Forecast of Demand for Sharing Bikes in Seoul

Wanxin Qi (wq2161), Lesi He (lh3109), Ke Xu (kx2173)

Introduction

In recent years, bike-sharing has become a popular way of transportation in several countries. It is more environmentally friendly, more convenient, and less costly, especially in cities with congested roads. For example, the Citi Bike in New York, the Blue Bikes in Boston, the Capital Bikeshare in Washington, etc. [1] Until today, bike-sharing programs in China have experienced ups and downs due to improper inputs and inadequate research. After a series of adjustments, the current bike-sharing network has reached a balance between public availability and system management. What people learned from China's lesson is that the number of bicycles required in different situations varies, different places, times, and external factors, and models to accurately predict the number of bicycles to be distributed are in need. Thus, using the dataset of Seoul Bike Sharing Demand from the Seoul Public Data Park website, the object of the research is to obtain a model that could predict the hourly rented bike counts by several predictors about weather information and date information to explore the demand for bicycles in different situations to provide a stable supply. [2, 3]

The original dataset contains 14 variables and 8760 observations with no missing value. It recorded the hourly count of bikes rented from 12/01/2017 to 11/30/2018. First, the date variable was transferred to weekdays, i.e., Monday through Sunday, because the dataset has the predictor season to provide rough information about months, and the relationship between rented bike counts and each day of the week is more interesting to explore. The variables were converted to appropriate types respectively. Thus, there were 4 categorical variables, weekdays (Monday to Sunday), seasons (Winter, Spring, Summer, Autumn), holiday (Yes, No), and functional (the rental bike system is operating) day (Yes, No), and 9 numeric variables, the hour of the day, temperature (°C), humidity (%), wind speed (m/s), visibility (10m), the temperature of dew point (°C), solar radiation (MJ/m2), rainfall (mm), and snowfall (cm). 1000 observations were randomly sampled from the original dataset to be the dataset of the research, and they were partitioned into the training dataset with 800 observations and the testing dataset with 200 observations. After converting the categorical variables to dummy variables, there are 20 predictors in total.

Exploratory analysis/visualization

The plot of average hourly rented bike counts by the hour of the day across seasons (Fig. 1) shows that the distribution of average hourly rented bike counts has two peaks at 8 to 9 am and 6 to 7 pm and changes with the hour. Summer has the most average hourly rented bike counts, winter is significantly lower than other seasons, and the distribution of average hourly rented bike counts of each season has the same trend. By the distribution of hourly rented bike counts across the functional day and the exploration of the original dataset, all the 0 hourly rented bike counts appeared on non-functional days. Thus, there is no zero-inflation problem, and these observations and the functional day variable were kept.

By the correlation plot of all the continuous variables (Fig. 2), temperature is highly positively correlated with bike counts (0.54), and it is the only predictor that appears to have some potential

linear relationship with bike counts. The hour of the day and dew point temperature are moderately positively correlated with bike counts (0.41 and 0.38). Other correlations are weaker but still informative to building models. Among them, humidity, rainfall, and snowfall are negatively correlated with bike counts. A few predictors are somewhat correlated. The correlation between temperature and dew point temperature is up to 0.92, and the correlation of humidity with visibility and dew point temperature are -0.54 and 0.56, respectively.

Models

Linear models (LS, LASSO, and PLS), nonlinear model (MARS), and ensemble method (GBM) were applied to predict the rented bike count. The 20 predictors, hour of the day, temperature, humidity, wind speed, visibility, dew point temperature, solar radiation, rainfall, snowfall, the six dummy variables of weekdays, the three dummy variables of seasons, the dummy variable of holiday, and the dummy variable of functional day were all included.

Least Squares

The linear model using the least-squares method contains several assumptions, which are the linear relationship between the outcome variable and the independent predictor variable, no multicollinearity between predictors, and the independence, the homoscedasticity, and the normality of residuals. In the exploratory analysis, only a few variables can be seen to have a linear relationship with rented bike count, and temperature and dew point temperature are highly correlated, which violates the assumptions. This becomes a limitation of the least-squares model. By the summary of the model, with a threshold of p-value < 0.05, the hour of the day, temperature, humidity, radiation, rainfall, all the dummy variables of season, weekdays (Tuesday and Thursday), and functional day are statistically significant. The test RMSE is 396.6971.

Least Absolute Shrinkage and Selection Operator

To deal with the multicollinearity problem in the linear regression, the regularization method, least absolute shrinkage and selection operator (LASSO), was applied. LASSO requires the predictors to be scaled (to have sample variance 1) and centered (to have sample mean 0). By slightly increasing the bias and reducing the variance, it shrinks the coefficient estimates towards zero, thus excluding the variables having little effect on the outcome. Therefore, it contains a tuning parameter λ which controls the flexibility. By setting a grid of λ values and computing the cross-validation error for each value of λ, the best λ value 2.359821 was picked with the smallest cross-validation error. Based on the result, it excluded visibility, dew point temperature, and weekdays (Wednesday), which are not significant in the linear regression. The test RMSE is 396.3537. The limitation is that it may randomly select a variable among highly correlated variables, e.g., the temperature and the dew point temperature in this case, and there is no analytic solution.

Partial Least Squares

The dimension reduction method, partial least squares (PLS), was applied to deal with the multicollinearity problem as well. PLS also needs the predictor variable to be scaled and centered. It uses a small number of linear combinations of the original variables to fit linear regression. The new features identified not only maximally summarize the variability of predictors but also maximally correlate to the outcome. The number of new features is thus a tuning parameter. Based on the input range of component number and cross-validation, the optimal number of components

10 was picked with the smallest cross-validation error. The model explains 57.46% of the outcome, and the test RMSE is 395.9891. Overfitting may be a limitation of PLS.

Multivariate Adaptive Regression Splines

Since a few predictors have a linear relationship with the response variable, a nonlinear model, multivariate adaptive regression splines (MARS), was built. It is more flexible to capture the underlying truth than linear regression with no assumptions. It has two tuning parameters which are the degree of interactions and the number of retained terms. The best tuning parameters were picked by a grid search that identifies the optimal combination with the smallest cross-validation error, in which the degree of interactions is 5 and the number of retained terms is 26. By visualizing the variable importance (Fig. 3), 10 variables were included in the final model, and the temperature seems to be the most important variable. The test RMSE is 306.2679. The limitations are that it is more difficult to interpret than linear models, and it highly relies on the quality of the data.

Gradient Boosting Machine

To get a more accurate prediction than single models, a gradient boosting machine (GBM), as an ensemble method, was applied. This is a sequential technique such that a new decision tree is an improvement of previously grown trees, and it has no assumptions. There are several tuning parameters, including the number of trees, the number of splits in each tree, and the shrinkage parameter (a small positive number that controls the rate at which boosting learns). The optimal combination of tuning parameters was obtained by adjusting the tuning grid with the minimum cross-validation RMSE. As a result, the best number of trees is 584, the best number of splits in each tree is 14, and the best shrinkage parameter is 0.05. The variable importance is demonstrated by visualization (Fig. 4). Temperature and the hour of the day have significantly high importance among all the variables, indicating that they have the largest overall impact, while weekdays, holiday, and seasons show little importance. The test RMSE is 235.5603. One reason to be considered for this large RMSE is that the outcome variable is in fact a count, while it was treated as continuous. The limitations are the risk of overfitting and computationally expensive. [4] As a much more flexible method, it would produce the optimal solution based on all the data including outliers which may cause overfitting. The several tuning parameters require a wider grid of search, and the tuning process takes a longer time than statistical modeling.

By 10-fold cross-validation and resampling, with the minimum mean and median RMSE (Fig. 5), the gradient boosting machine was chosen as the final model. As a black-box model, to further quantify global feature importance, partial dependence plots (PDPs) and individual conditional expectation (ICE) curves were generated.

