September 2023

2023–24 BFI-MEBDI Machine Learning Competition

This document describes the 2023–24 BFI-MEBDI ML competition organized by Big Data Initiative at the Becker-Friedman Institute Institute at the University of Chicago and the Minnesota Economics Big Data Institute at the University of Minnesota. This document and all other competition-related materials will be posted on MEBDI's ML Competition web page at https://mebdi.org/ml-competition-2023 and BFI's big data initiative web page at https://bfi.uchicago.edu/initiative/big-data

ML Problem Description

- **Goal:** Design a computer algorithm that provides the best out-of-sample prediction (imputation) of individual earnings above a topcoding threshold using information from individuals' (topcoded) earnings dynamics in years preceding the imputation.
- **Eligibility:** To be eligible, as of September 20203, you must be in the 2nd year or later (i.e., entered the PhD program in the Fall of 2022 or before) in the economics PhD programs at the University of Minnesota Economics Department or University of Chicago (Economics Department, Booth School, or Harris School) and be in good standing. Each student can enter the competition individually or form a team with one other eligible student from the same school (team of two). To enter the competition, teams must notify their decision to participate to organizers via email at `BFI.MEBDI.MLCompetition2023@gmail.com` by October 31, 2023. Teams that submit identical or near identical algorithms will both be disqualified.
- **Deadline:** All deliverables listed below must be submitted (to the email address given above) by February 16, 2024 at Noon US Central Time. Late entries will not be considered.
- **Prizes:** First place: $7,500. Second place: $2,500.

See next page for details.

1 Details

Problem Description

Objective: Design a Machine Learning algorithm with the best out-of-sample performance (as defined below) for imputing individual earnings above a topcoding threshold using the history of topcoded earnings for the same set of individuals. Here are the specifics:

You are given a balanced panel of individual earnings for $N = 78,723$ individuals over 6 consecutive years. Earnings above a threshold (that changes every year) are topcoded in each year, *except* in the first year.[1] You are also given the age of each individual (between 25 and 55 years old) in the first year of the sample.

The objective is to impute earnings of the subsample of individuals whose earnings are topcoded in years $T = 5$ and $T = 6$ to get the best goodness of fit measure(s) described below.

Data and Variables

The official sample to be used for the competition is posted at this link: https://mebdi.org/s/BFi-MEBDI-Offical-Sample.csv. The dimensions of the dataset you are given is $N \times (T + 2)$ with individual IDs in the first column, age in the second column, and earnings observations in the remaining $T = 6$ columns.

The panel is balanced and is constructed by taking a cross-sectional sample of individuals in year 1 and following them in all subsequent years. Individuals who have zero income in 4 years or more have been dropped from the sample.

Software and Packages

You are allowed to use any computational language (Fortran, C, Python, Matlab, Julia, etc.) as well as outside libraries and packages as long as those libraries/packages are publicly available (possibly at a reasonable cost). If you do use any extra packages or libraries that are not part of the base programming language, they must either 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 obtain/access those packages and libraries. The main requirement is that others (including the competition judges) should be able to run the code you submit for the competition and replicate the results you submit (obtaining the libraries/packages if needed).

2 Deliverables and Criteria for Win

Deliverables

To be admissible, your submission must include the following:

[1] The topcode thresholds in years 2 to 6 are 140,000, 141,000, 142,000,143,000, and 144,000, respectively.

1. An $N \times 3$ dataset that contains: Individual IDs in column 1 and the predicted (imputed) income values for years $T = 5$ and $T = 6$ in columns 2 and 3. (Of course, for nontopcoded observations, your dataset will contain the actual values.)

2. A clearly written, concise report (between 3 and 5 pages) that

   (a) contains an executive summary (half-page or so) of the method(s) you have used in producing your imputations.

   (b) a detailed description of the ML algorithm(s) you have used, describing all the necessary modifications and all the specific choices you have made in every step. Someone who reads this report (e.g., the competition judges and others) should be able to write a code based on your description and replicate exactly what you did.[2]

3. All the source code for your submission (in one zip file). 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

The dataset has been artificially topcoded, and we (the judges) have access to the original observations. Let y^*_{it} denote the uncensored log earnings of individual i in year t. You observe $\widetilde{y}_{it} = y^*_{it}$ when the (i,t) observation is not top coded, but you only observe $\widetilde{y}_{it} = \overline{y}$ (threshold) when the (i,t) observation is top coded. Our measures of success are based on comparisons between the uncensored $y_{it}{}^*$ and your predictions, $\hat{y}_{it}$. Specifically, we consider three performance measures of out-of-sample prediction. Two of the performance measures are based on imputed earnings in year 6 alone. The third one is based on imputations both in years 5 and 6. Let $\mathcal{S}_5$ and $\mathcal{S}_6$ denote the subsamples of observations that are topcoded in years 5 and 6, respectively, with corresponding sample sizes of S_5 and S_6.

1. **RMS-log:** the scaled root mean square (RMS) of prediction errors (in natural log of earnings) in subsample $\mathcal{S}_6$

$$\text{RMS-log} = \sqrt{\frac{1}{S_6}\sum_{i\in\mathcal{S}_6}(y^*_{i6} - \widehat{y}_{i6})^2} \tag{1}$$

   where $\widehat{y}_{i6}$ is predicted income in year 6 and $\widetilde{y}_{i6}$ are the (topcoded) observations available to you.

2. **RMS-level:** Replace the logs with levels of income in equation (1):

$$\text{RMS-level} = \sqrt{\frac{1}{S_6}\sum_{i\in\mathcal{S}_6}(\exp(y^*_{i6}) - \exp(\widehat{y}_{i6}))^2} \tag{2}$$

[2]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.

3. **Variance of earnings growth:**

$$VEG = \sqrt{|(var(y^*_{i6} - y^*_{i5}) - var(\hat{y}_{i6} - \hat{y}_{i5}))|} \tag{3}$$

where the variances are computed over the subsample of individuals who are either in $\mathcal{S}_5$ or $\mathcal{S}_6$ but not both.

Goodness of Fit Measures:

$$GOF_1 = \frac{4}{5}\text{RMS-log} + \frac{1}{5}\text{VEG} \tag{4}$$

$$GOF_2 = \frac{1}{5}\text{RMS-level} + \frac{4}{5}\text{VEG} \tag{5}$$

Win Rule:

To win the competition, an algorithm must deliver a GOF_1 measure for imputed observations that is at least 1% better (lower) than the next best entry.

Tie-breakers: If there is more than one team within 0.99% of the lowest GOF_1 measure, then the team with the lowest GOF_2 measure wins IF their GOF_2 measure is at least 1.5% lower than the next best team. If there is still a tie, then if a team has a combined GOF_1 + GOF_2 score that is at least 0.5% lower than the second lowest combined score, the team with the lowest combined score wins.

If two or more teams are still tied after the tiebreakers, they will be declared joint winners and will equally share the total prize money ($10,000).

The submissions will be evaluated by a committee of faculty members who will also have the authority to interpret the rules if needed.

Organizers
Stephane Bonhomme (U Chicago)
Fatih Guvenen (UMN)
Kirill Ponomarev (U Chicago)
David Wiczer (FRB Atlanta)