Prediction with Machine Learning for Economist 20.01.2023
Assignment 1 – Report
Mehmet Ataberk Atan

The dataset: I chose the occupations:

- 830: 'credit analysts',
- 840: 'financial analyst',
- 950: 'unclassified financial',
- 1610: 'biologists',
- 1700: 'physicists',
- 1720: 'chemists',
- 1760: 'physical scientists, other',
- 1910: 'biological technicians',
- 1920: 'chemical technicians',
- 1965: 'misc. science technicians'.

Variables: Overall, I used the variables; age, sex, race, marital status, presence of own child, class of the employer (private, public etc.) and occupation indicators themselves. For the categorical variables I created binary dummies. For features, I transformed and interacted a subset of the variables where appropriate. Instead of an algorithmic approach I relied mostly on human judgement for feature engineering since the number of variables are not 'too high'.

Age and sex are well established correlates of earnings in the literature. I tried to make use of the quadratic relationship (positive first derivative, negative second derivative) between earnings and age by adding its square in Model 2. Even the cubed age is added in Model 4 too boost model complexity for the sake of this exercise.

An interaction of age and sex is also added in Model 3. Reasoning is that effect of aging on earnings may differ for men and women. Or similarly the gap in earnings between men and women varies with age.

Earnings also varies with race (possibly through many mechanisms) and other household/individual characteristics included in the models.

Finally, in addition individual/household characteristics, I added occupational indicators and state/federal/private ownership indicators of the employers. The subset of vocations chosen in the dataset is likely to vary with earnings since there are dissimilar vocations involved despite the two main groups – analysts and scientists.

Results: I evaluated the performance of four models based on their mean cross validated RMSE, mean RMSE, and BIC values. I find that the performance of the models appears to improve with increased complexity, apart from BIC. The RMSE and BIC values were calculated on a test dataset that the models had not previously seen. However, it is acknowledged that as model complexity increases, performance on hold-out data should decrease due to overfitting. The results suggest that there may still be room for adding more features to improve performance on live data, as the performance on the test set does not yet indicate overfitting. They also suggest that the results should be tested for robustness using different seeds, different train/test split proportions, and algorithmic feature selection. Finally, I should admit that the calculation of BIC is uncertain and the increase even from Model 1 to Model 2 seems counter intuitive. However, the conclusions drawn from the results still hold.