Prediction with Machine Learning // The Assignment 1

This is an individual assignment.

Work individually. But you may collaborate in your support group, check and comment (add issues) on each other code.

The dataset

Consider the cps-earnings dataset at https://osf.io/g8p9j/ (Cross section. N=149 316 individuals)

- Pick an occupation individually and filter data accordingly.
- Occupation codes are here: https://osf.io/57n9q/
- Instead if a single occupation, you may merge occupations as you see fit (ie all tax/insurance specialists, etc).

You can see some ideas working with this code here.

- https://github.com/gabors-data-analysis/da_case_studies/tree/master/ch09-gender-age-earnings
- https://github.com/gabors-data-analysis/da_case_studies/tree/master/ch10-gender-earnings-understand

Tasks

Build 4 predictive models using linear regression for earnings per hour.

1. Models: the target variable is earnings per hour, all others would be predictors.
2. Model 1 shall be the simplest, model 4 the most complex. All models shall be OLS. You shall explain your choice of predictors.
3. Compare model performance of these models (a) RMSE in the full sample, (2) cross-validated RMSE and (c) BIC in the full sample.
4. Discuss the relationship between model complexity and performance. You may use visual aids.

Output to submit

You shall work in a markdown notebook / Quarto. You should submit your notebook in Github and a pdf report on Moodle.

- In the notebook you export to pdf hide unnecessary output. You shall show key cleaning / filtering decisions, your model estimation and results.
- Code quality is important. Use functions, code style guides
- Make sure output looks pretty

Hints re Git and commit

- Committing is a habit, and people may have different ways.
- Some people commit very frequently, others less so.
- We basically expect you to have a few commits, one per major parts of the exercise. The first commit will set up the folder/file for A1.
- Then you can commit, say data work, descriptive stats, graphics, and regressions. And then, commit your edits.
- Make sure the commit text is short but meaningful: Good: "adding graphs", "calculate RMSE", "edit typos". Bad: "update"

Grading

This assignment is worth 20 points.

- 3 points will be for Git use.
- 10 points will be technical aspects the analysis, and the discussion of your steps
- 7 points will be based on your report including code quality