Tasks chapter8: Random forests – predicting ozone concentration

David Kurz, Malte Hildebrandt
2023S707737 VU Geostatistik
Universität Innsbruck

June 20, 2023

We move on from a single tree to an ensemble of trees - a forest. The dataset remains the same and also the package partykit to fit the forest to the data.

1. Data preparation

   - The ozone measurements are in the package datasets and can be loaded with data("airquality").
   - Convert Month and Day to day of year with as.POSIXlt(paste(1973, airquality$Month, airquality$Day), format = "%Y %m %d")$yday
   - Convert the units to SI: Langley to Joule per square meter, miles per hour to meters per second, and Fahrenheit to Celsius.
   - Subset to non-missing ozone data (other variables can be missing) and drop Month and Day, using subset() and assign to airq. (Dropping works with the option select = -c(Month, Day)).

   After data preparation our airq data frame consits of 116 observations of 5 variables. These are: Ozone with datatype num, Solar.R with datatype int, Wind with datatype num, Temp with datatype num and DayOfYear with datatype int.

2. Fit forest

   - Set a seed of 2908 to make the results reproducible
   - Fit a conditional random forest consisting of ntrees = 100 trees with partykit::cforest(formula, data = ...) using all variables and assign to airQcforest. For the usual formula syntax, remember the abbreviation ".", which uses all variables in the data set except for the response variable to the left of the "∼" symbol. (In case you're wondering why the package is explicitly specified: it makes absolutely sure that you don't get cforest from the party package.)

3. Examine and plot the resulting forest model

"Random forests are a widely used ensemble learning method for classification or regression tasks. However, they are typically used as a black box prediction method that offers only little insight into their inner workings. . . . the stablelearner package can be used to gain insight into this black box by visualizing and summarizing the variable and cutpoint selection of the trees within a random forest." (Quote from the vignette https://cran.rproject.org/web/packages/stablelearner/vignettes/forests.html). Reading the vignette will help you with the interpretation asked of you in the following parts.

- Load the stabletree package.
- Convert the conditional random forest model in airQcforest to a stabletree object with as.stabletree(airQcforest) and assign to airQcforest_st.
- Do a summary of airQcforest_st and interpret.

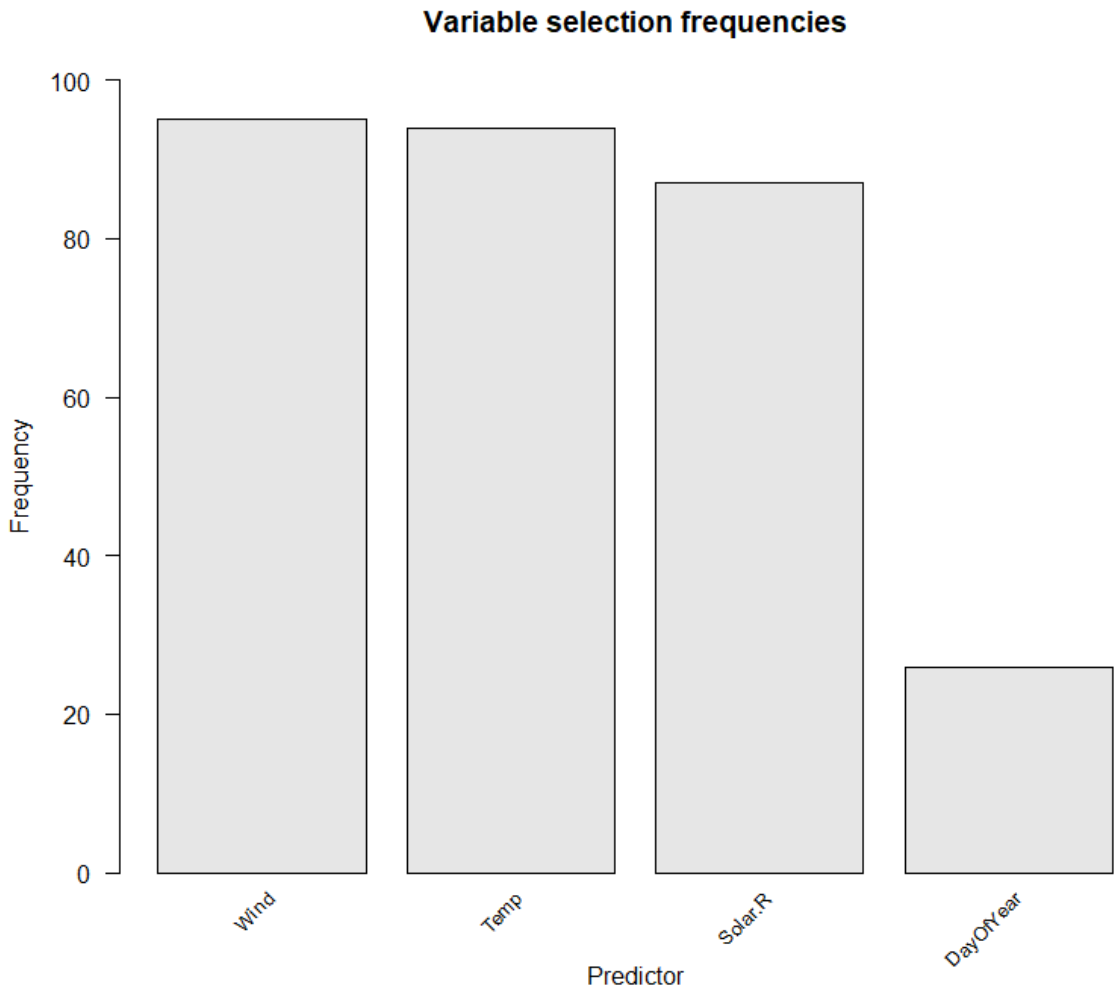
  Call:
  partykit::cforest(formula = Ozone ~ ., data = airq, ntree = 100)

