Forests

We analyse the browsing and online purchasing behaviour of households using Comscore's web browser data. The data file `browser_2006.csv` contains 8,000 households that spent at least 1 US-dollar online in 2006. The variable `spend` is the online spending (in US-dollars) of a household. Furthermore, the data contains the browser history of households for the 1,000 most heavily trafficked websites (see the list of websites in `browser-sites.txt`). In particular, the data contains the percentage of time spent on specific websites from the total time spent online. Table 1 shows a short overview of the `browser_2006.csv` dataset. Additionally, we have access to the file `browser_new.csv`, which contains the browser history of 2,000 new households, but not the online spending.

Table 1: Overview of the Structure of the `browser_2006.csv` Dataset

	id	spend	atdmt.com	yahoo.com	whenu.com
1	1297	177	10.88	14.97	0.00
2	537	3396	8.00	10.69	10.07
3	6422	1895	5.12	1.47	0.00
4	3399	682	10.38	6.60	0.00
5	5292	610	11.73	18.62	0.29

Group Home Assignment (max. 3 points)

The mandatory group home assignment has to be submitted before 12:00 o'clock prior PC-session 4. It is obligatory to solve the assignment in R Markdown with `echo = TRUE` option for every code chunk and generate a PDF file with the solution. Generating an HTML file from the R Markdown and converting it into the PDF is also possible. Make sure that the final PDF file has no readability issues. The PDF with the answers to the four questions below as well as the file with the R Markdown code has to be submitted via Canvas.
Download the data sets `browser_2006.csv` and `browser_new.csv` from Canvas. Load the data into R. Generate matrices for the outcome, control, as well as id variables for the 2006 and the new data. Install and load the packages `grf`, `DiagrammeR`, and `glmnet`.

1. How much is the average online spending in 2006? (0.5 points)

2. On which webpage is the household with id = 1297 (first row of the 2006 sample) most of the time? (0.5 points)

3. Which two webpages are together the best linear predictors for online spendings in 2006? You can use a Lasso with only two active control variables to answer the question. (1 point)

4. Estimate a post-Lasso model on your solution to the previous exercise. Use this model and a second OLS model that includes all 1000 websites as variables to predict the spendings for the households in the file `browser_new.csv`. Plot the predictions against each other and calculate the correlation of your predictions. Do you think Lasso has selected a reasonable set of variables? (1 point)
   If you have not identified an optimal pair of webpages in the previous exercise, choose any two for this exercise.

The exercise on the next page will be solved during the PC session. It is not part of the group home assignment.

Exercise: Online Spendings

1. Generate a variable for log online spendings. Plot the cumulative distribution of online spendings and log online spendings.

2. Randomly partition the 2006 data into a training and test sample of equal size. For this purpose, generate a variable that indicates the rows that are included in the training sample (using the `sample` command).

3. Build a random forest to estimate online spending in the training sample. The forest should contain 1000 trees. Each tree should use a 50% subsample of the training data, square root of the covariates, and restrict the `min.node.size` to 500.

   (a) Plot a tree of the forest.

   (b) Print the variable importance. Why do we have to be cautious when interpreting the variable importance?

   (c) Use the forest to predict the online spendings in the test sample. Evaluate the performance of the random forest using the R^2.

4. Draw a graph of out-of-sample R^2 with regard to the number of trees in the forest.

5. Build different forests and try to find the specification which minimizes the R^2 in the test sample. Try the following model specifications:

   (a) Set the `min.node.size` such that the forest can use deep trees. Plot one of the deep trees.

   (b) Use the log online spending instead of the level of online spending as outcome variable. Are the same variables important?

   (c) Use the specification for `honest` trees.

6. Use the data `browser_new.csv`, which contains the browsing behaviour of new potential customers. Predict the online spending in the new data using the forest model that performs best in the test sample. These predictions might help you to target marketing campaigns at the new potential customers with the highest (or lowest) expected online spending.

7. Save the id's and the predicted spendings of the new customers in a csv-file.

Useful links:

- A description of the `grf` package is here: `https://cran.r-project.org/web/packages/grf/grf.pdf`.