Prediction with ML

Anton Shestakov

February 2025

Coding: https://github.com/Anton21a/Machine-Learning.git

This report presents an analysis of hourly earnings prediction using the cps-earnings dataset. The focus is on drivers/sells worker and truck drivers (9130 code), selected from the dataset. The analysis involves four models with different sets of predictors.

The general model specification follows:

$$\ln(\text{Hourly Wage}) = \beta_0 + \sum \beta_i X_i + \varepsilon, \quad \text{i.i.d.} \tag{1}$$

We estimate four regression models:

- **Model 1:** Basic model with race and gender (assuming that gender distribution across drivers (especially truck drivers is skewed. Analogical disproportion might refer to race in the context of the US)
- **Model 2:** Adds age and age^2 (standard variable with concave functional form)
- **Model 3:** Incorporates marital status (given individual family status, a person might be motivated to exert different extent of effort at work)
- **Model 4:** Full model, adding education levels (suggesting that more educated people are less likely to work as drivers)

The models were evaluated on their RMSE across 5-fold cross-validation, with results summarized below:

Model	RMSE (Full Sample)	BIC (Full Sample)	5-Fold Cross-Validated mean RMSE
Model 1	0.4716895	4305.7	0.4712718
Model 2	0.4542617	4081.8	0.4541101
Model 3	0.4540238	4094.6	0.4541925
Model 4	0.4499082	4060.7	0.4505286

As expected, increasing the complexity of the model led to lower RMSE values, both on the full sample and in cross-validation. The Figure 1 shows the average RMSE after 5 fold CV processing. Moving from the first to the second model significantly improves mean RMSE by adding new confounder, but then the effect becomes rather diminishing. The Model 4 (the most complex) had the best predictive performance for each fold (Figure 2). The small gap between training and cross-validated RMSE suggests that overfitting is not a major concern.

Figure 1:

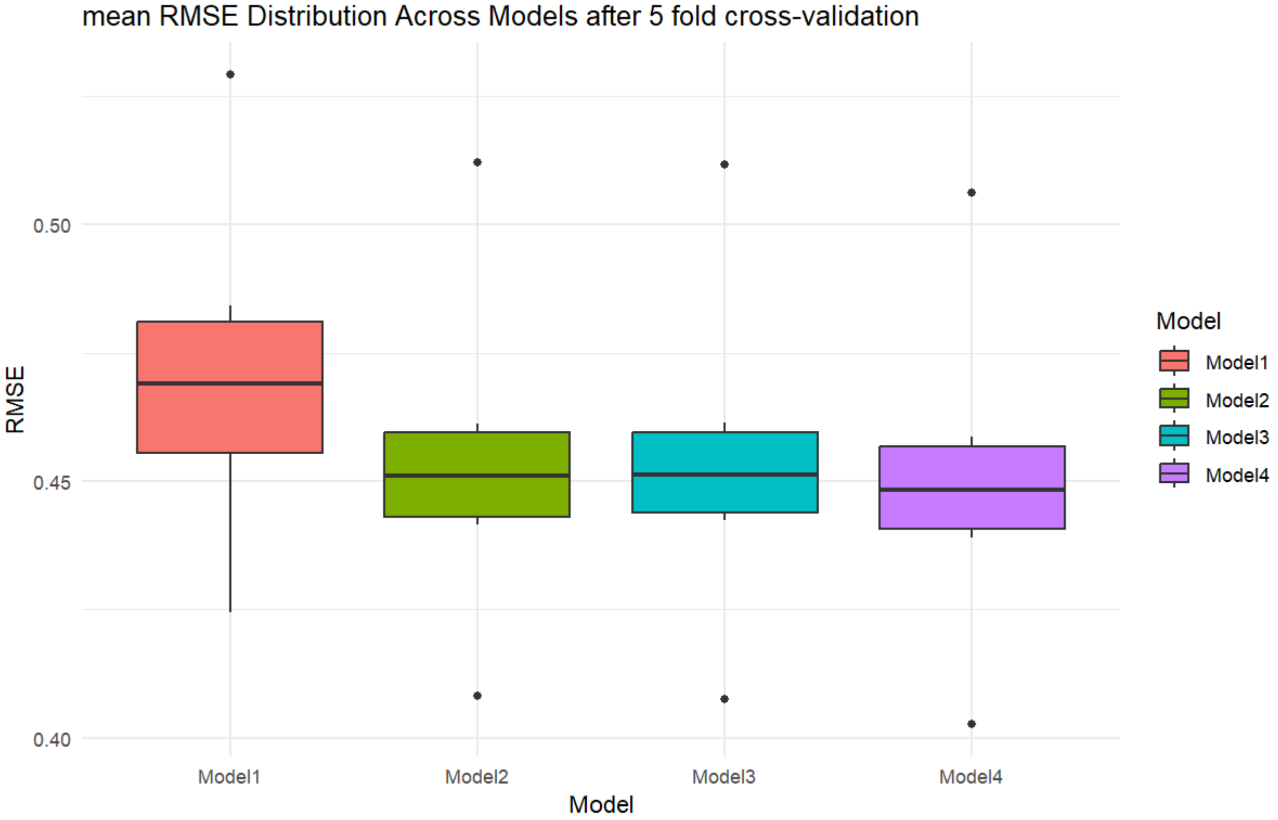

Figure 2: