Trees and Forests

We analyse the browsing and online purchasing behaviour of households using Comscore's web browser data. The data file `browser_2006.csv` contains 1,500 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. Additionally, we have access to the file `browser_new.csv`, which contains the browser history of 500 new households, but not the online spending.

Download the data sets `browser_2006.csv` and `browser_new.csv` from Github. Load the data into R. Install and load the packages `grf`, `rpart`, `rpart.plot`, `DiagrammeR`, and `glmnet`.

Exercise 1: Data Description and Preparation

1. Generate the ID, outcome and control variables in matrix format.
2. How high is the average online spending in 2006?
3. On which webpage is the household with id = 921 most of the time?
4. Generate a variable for log online spendings. Plot the cumulative distribution of online spendings and log online spendings.
5. 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).

Exercise 2: Trees

1. Build in the training sample a shallow tree (terminal leaves should contain at least 100 observations) with the outcome log online spendings. Plot the structure of the shallow tree.
2. Build in the training sample a deep tree (terminal leaves should contain at least 10 observations) with the outcome log online spendings. Plot the structure of the shallow tree. Plot the cross-validated MSE.
3. Determine the optimal number of terminal leaves.
4. Prune the deep tree and plot the structure of the pruned tree.
5. Calculate the R^2 in the test sample.

Exercise 3: Forests

1. Build in the training sample a random forest to predict log online spending. The forest should contain 1000 trees. Each tree should use a 50% subsample of the training data, 1/3 of the covariates, and restrict the `min.node.size` to 100.
   (a) Plot a tree of the forest.
   (b) Plot 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.
2. Draw an area under the curve (AUC) graph with regard to the number of trees in the forest.
3. Build a forest with smaller `min.node.size` (= 10) and test if this improves the R^2 in the test sample.
4. 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 prediction 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.
5. Download the id's and the predicted spendings of the new customers in a csv-file.

Useful links:

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