Since temperature and the hour of the day have the largest overall impact, their marginal effects on the change in the average predicted outcome was investigated by PDPs, as shown in the top two plots of Figure 6. By the left plot, when all the other variables are controlled at their mean, the average predicted rented bike count changes over time. It shows a clear positive relationship from 4 am to 8 am and 10 am to 6 pm, where 8 am and 6 pm are the two peaks. The right plot shows the association of the average predicted rented bike count with temperature holding other variables at their mean. The relationship is positive when the temperature is lower than around 28 °C and negative when higher. The bottom two plots of Figure 6 show the influence of the interaction of the hour of the day and rainfall and the interaction of humidity and temperature on predicted rented

bike count when all the other variables are fixed at their mean. Based on the left plot, the change of the average predicted rented bike count over time appears the same as the PDP of the hour of the day, while the difference is that the rented bike count is significantly higher for 0 rainfall than for positive rainfall. However, the rainfall doesn't seem to influence the predicted outcome, as the plot doesn't show a vertical change. Based on the right plot, for temperatures at around 20 °C to 35 °C with humidity less than 75%, the model predicts on average a high demand for bikes. There is a weak relationship that the predicted rented bike count increases as the temperature increases up to 28 °C and as humidity decreases.

To uncover the heterogeneous effects of the hour of the day and temperature on the predicted rented bike count by fixing each observation of all the other variables, the ICE curves were plotted (Fig. 7) and they were centered to remove level effects. Based on the left plot, compared to the PDP (Fig. 6 upper left), the same conclusion can be drawn since all the single curves seem to be parallel and have the same trend. Similarly, the ICE curves of temperature (Fig. 7 right) show that most instance lines have the same trend as the average line, with only a few lines showing a decreasing trend for the temperature at 10 to 20 °C. Besides, when the temperature is low (e.g., 5 °C and lower), the variance of the predicted rented bike count decreases and the predicted rented bike count is close to 0, indicating that the demand for bikes is very low when it is cold, despite other conditions. The variances of predicted rented bike count increase as the temperature increases, indicating potential interactions with other variables, which need further exploration.

Conclusion

By utilizing least squares, LASSO, PLS, MARS, and GBM to predict the demand for sharing bikes in Seoul using weather and date information, the gradient boosting machine performed the best with the lowest cross-validation RMSE. Important variables of the prediction model include the hour of the day and temperature with significantly higher importance than other variables, followed by humidity, solar radiation, and functional day. The partial dependence plots and individual conditional expectation curves further reveal the impact of the hour of the day and temperature on the response variable and uncover insights into the influences of rental bike count. The reason that functional day has such importance to the response variable is that all the 0 hourly rented bike count was due to non-operating renting systems. By the order of importance of the predictors, weather information is more important than date information, which is as expected. Instead of the established date information, bike-sharing programs should pay more attention to the weather forecast to distribute the appropriate number of bikes in the right locations. The next step of the research could be narrowing geographically to get a more accurate prediction model for different places and to take more related factors into the research such as user feedback.

References

1. Brian Martucci. “6 Best Bike-Share Programs in the U.S. & Canada.” Money Crashers, 14 September, 2021
2. Sathishkumar V E, Jangwoo Park, and Yongyun Cho. “Using data mining techniques for bike sharing demand prediction in metropolitan city.” Computer Communications, Vol.153, pp.353-366, March, 2020
3. Sathishkumar V E and Yongyun Cho. “A rule-based model for Seoul Bike sharing demand prediction using weather data” European Journal of Remote Sensing, pp. 1-18, February, 2020
4. Vihar Kurama. “Gradient Boosting In Classification: Not a Black Box Anymore!” PaperspaceBlog, 2020

Figures and Tables

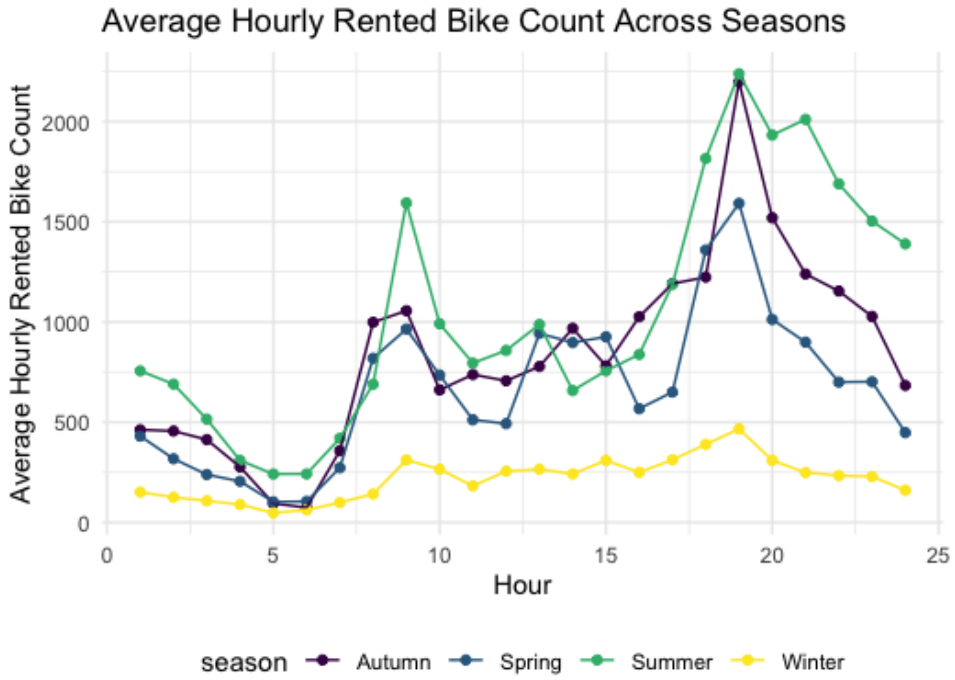


Fig. 1. Average Hourly Rented Bike Counts by the Hour of the Day across Seasons.

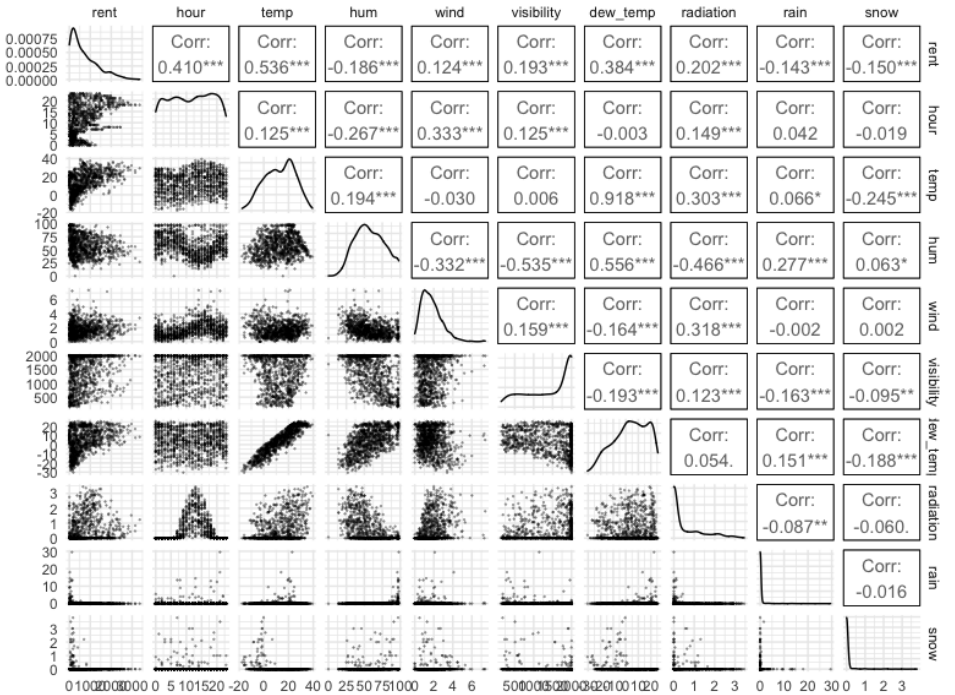


Fig. 2. The Correlation Matrix and Scatterplots of All the Continuous Variables.

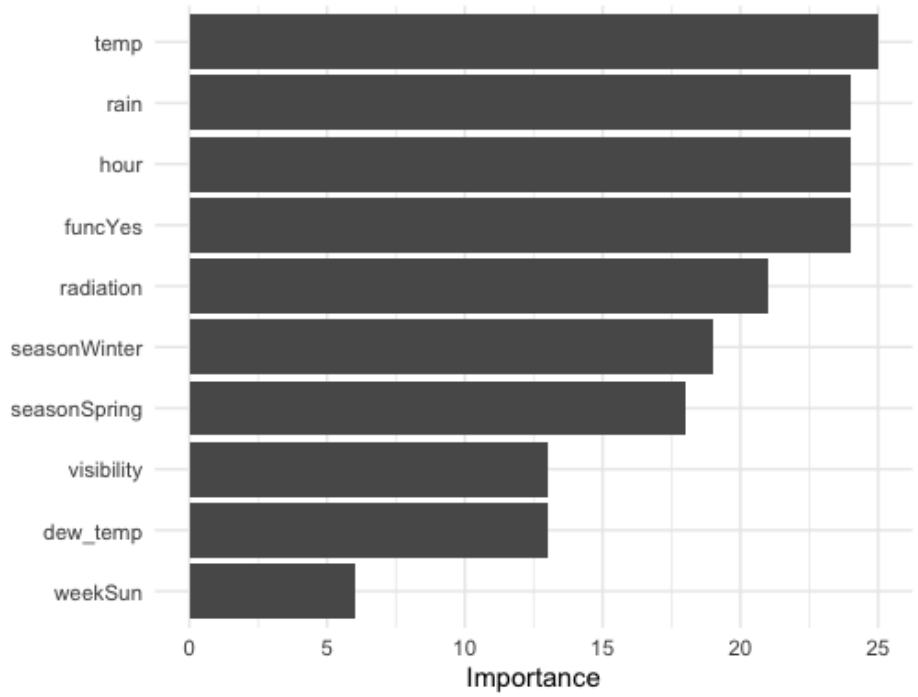


Fig. 3. The Variable Importance Plot of the MARS Model.

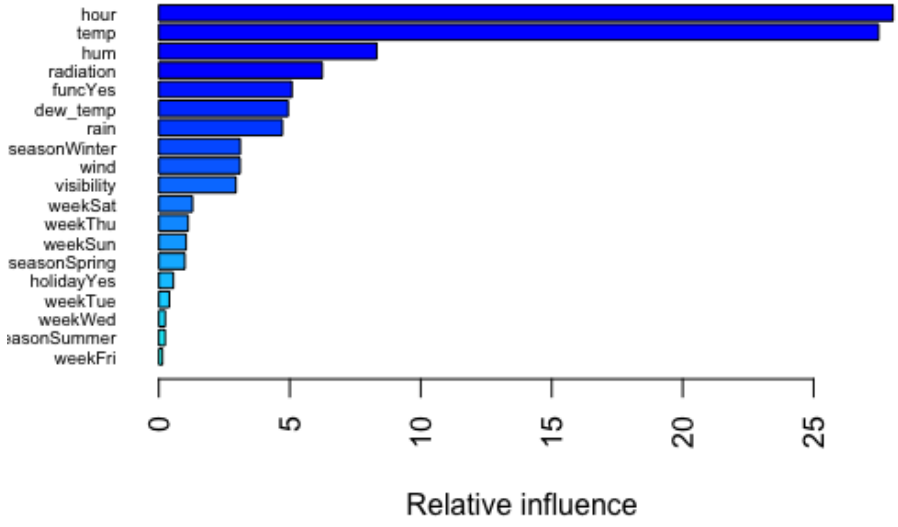


Fig. 4. The Variable Importance Plot of the GBM Model.

