Final project: Predicting Emigration

Version: 2025-05-20

The aim for this final project is to build your own prediction model.

We have created an *artificial* data set of 8400 observations. The artificial data set was designed based on the Arab Barometer Wave IV (www.arabbarometer.org) which has been conducted in Algeria, Egypt, Jordan, Lebanon, Morocco, Palestine, and Tunisia over 2016-17.

Your training data (`AB4x_train.Rdata`) is a 80% random sample of our fake Arab Barometer (6720 observations). We hold back 20% of the remaining observations (1680 observations) for evaluation (`AB4x_eval.Rdata`). Your task is to build a model to predict if a respondent reports any plans to emigrate (Variable `emig`). Your model's performance will be assessed using the holdout data.

Note: while our fake Arab Barometer is designed based on the real survey and the data structure is similar, the statistical properties are fundamentally different from the original. That is, it is not possible to reconstruct the holdout data from the original survey.

We allow for groups of up to 3 students but individual submissions are also fine. We expect you to submit the following files (compressed in a ZIP file):

- `training.R`: R-code of your training process
- `final_model.Rdata`: Your preferred model as a `.Rdata` file
- `evaluation.R`: R-code that evaluates the performance of your preferred model on `AB4x_eval.Rdata`
- `report.pdf`: A brief report (800-1200 words, no more than 2 tables/figures in total) about which models you trained, their performance and how you decided which model to submit as your preferred model

To submit the R-code, we require you to use the templates `training.R` and `evaluation.R` that come with this assignment. The file `training.R` should include all code used for the training process and save your preferred model as a `final_model.Rdata` file at the end. The file `evaluation.R` loads this preferred model and generates predictions for `AB4x_eval.Rdata`. Please include necessary pre-processing steps in the file `evaluation.R`. We will use the file `evaluation.R` together with your final model to assess your model's performance.

To facilitate your submission of R-code that runs through without interference, we provide you with a trial evaluation dataset (`AB4x_eval_mock.Rdata`). This dataset contains no information, but is exactly identical in size and structure to the actual evaluation dataset. This dataset was generated by shuffling the observation within each variable separately, thereby breaking any correlation structure between variables and thereby rendering it useless for model evaluation. We also provide you with a codebook for the recoded data (`AB4x_codebook.txt`).

Your grade will be a function of the quality of the final report (50%) and the performance of your preferred model on the evaluation dataset (50%).

Deadline to submit questions: Wednesday, May 28, 2 PM

Deadline to submit results: Monday, June 2, 2 PM