PC project 3: Causal Trees and Forests

You are expected to solve this PC-Project until and including Exercise 1c until 14.03.2022. Submit the PDF with your answers as well as your Python files with a reproducible code in pc3.py format. All functions should be submitted as pc3_functions.py. Upload your solution in a zip file named pc3_yournames.zip to the designated module in Canvas. Please, provide all 3 files, namely the pc3.py, pc3_functions.py and pc3.pdf. Both the codes and the written answers are graded.

Part 1: Trees and Forests

Download the dataset *data_pc3.csv* and save it in a folder on your computer. Download also *pc3.py* and *pc3_functions.py* and save them in the same folder on your computer. You may re-use the functions from previous PC projects and include them in *pc3_functions.py*.

a. Open *pc3.py* in Spyder. Specify the path variable accordingly and make sure that *pc3_functions.py* is imported. Load the *data_pc3.csv* into your environment as Pandas DataFrame object. Report descriptive statistics.

b. Explain the idea behind the general algorithm for a predictive regression tree. What would be a tree prediction for outcome *Y* with only the root node? Code your own function to compute the SSE (sum of squared errors). Based on SSE, compute the MSE (mean squared error) of such prediction.

c. Predict the outcome *Y* with the covariate *X*. For this purpose, code your own function to find out where an SSE optimizing Regression Tree algorithm would place the first split. Code the function such that it gives back the best splitting value of the covariate *X*, the resulting optimal SSE splitting value and the row index of the corresponding optimal *X* value. Include an option to make sure that the resulting tree leaves contain at least certain number of observations. Report the results for the first split with minimum number of observations in the leaves equal to 10 and calculate the corresponding MSE. (Hint: sort the data in an increasing order according to *X* and reset the indices using the Pandas module.)

d. Proceed further with splitting using the SSE minimizing rule. Where would the tree place the second and third split? Discuss the procedure and the

corresponding results for the optimal tree. Report an overview of the optimal tree splitting values and the resulting size of the leaves as well as the MSE.

Part 2: Causal Forest estimation to analyze causal heterogeneity

Lechner (2018) suggests to use a so called Modified Causal Forest algorithm that builds upon the ideas of Causal Trees and Forests. The algorithm splits the data such that it removes the selection bias and uncovers effect heterogeneity. Additionally, the splitting criterion considers the MSE of the "causal" problem directly. This feature allows to consider also multiple treatments already in the splitting procedure.

For this exercise we use the dataset about newborn's birthweight from the PC project 2. In particular, we use the cleaned dataset we prepared in the last project, namely *data_pc2_clean.csv*, as an input for the Modified Causal Forest (*mcf.zip*).

a. Download the mcf package via *pip install mcf*. Open the dataset (*data_pc2_clean.csv*). Visit the mcf documentation https://mcfpy.github.io/mcf/#/ and familiarize yourself with the basic functionalities of the package. What are the core functions in the package? Summarize the functionality of the core functions in one sentence. Suppose you have a question or have detected a bug. How would you contact the developers?

b. Run the code with the defaults and make sure that you have a saved version of the output in your output subfolder.

c. Inspect the Modified Causal Forest output and infer the following information:

- How many observations are used to build the Modified Causal Forest?
- How many trees are in the forest? What is the subsampling share? What is the minimum leaf size? What is the number of splitting variables? Comment on the meaning of these tuning parameters.
- Report the estimated ATE and compare this value to the results in PC project 2. Do you need different set of identification assumptions for the causal forest?
- Report the distribution of the IATEs and comment on the results. Describe the difference between the ATE and the IATE estimates.

- Interpret the sorted effects plots. Do you think there is a large effect heterogeneity between most and least affected mothers? Comment on the precision of estimated IATEs and the ATE.

d. Now continue with your *pc3.py* script and load the following file *data_pc2_cleanPredpred.csv* which includes the estimated IATEs from the Modified Causal Forest from your "data" subfolder. Plot the IATEs against the mother's age. What do you conclude from this analysis?

e. Analyze the effect heterogeneity along the mothers age formally based on the estimation of the Group Average Treatment Effects (GATEs). Interpret the GATEs plot for the *mage* variable. Do you observe the same decreasing effect as in 2d? Comment on the precision of the estimated GATEs.

References

Lechner, M. (2018). Modified causal forests for estimating heterogeneous causal effects. *arXiv preprint arXiv:1812.09487*.