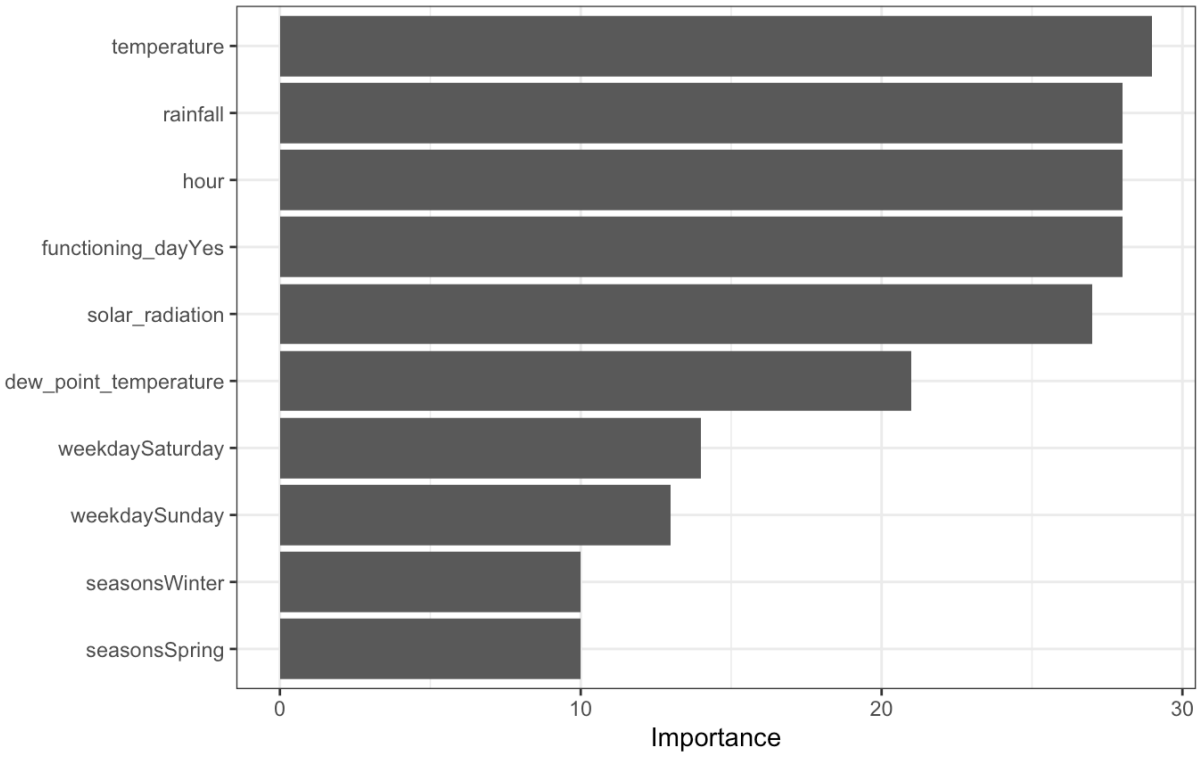

Figure 5. Partial Plot