STATISTICS 452/652: Statistical Learning and Prediction

Assignment 2

Refer to the Air Quality data described previously, and the analyses we have done with `Ozone` as the response variable, and the five explanatory variables (including the two engineered features from Assignment 1 as below).

- `AQ$TWcp = AQ$Temp*AQ$Wind`
- `AQ$TWrat = AQ$Temp/AQ$Wind`

1. Fit a default Random Forest (RF) to *only the three main variables in the data*—`Temp`, `Wind`, and `Solar.R`—and not the two extra ones that we engineered. A RF should be able to detect interactions automatically if needed.
   (a) **Report the OOB error.**
   (b) **Produce variable importance measures and comment on the relative importance of the variables. How do they compare to what we have seen in earlier analyses of these data?**
2. Repeat the exercise in question 1, adding the two engineered features into the data.
   (a) **Report the OOB error. Does it improve much compared to the previous RF analysis without the variables?**
   (b) **Produce variable importance measures. Are the two engineered features particularly important?**
3. Use boosting to model the relationship between `Ozone` and all ONLY THE THREE ORIGINAL VARIABLES. Tune on an initial grid of $\lambda = 0.001, 0.005, 0.025, 0.125$ and $d = 2, 4, 6$, and select trees optimally using twice the number suggested by OOB error. Use two reps of 5-fold CV (refer to the lecture note and R code to understand how to do this).
   (a) **Report the mean root-MSPE for each combination of** λ and d
   (b) **Show relative root-MSPE boxplots for each combination of** λ and d
   (c) **What combination of** λ and d **do you prefer?**