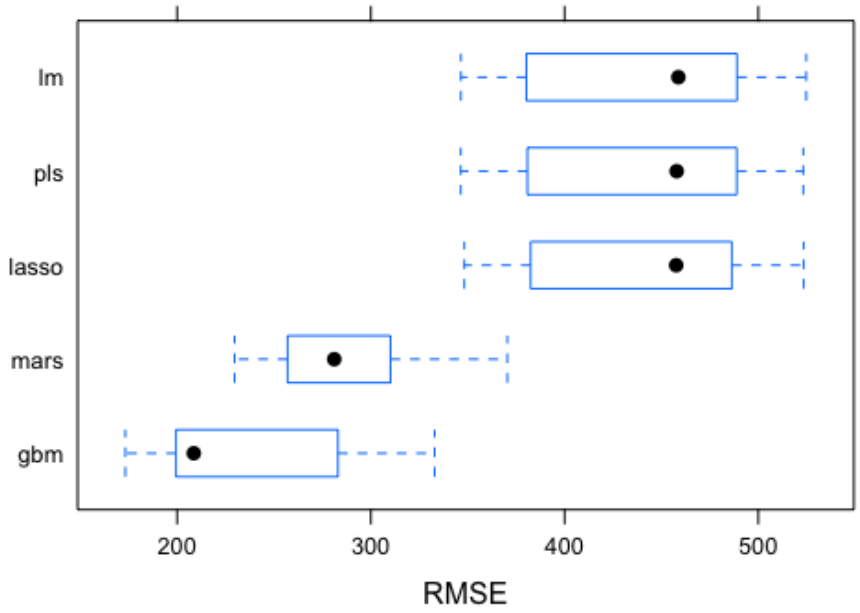


Fig. 5. The Distribution of RMSE by Resampling Each Model.

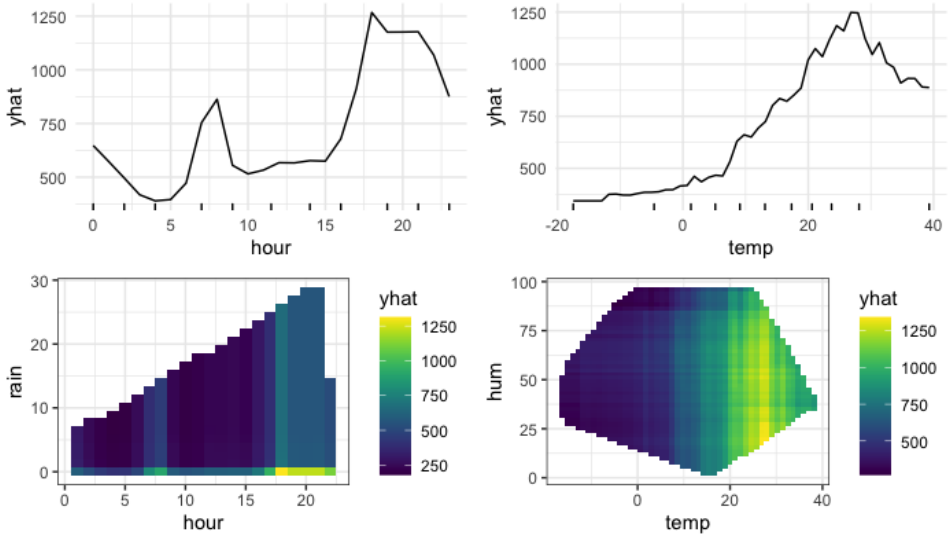


Fig. 6. Partial Dependence Plots of the Hour of the Day (Upper Left), Temperature (Upper Right), the Interaction of the Hour of the Day and Rainfall (Bottom Left), the Interaction of Temperature and Humidity (Bottom Right).

