2020 MEBDI Machine Learning Competition

January 10, 2020

This document describes the 2020 MEBDI ML competition and its rules. This document and all other competition-related material will be posted on MEBDI's website at https://mebdi.org/ml-competition. These rules can be modified to handle unforeseen contingencies, in which case the changes will be announced via email to all graduate students.

ML Problem Description

- **Goal:** Design a computer algorithm that provides the best out-of-sample prediction for wage income of individuals in a sample drawn from the Current Population Survey. More details below.
- **Eligibility:** To be eligible you must have entered the program in Fall 2017 or before, be in good academic standing, and must not be on the job market this year (i.e, you must plan to continue as a full time student in Fall 2020). Each student can enter the competition individually or form a team with one other eligible student (team of two). You are allowed to enter the competition again in future years if you satisfy the eligibility conditions for that year. Teams that submit identical or near identical algorithms will both be disqualified.
- **Evaluation:** The submissions will be reviewed by a panel of professors, who (if need be) will also have the final say on the interpretation of rules.
- **Deadline:** Noon CT on Friday, April 17, 2020. Late entries will not be considered.
- **Prize:** $5,000 total, split into two parts (Part I prize: $1,000 + Part II prize: $4,000).

Details

Problem Description

Start with the following version of the Mincer wage regression, where $y_i \equiv \log(Y_i)$ is log labor income:

$$y_i = d_{\text{age}} + d_{\text{educ}} + d_{\text{age}} \times d_{\text{educ}} + d_{\text{gender}} + d_{\text{state}} + d_{\text{race}}, \tag{1}$$

where d_j indicates a dummy for variable j, and the regression is run in a single cross section (2012). Education is the last degree completed and is a categorical variable with 4 possible values: (i) high school degree; (ii) some college; (iii) college degree; and (iv) graduate education, or at least some post-baccalaureate education. (The dummy for less-than-highschool education is omitted for normalization.)

The competition has two parts:

- Part I: Design an ML algorithm with the best out-of-sample prediction performance (as defined below) for y_i using only the variables in the Mincer regression (1). The actual levels of variables can be used (instead or in addition to dummies) when when applicable.
- Part II: Design an ML algorithm with the best out-of-sample prediction performance (as defined below) for y_i using an expanded set of variables described below.

Data and Variables

The official sample to be used in the competition can be downloaded from the MEBDI competition web site as a zip file: https://mebdi.org/ml-competition/. A brief description of the dataset and variables is as follows:

- March Current Population Survey, 2012, ASEC supplement (so income is for year 2011).
- Downloaded from IPUMS.
- CPS variable for income: log of INCWAGE (last year's total wage & salary income).

Sample Selection

- Ages 25 to 60, inclusive (in 2011)
- Drop individuals with imputed annual earnings, or with missing data on age, education, and gender. Keep if state or race is missing.
- Drop if INCWAGE is less than \$2,000 (2012 dollars) or if either the income from main job (INCLONGJ) or from other jobs (OINCWAGE) exceeds the top code minus \$1,000 (corresponding to \$249,000 and \$49,000 respectively for each variable) .
- Drop if hourly wage $>$ \$500, annual hours $<$ 50 hrs. Hourly wage is INCWAGE divided by the product of "usual weekly hours" and "weeks worked".
- Once the main sample is selected, it is split into training subsample containing 80% of observation, with the remaining 20% constituting the test sample–for out-of-sample prediction. These samples are already fixed in the sample you will download.

Variables Allowed As Predictors in Part II

You can use any variable available in the 2012 CPS, with the exception of those that are directly linked to the income variable that is being predicted. The list of disallowed variables includes:

- All variables in ASEC's INCOME and TAX modules
- All variables in the WORK module with the exception of occupation and industry of employment, which can be used
- Variables about alimony paid, child support, and the like.

It is not practical to list all the variables that that are disallowed but if you are unsure you are welcomed to ask for clarification about specific variables.

Software and Packages

- You are allowed to use two compiled languages (Fortran or versions of C) or the following five high-level programming languages: Matlab, Python, R, Stata, or Julia, or any combination of them. You can also use outside libraries and packages that work with these languages as long as the source code is freely and easily available (for inspection). Libraries that are part of a language, such as Matlab's Statistics and Machine Learning Toolbox, are allowed. Because it is impossible to anticipate every contingency, if you are unsure, please check with Fatih if the software is eligible before you start using them.
- If you are using any extra packages or libraries that is not part of the base programming language, they must be included with the source code or if that is not feasible, they must be described in the report with all the information necessary for the committee to access those packages and libraries.

Deliverables

To be admissible, your submission must include the following:

1. A clearly written, concise **report** (ideally between 2 and 4 pages) that
   - summarizes the bottom line result, including the three performance measures of out-of-sample prediction described below.
     - A detailed description of the ML algorithm(s) used, describing all the necessary modifications and all the specific choices you have made in every step. Someone who reads this report (and the committee will read it) should be able to obtain the code based on your description replicate exactly what you did and get the same out-of-sample measures. [1]

   (a) All the source code for your entry (in one zip file) and the executable program if you are using a compiled language. If you are using any extra packages or libraries that is not part of the base programming language, they must be included with the source code or if that is not feasible, they must be described in the report with all the information necessary for the committee to access those packages and libraries.

Success Criteria for Out-Sample-Prediction

For both Parts I and II, there are three performance measures of out-of-sample prediction that will be considered.

1. **RMS-log:** the root mean square (RMS) of prediction errors (in natural log of earnings) in the test sample:
$$\text{RMS} = \sqrt{\sum_{i \in \mathcal{T}} (y_i - \widehat{y}_i)^2} \tag{2}$$
where $\widehat{y}_i$ is predicted income and $\mathcal{T}$is the training sample.
2. **RMS-level:** Replace the logs with levels of income in equation (2).
3. **AbsDev-log:** maximum absolute prediction error in logs (taken across all observations in the prediction sample):
$$\max_{i \in \mathcal{T}} |y_i - \widehat{y}_i|$$

[1] For example: "We use the R code for elastic-net regularized linear models written by Robert Tibshirani et al available for download at https://cran.r-project.org/web/packages/glmnet/index.html." Then describe all the user-specific choices you made, etc.

Win Rule: To win either part of the competition, an algorithm must satisfy two conditions:

1. reduce the RMS-log measure for the test sample by at least 5% relative to the Mincer regression (1), and
2. produce an RMS-log measure for the test sample that is at least 1% lower lower than the next best entry.
   (a) *Tie breaker 1:* If no algorithm satisfies (2) (so there is more than one algorithm within 1% of the lowest RMS-log measure), but one team has an RMS-level measure that is 1% lower than the others, that team wins.
   (b) *Tie breaker 2:* If 2.a also fails but one algorithm has an AbsDev-log measure that is 1% lower than the other team(s), that team wins.
3. If two or more teams are tied as described for all three measures, they will be declared winners and will share the prize.