Problem Set 3. PSID Income Prediction Challenge

Info 3370. Studying Social Inequality with Data Science. Spring 2023

Due: 5pm on 24 Mar 2023. Submit on Canvas.

Welcome to the third problem set!

- Use this .Rmd template to complete the problem set
- If you want to print the assignment, here is a pdf
- In Canvas, you will upload the PDF produced by your .Rmd file
- Don't put your name on the problem set. We want anonymous grading to be possible
- We're here to help! Reach out using Ed Discussion or office hours

This problem set is an individual extension of the PSID Income Prediction Challenge that has been our focus in class. In groups or individually, you have been uploading submissions that contain the individual identifier `g3_id` as well as the predicted outcome `g3_log_income`.

The problem set is an opportunity to take your .R code for that exercise and turn it into a beautifully commented and clean .Rmd file. There are a few requirements beyond the requirements from the class activity.

- Make a conceptual argument for how you chose your candidate learner(s)
  - Example: Why did you choose these predictors? How did you decide about interactions?
- Conduct your own sample split within the learning set
  - Report the test sample mean squared error (MSE) of your candidate learner(s)
- **MPS students.** Include at least one OLS specification and at least one non-OLS specification. Comment on the magnitude of the difference in your test-set MSE between these learners. Full points can be given regardless of whether the difference is large or small.

Collaboration note. Working on prediction code in class in teams is absolutely welcome. Then, you should individually write the .Rmd for the homework explaining what is in your code. The code can be as similar or different to what you worked on in class as you like. Multiple students may submit nearly identical code. The explanations in your .Rmd should be your own independent work.

Grading note. See rubric on Canvas. It is most important to explain your ideas and document your code well. It is less important (and definitely not necessary) to attempt many learners.

Part 1 (25 points). Overview of your approach.

You overview should tell us in English rather than in code:

- how you chose your base learners
- how you conducted a sample split
- what you ultimately chose to make predictions

Unlike previous problem sets, there is no correct and incorrect answer. You will be graded on the clarity of your summary of your approach.

Part 2 (25 points). Commented code

This section will contain your code. We want your code to be clean!

- explain each chunk of code in English
- when piping `%>%`, use at most one pipe per line
- avoid lines running off the page
- indent your code appropriately
  - RStudio will do this for you. Just highlight the code and type `CMD+I` if you get out of line

- we suggest the tidyverse style guide (though perfectly following this is not required)
- report your estimated test-sample MSE for any learner(s) you consider

We recommend that you explain what a bit of code will do in words, then have the relevant code chunk.

```
# example code chunk
```

Then explain the next thing in words, and have a code chunk

```
# example code chunk
```

The goal is to be maximally clear for the reader to understand. We will learn from one another's code!

Computing environment

Leave this at the bottom of your file, and it will record information such as your operating system, R version, and package versions. This is helpful for resolving any differences in results across people.

```
sessionInfo()
```

```
## R version 4.2.3 (2023-03-15)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.2.3  fastmap_1.1.1   cli_3.6.0       tools_4.2.3
##  [5] htmltools_0.5.4 rstudioapi_0.14 yaml_2.3.7      rmarkdown_2.20
##  [9] knitr_1.42      xfun_0.37       digest_0.6.31   rlang_1.1.0
## [13] evaluate_0.20
```