REPORT - HW1 for class Predictions with Machine Learning for Economists

Goal: The aim of my project is to predict earnings per hour for chosen occupation - “doctors”.

Sample design: For my analysis I use the CPS data set. In this data set there is a defined group “healthcare practitioners and technical occupations”. I am interested only in healthcare workers whose daily job includes treating or diagnosing people, so I exclude from these groups health industry technicians, veterinarians, scientists, pharmacologists. Moreover, I exclude nurses and opticians since they differ from doctors in terms of background and required education (I leave only professions which require at least an MA degree. This way I only have specialization of doctors which require similar **minimum** educational background.I exclude codes that include “all other” to avoid including occupations which in reality would not classify as doctors. I drop from the observations those doctors who are employed but absent from work since I suspect they may be too different from others. Moreover, these reasons are unknown to me.

Selection of variables: Standard labour theory suggests that education, age and gender have an effect on salaries. I include these variables. Moreover, age may have a non-linear relationship on wages. Thus, I also include age squared. Labour theory also suggests that the number of children might have an effect on wages. However, in my distribution of children that people have I see that the distribution is significantly concentrated near 1 (most people have 1 child). Thus, I decided to create dummy variable children where 1 would stand for having at least 1 child, 0 - having no children. I suspect that for doctors whether the doctor is employed in the private sector or not or whether he is a citizen (for legal reasons) might affect his salary. I create dummy variables for this. I also include a dummy variable for whether the person is married and his spouse is present or not. I suspect that marriage status can have an effect on work performance and, as a result, on hourly wage. Last but not least, I include the interaction of gender variable and gender. I suspect that if there is some gender differential, it may vary by age. For more complex models I evaluate interactions of the following pairs of dummy variables: gender and children, gender and spouse, gender and citizen, private and citizen. Based on analysis of barplots for interaction effects (see graphs on github) I exclude from this list interaction of private and citizen since it does not show significant interaction effect. I on purpose didn’t consider presence in union although for some other professions it might be an important factor. Not all doctors in the U.S. are allowed to form a union, so that is not an appropriate variable.

**The linear relationship of included variables to ln wage is tested when I run linear regression models of different specifications later. The only thing that may be a concern is the “private” variable (whether the doctor works for a private organization or government) which didn’t appear significant in any of the models where it was included. Perhaps for future projects on this topic, I should have excluded private variable from all mode*ls

Predicted variable: ln earning per hour. I chose the ln form after I built both distribution for hourly wage and ln hourly wage (see github graphs). I saw that distribution for hourly wage isn’t normal while ln hourly wage is much closer to normal.

Models: I build the following 4 models for predicted variable:

Model 1: age, age^2
Model 2: age, age^2, educ, female
Model 3: age, age^2, educ, female, fem_age, private, citizen, spouse, children
Model 4: age, age^2, educ, female, fem_age, private, citizen, spouse, children, fem_child, fem_citizen, fem_spouse

*female stands for gender where 1-female, 0-male

Comparison of RMSE, RMSE-CV and BIC for all these 4 models.

Model	Model 1	Model 2	Model 3	Model 4
RMSE	0.6988471560149766	0.6796746340746923	0.6750037388641617	0.6743190445350765
BIC	3478	3402	3416	3435
RMSE-CV	0.694783	0.675001	0.670105	0.669177

Based on RMSE: Model 4 is better (the lowest RMSE), Model 3 is very close
Based on BIC: Model 2(first best, the lowest BIC) and Model 3 are better
Based on RMSE-CV: Model 3 and Model 4 are better, being very close in average RMSE-CV value

Complexity evaluation: Since I add more variables, the complexity of the models is improving from Model 1 to Model 4.

Performance evaluation: As expected by econometrics theory, RMSE improves with complexity of the model (from 1 to 4 RMSE is decreasing). Adjusted R-squared improves (also as expected in econometrics) from Model 1 to Model 3. However, it is the same between Model 3 and Model 4. I think it can be explained by the fact that in my models I didn't add new variables when I changed from Model 3 to Model 4, I only added interactions. I can see that if going from Model 1 to Model 4 BIC stops improving after a while. Smaller BIC is associated with better performance, so starting from Model 3 as Models become more complex, performance starts to decrease. As RMSE, RMSE-CV also improves with model complexity. However, in both cases relative improvement gradually decreases and the difference in both RMSE-s between Model 3 and Model 4 is almost neglectable.

Model	Model 1	Model 2	Model 3	Model 4
Adjusted R-sq	0.066	0.115	0.125	0.125

Decision: The fact that R^2 contradicts BIC isn’t that crucial since BIC better accounts for risk of overfitting in more complex models. What is crucial, is the contradiction between BIC and cross validation. Although, in some situations we are suggested to use cross validation over BIC (because cross validation isn’t based on auxiliary assumptions) when they produce conflicting results, personally, here I would choose Model 3 since it was identified as a good one by all criterias. Moreover, improvement in BIC from Model 2 to Model 3 is rather small (compared to change between, for example, Model 1 and Model 2), so most likely I won’t make a big mistake by choosing Model 3.