MEBDI Machine Learning Competition: Judge's Summary

The Judges evaluated submissions from three teams: (1) Lejvi Dalani and Hasan Tosun , (2) Egor Malkov and Filip Premik, and (3) Dhananjey Ghei and Sang Min Lee.

Part 1: Predicting CPS Wages with limited covariates

Despite submissions employing an array of strategies (boosted trees, deep learning, lasso, random forests, bagging trees, and support vector machine regression), no team was able to make substantial improvements (more than 5%) in fit relative to the baseline Mincer regression. Further, all three teams were within 1% of each other according to all three goodness-of-fit criteria (Log-RMSE, RMSE, and absolute deviation). As such, we have decided to split the prize money for this portion of the problem evenly amongst participants. While it may seem puzzling that no machine learning technique can improve on the Mincer regression, the reason is quite intuitive. As Malkov and Premik point out in their submission, since each of the covariates in this limited set of variables are categorical variables, a nonparametric regression can be run by including all interaction terms, and will by definition produce the estimate with the best mean squared error in large samples. Since the Mincer regression is not far removed from this nonparametric regression, the gains in prediction will be small.

Part 2: Predicting CPS Wages with expanded covariates

In this section, Ghei and Lee provided the winning submission with Log-RMSE of 0.25 on the training dataset and 0.36 on the test dataset. All submissions employed an ensemble method using some combination of boosted trees, deep learning, lasso regression and random forests, though there seems to have been more variation in strategies for data preparation and variable selection. In their winning submission, Ghei and Lee first create additional variables using principal component analysis and taking means of continuous variables by broad occupation, industry, and geographic area. They next use LASSO to select the subset of variables to feed into the final machine learning algorithm. In their final step, they perform separate fitting exercises using XGBoost, Light GBM, and Deep Learning algorithms. In each case, hyperparameters were chosen using 5-fold cross-validation with random grid search. The best-fitting models from each approach were then combined using a stacking method. The Judges were able to successfully replicate the results of this submission using 20 cores on the LATIS cluster.