Causal Machine Learning – Assignment

Instructions

- The assignment consists of 3 tasks. The maximum score you can achieve is 6 points.
- You may work in pairs of two or on your own. If you work in pairs of two, please submit only one solution.
- Send your answers in one well commented R-script to m.schalberger@fu-berlin.de by 11:59 p.m. on Jan. 7, 2026
- Set a random seed for all stochastic procedures.
- Document the packages used.

Economic Context

Marcus et al. [2022] study a sports club voucher program and use regional and cohort policy variation in a difference-in-differences design to estimate the causal long-run effect of subsidies on children's sports participation and health. In contrast, your exercise does not exploit this exogenous policy variation and instead treats current sports club membership as the treatment, with very good health as the outcome. Thus, you are not answering the paper's original policy question, but rather examining the question: how membership in a sports club is associated with the probability of being in very good health, conditional on observed characteristics.

The associated dataset is sportsclub.csv (which you can find in Blackboard). In this assignment you should use the following variables from it:

- Outcome variable
  - health1: very good health
- Treatment variable
  - sportsclub: member in sports club
- Control variable
  - female
  - siblings: has siblings
  - born_germany
  - parent_nongermany
  - newspaper: newspaper at home
  - academictrack
  - urban: living in a city
  - age
  - deutsch: German citizenship
  - bula: Federal state

- obese
- eversmoked
- currentsmoking
- everalc

Load the dataset, select the relevant variables and discard all rows that include NA's (if any).

Task 1

a) Compute the naive estimator for the average treatment effect (ATE) and the 95% confidence interval (you can assume a normal distribution). **(0.5 points)**

b) Estimate the probability of treatment using all control variables (probit). **(0.25 points)**

c) Estimate the Average Treatment Effect on the Treated (ATT) using nearest-neighbor matching with replacement. Then, use propensity score weighting for the estimation of the ATT. **(1 point)**

d) Identify control variables that could be problematic for estimating the treatment effect. Explain why they may cause issues in the analysis and describe briefly whether using propensity score methods addresses these concerns. **(0.25 points)**

Task 2

a) Compute and compare the 10-fold and 5-fold cross-validation errors resulting from fitting a logistic regression model with control variables deemed unproblematic in 1d). **(0.5 points)**

b) Split the data into a 70% training and 30% test set. Estimate lasso, ridge, and elastic net models using cross-validation to choose penalty parameters. **(1 point)**

c) Evaluate all three models on the test set and compare their prediction errors. **(0.25 points)**

d) Compare the coefficients across the three penalized models and comment briefly on differences. **(0.25 points)**

Task 3

a) Fit a classification tree to the training data. Use Gini as the splitting criterion. Display the summary statistics of the tree and describe the results obtained. **(0.25 points)**

b) Now apply bagging and a random forest to the training data. Calculate the misclassification error for each method. Which method do you prefer? Explain briefly. **(1 point)**

c) Perform boosting on the training data using 1,000 trees. Then, calculate the misclassification error. **(0.5 points)**

d) Which variables appear to be the most important predictors in the boosted model? **(0.25 points)**

References

Jan Marcus, Thomas Siedler, and Nicolas R Ziebarth. The long-run effects of sports club vouchers for primary school children. *American Economic Journal: Economic Policy*, 14(3):128–165, 2022.