Self-Study in the Course:

Data Analytics I: Predictive Econometrics

Submission Deadline: February 15, 2023, midnight

Name: ____

Student ID: ____

The data for the self-study are available from December 21 (evening) on Canvas. You can achieve up-to 25 points for this self-study, which has a weight of 25% for the final grade.

You work for a health insurance company in South America. The data file obesity.Rdata contains the current Body Mass Index (BMI) of a sample of adults from Mexico, Peru and Colombia together with characteristics about their eating habits, physical condition and other variables collected via a survey.

The insurance company wants to develop a model to predict BMI based on the survey questionnaire. The BMI predictions will serve for further cost prediction purposes as patients with obesity incur significantly higher medical costs. The file obesity_predict.Rdata contains adults for whom you are asked to deliver first BMI predictions.

Exercise:
Predict the BMI using any method or approach that we either studied during the course or you potentially studied by yourself. **Implement max. three different estimation approaches that you know (e.g. OLS, Lasso, Ridge, Tree, Forest, etc.)**. Select one approach that you believe gives the best predictions. Save these predictions into a file firstnames_lastnames.csv (replace with your name) and use comma as separator for the csv file. Additionally, submit your R code (containing also the approaches that you have implemented but have not used in your preferred prediction) and the answers to the questions below in a PDF format. Please submit all files via Canvas.

Important: Submit only predictions for BMI from one method in one csv file. I.e. the file will contain one column with the predicted BMI (do not change the order of the observations).

The points for the self-study depend mainly on the answers to the questions below and the R code. Additionally, the accuracy of your predictions will be assessed based on the R^2 of the predicted BMI and will be used only marginally for the final score.

Questions:

Write your answers in the boxes below. Give short descriptions to each question.

1. Are there any issues with the data (missing values, problematic labels, small groups in some categories, useless predictors, etc.) Describe your data cleaning procedure.

2. Describe your preferred prediction method/approach. Did you transform variables in training and test data? Did you partition the data? How did you select the tuning parameters? Which estimation method did you use? Why did you select this method/approach for your preferred specification?

3. Besides your preferred specification, did you test additional variable transformations?

4. Besides your preferred specification, did you test additional ways to select the tuning parameters and/or partition the sample?

5. Besides your preferred specification, did you test additional estimation methods/approaches?

6. Did you use or test methods/approaches that were not covered in the lecture?