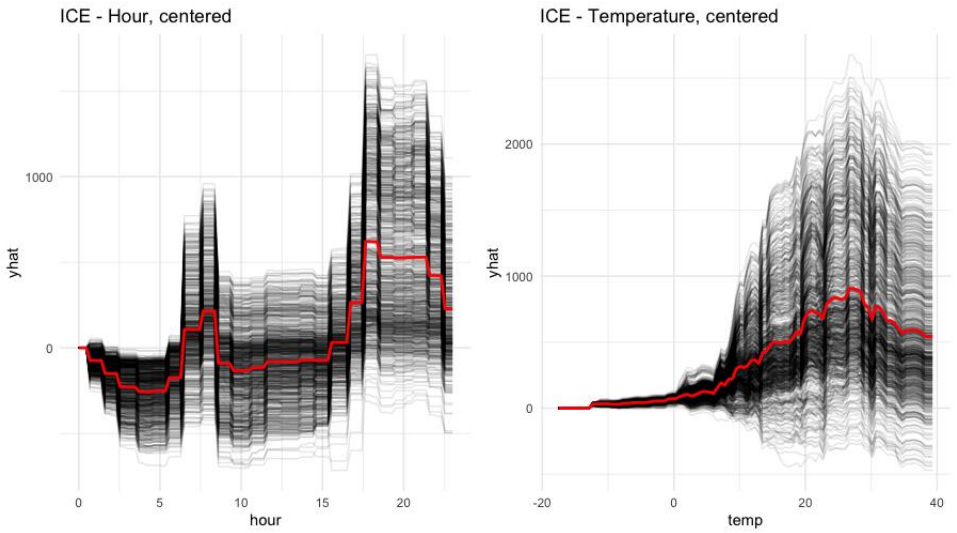
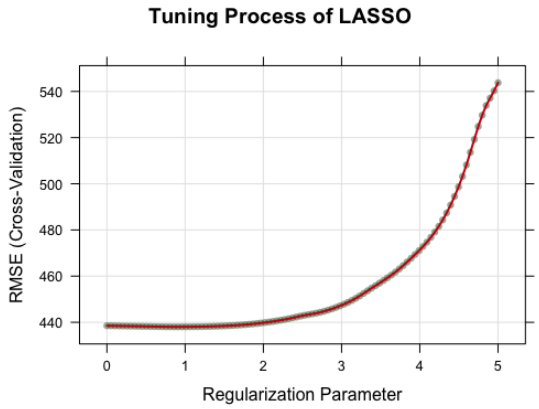
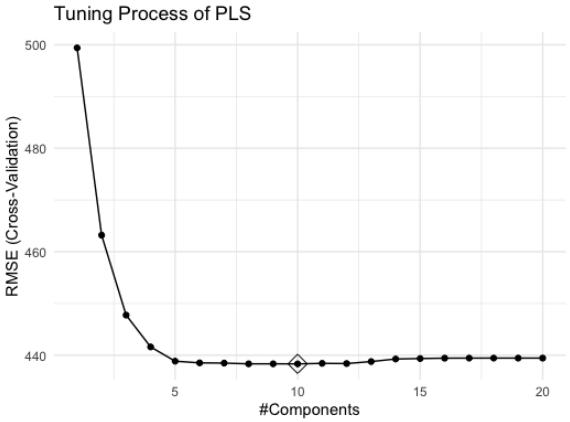


Fig. 7. Centered Individual Conditional Expectation Curves of the Hour of the Day (Left) and Temperature (Right).

