Prediction with Machine Learning for Economists - Assignment 1

For this assignment, I select the occupation of Financial managers (occ2012 = 120).

1. Models Specifications

In all Models, the target variable is earnings per hour, all others would be predictors.

Model 1: I choose gender, age, and the interaction between gender and age as predictors. There is no doubt that gender and age have an impact on wages. At the same time, the marginal impact on wages may differ for males and females as age increases. Therefore, I also include the interaction term of gender and age in the regression.

Model 2: On the basis of model 1, I additionally added the square term of age, race, and prcitshp (citizenship status). The marginal effect of age on wages is not constant and I believe that race and citizenship status discriminate against job categories, which in turn affects wages.

Model 3: Based on model 2, I further added the factor of the family (ownchild and marital) as well as education level (grade92 and female*grade92). First of all, the number of children and whether or not they are married affects people's work effort status. For example, when they get married or have children, people have to increase their efforts to obtain higher wages to support the family. The impact of education on wages is self-evident, and at the same time, the marginal impact of education differs for males and females. Therefore, I also include the interaction term of education and gender in the regression.

Model 4: On the basis of model 3, I added stfips (states) and age cubes, age quadratic and their interaction term with gender. The salary level may vary from region to region due to policy, economic development level, and other factors. Additional predictors are added in preparation for the following analysis.

2. Compare model performance of these models (a) RMSE in the full sample, (2) cross-validated RMSE and (c) BIC in the full sample.

(a) As the complexity of the model increases, the RMSE in the full sample becomes smaller.

(c) BIC in the full sample of model 3 (14 coefficients) is the smallest. BICs in the full sample of model 1 (4 coefficients) and model 2 (10 coefficients) are similar, and BIC in the full sample of model 4 is the biggest (68 coefficients)

(b) The average of cross-validated RMSE in model 3 is the smallest.

Based on the above comparisons, I think model 3 is the best model for prediction.

3. Discuss the relationship between model complexity and performance.

According to the last figure inside the python code, we can see that the RMSE of the Training Sample decreases as the model complexity increases. However, the RMSE of the Test Sample has a minimum value as the complexity of the model increases, after which the RMSE increases with the increasing complexity of the model. It indicates that there exists an overfitting problem of the Test sample (or live data) in model 4.

The complexity of the model is specific to the original data, but fitting patterns in the original data that are not there in the population, or general pattern, it represents. Therefore, we need to pay attention to the overfitting problem of the complex model since it may not always give a better prediction.