  Sampler:
  B = 100
  Method = partykit::cforest

Variable selection overview

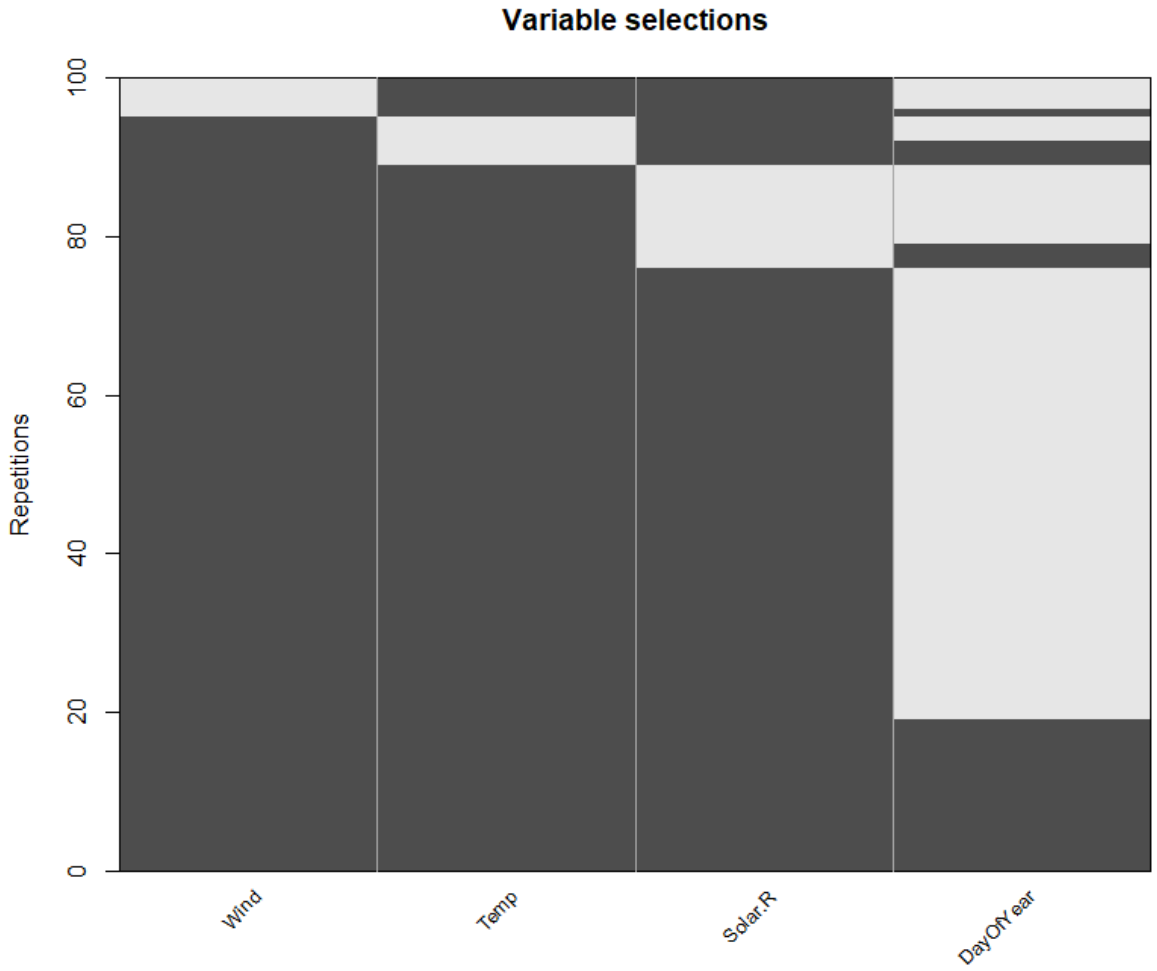
	frequency	mean
wind	0.95	1.77
temp.	0.94	1.82
rad.	0.87	1.23
day	0.26	0.26

- Use a barplot() to show how frequently the predictors were selected in each of the ntrees and interpret.



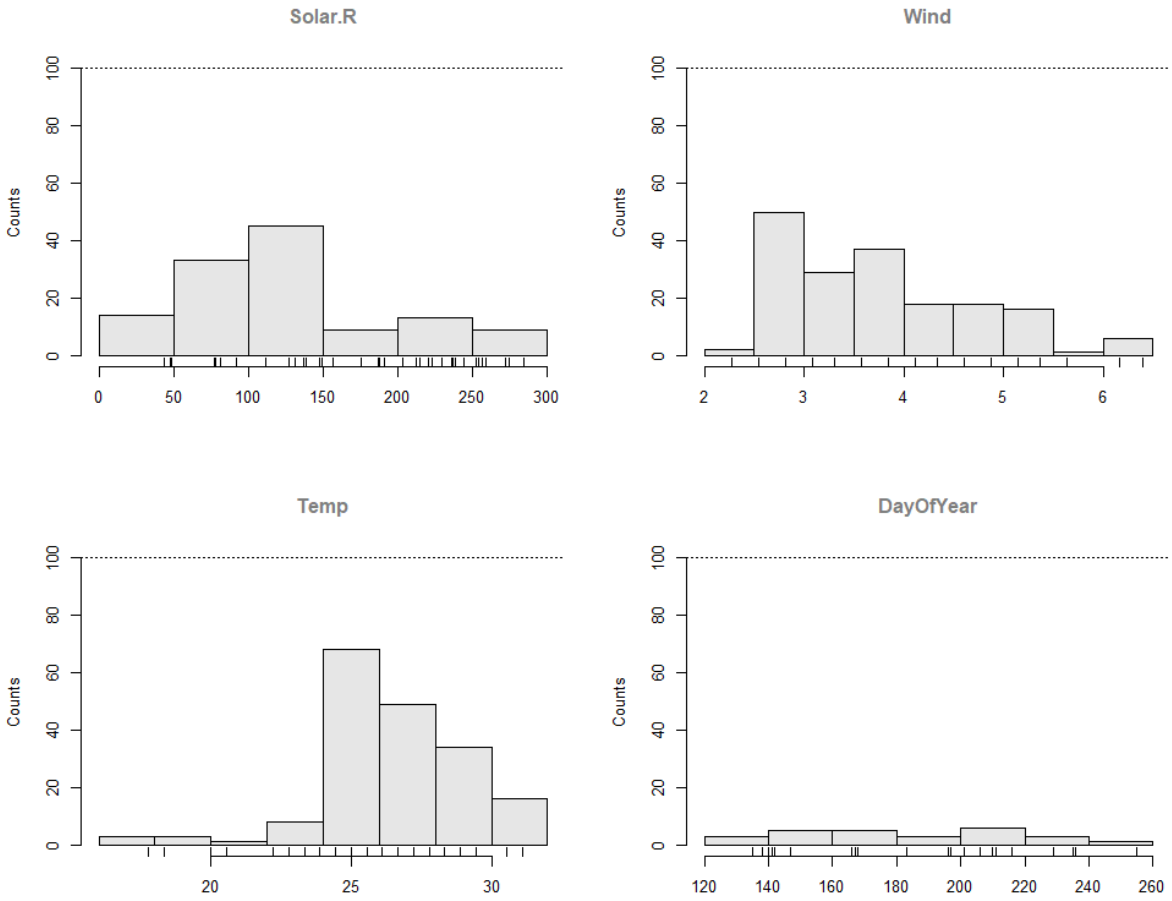
Random forests provide a measure of feature importance, depicted here in the frequency of each predictor. We can derive how much each feature contributes to the model's predictive accuracy based on these frequency scores. They can be used to understand the relative influence of each predictor on the model's predictions. Wind and temperature were frequently selected, indicating that the prediction of ozone concentration is mainly based on or influenced by these features. Less important is the incoming solar radiation, and least important is the day of the year.

- Use image() to show which variables were selected when a particular variable was NOT selected and interpret.



The graph represents the frequency of plots on the y-axis, and the predictors are listed. The black shaded area indicates when a predictor was selected to predict ozone, while the grey shaded area represents the opposite. We can observe the different variable or feature selections for each predictor. The wind predictor does not select a few variables, whereas the temperature predictor selects the exact same variables that are "missing" from wind. Additionally, the solar radiation predictor selects the exact same number of variables that are not selected by temperature and wind combined. This suggests that these first three predictors effectively cover the prediction of ozone by utilizing variables that the others do not include. On the other hand, the day of the year predictor does not show a clear and useful connection to the other predictors. It only uses a few variables from the other predictors. Overall, the day of the year is not a good predictor for ozone because it does not take into account the weather situation like wind, temperature, and solar radiation do.

- Use plot() to inspect the cutpoints and resulting partitions for each variable over all ntrees. Which predictor variables are well suited for random forests and for which ones might a simple (generalized) linear regression have been fine, too?



Solar radiation exhibits a wide range of variable counts for predicting ozone when it is below 150 W/m^2. The small stripes above the x-axis indicate the cutting points, which display a non-linear and somewhat random behavior. This makes solar radiation a suitable candidate for a random forest model. Similarly, the cutting points in the day of the year plot also exhibit random behavior, indicating its potential suitability for a random forest approach. However, when considering the very low counts of variables selected for ozone prediction, this predictor may not be particularly effective.

On the other hand, wind and temperature display a consistent frequency of cutting points within their specific range of values. As a result, a simple linear regression model would also be suitable for predicting ozone. The temperature predictor demonstrates high variable selection counts when the temperature is around 25°C, approximately in the range of tens. Similarly, for wind as a predictor, this occurs at wind speeds between 2.5 to 3 m/s.