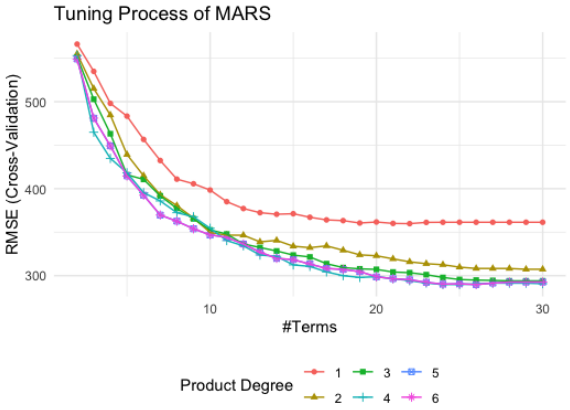
Appendix



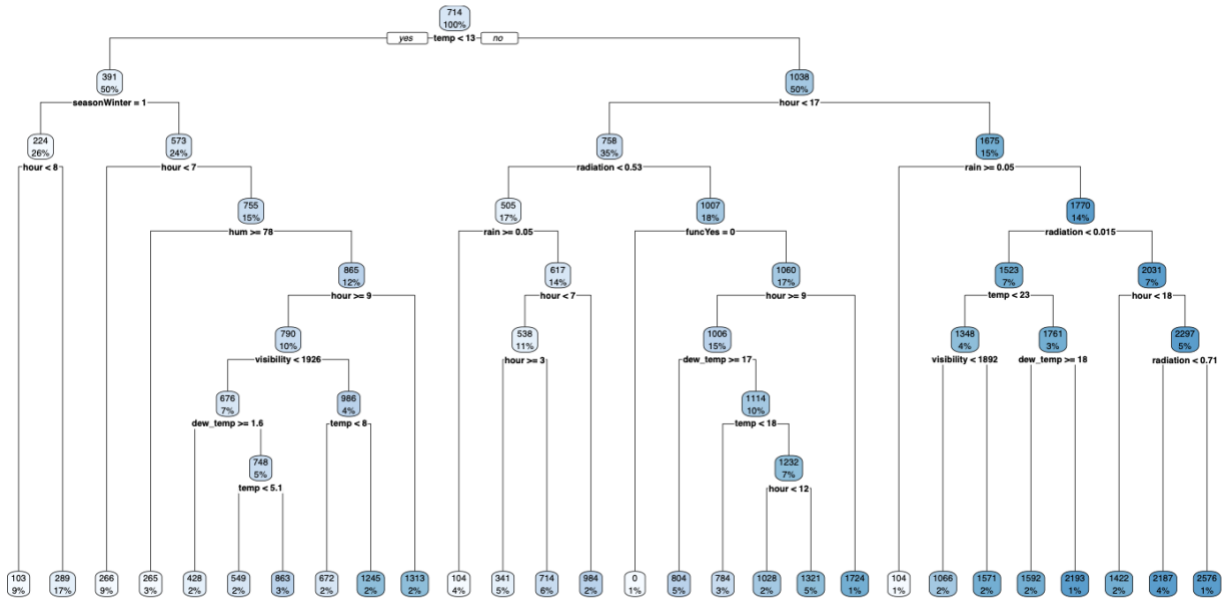
Tuning Process of LASSO Model



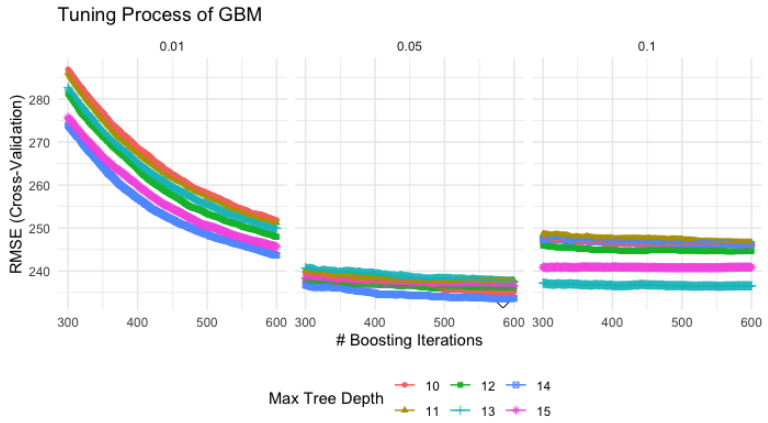
Tuning Process of PLS Model



Tuning Process of MARS Model



The Plot of Regression Tree



Tuning Process of GBM Model