DA3 - Assignment 1

Adam Kovacs

26 January 2022

Introduction

In this project, I analyze the **cps-earnings dataset** to build predictive models targeting the earnings per hour of advertising, promotions, marketing and sales managers. Altogether 4 models are built, all OLS, but increasing in complexity. The performance of these models are evaluated through RMSE and BIC in the full sample and through cross-validated RMSE. Finally, the relationship between model complexity and performance is illustrated with the help of a visual aid.

Preparatory steps

Having filtered the dataset to the chosen occupations, label engineering was needed. The required target variable is earnings per hour, but we have data on weekly earnings and usual weekly working hours. Thus, I calculated the target variable from these two through a simple division. Next, descriptive statistics of earnings per hour were looked at to identify potential extreme values. I detected an extremely low value and since earning such a low wage is not even allowed by law, a filter of at least 1 USD hourly wage was applied (losing one observation). Descriptive statistics after filtering is available in Table 1 of the Appendix.

Next came feature engineering: From the categorical variable on highest education, 6 dummies were created, namely did not finish high school, finished high school, have an associate degree, a bachelors degree, a masters degree or professional degree. Dummies are created also for gender (female as 1), for union membership, whether the person is native, whether the workplace is in the private sector and the employment status. As for functional forms, an important numerical variable where this is important is age. Between age and earnings, the lowess on Table 2 of the Appendix suggests a non-linear, quadratic relationship, so we create age squared as a feature as well.

Modelling and evaluation

As noted above, four models are built. In the first, simplest model, only one variable, age (and its square) are used as predictors, which are good proxies for experience. In the second model, we include the traditional university degrees (bachelors and masters) that are also highly valued by firms when deciding on wages. In the third model, besides adding professional degree as well, important characteristics of people (gender, race, union membership) and firms are also added (whether it is in private sector). Finally, in the most complex model, all (seemingly less valuable) highest education dummies and further personal traits are also included (marital status, number of children, whether they are native).

First, we evaluate the models using the full sample. The exact values are displayed on Table 2 of the Appendix. Based on the Bayesian information criterion, the best model is the third one, which has the lowest number. Looking at the RMSE, however, it has its minimum at the fourth, most complex model.

A more robust evaluation method is also looked at, using 4-fold cross-validation. Based on the average RMSE of the four folds, it leans towards the suggestion of the RMSE in the full sample, not the BIC. The fourth model has the lowest value, beating the third one by quite some margin (14.87 compared to 14.95).

The relationship between model complexity and the performance of the models is also illustrated through a visual aid on Figure 2 of the Appendix. Going from the first to the second model improved the predictive power substantially: traditional university degree seems to be indeed very important. Next, adding further features describing the characteristics of people and the firm improved the RMSE further, but to a lesser extent. Finally, adding all remaining predictors resulted in a slightly even better performance after all, but the tendency of smaller improvement with higher complexity is evident. Adding even more functional forms/interaction terms could easily even worsen out-of-sample performance.

Conclusion

To conclude, the task was predicting the hourly wage of advertising & promotions managers and marketing & sales managers. Four models were built with increasing complexity containing 2, 4, 9 and 16 predictors respectively. Based on the full-sample and average cross-validated RMSE, the most complex one performed the best. However, based on the BIC, the less complex third model turned out to have the best predictive power.

Appendix

Table 1: Hourly wage of Advertising and Sales Managers

	Mean	Median	SD	Min	Max	P05	P95
Hourly wage	33.36	30.00	16.54	2.78	100.00	11.53	64.10

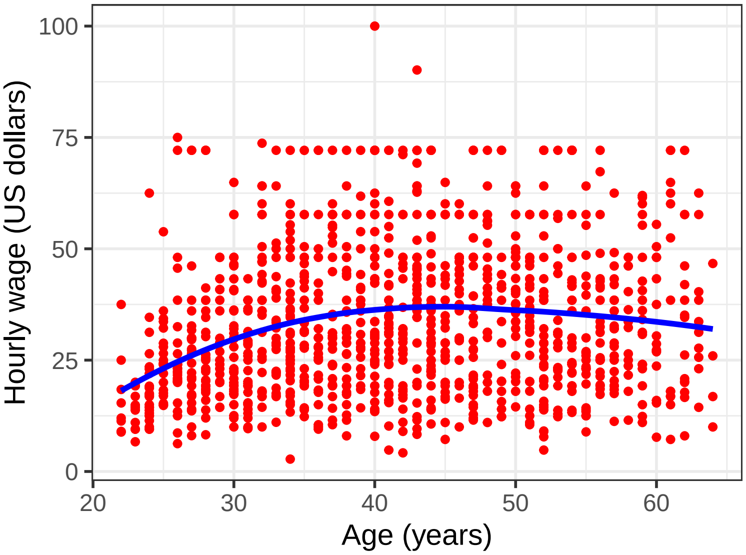

Figure 1: Relationship between age and hourly wage

Table 2: Summary of evaluation on full sample

	Model	N predictors	R-squared	Training RMSE	BIC
1	(1)	2	0.06	15.99	9, 164.68
2	(2)	4	0.15	15.20	9, 067.52
3	(3)	9	0.19	14.87	9, 055.11
4	(4)	16	0.21	14.68	9, 076.34

Table 3: Summary of cross-validation obtained RMSEs

	Resample	Model1	Model2	Model3	Model4
1	Fold1	16.32	15.60	15.58	15.53
2	Fold2	15.78	15.21	14.89	14.74
3	Fold3	15.84	14.91	14.70	14.53
4	Fold4	16.06	15.13	14.63	14.67
5	Average	16.00	15.22	14.95	14.87

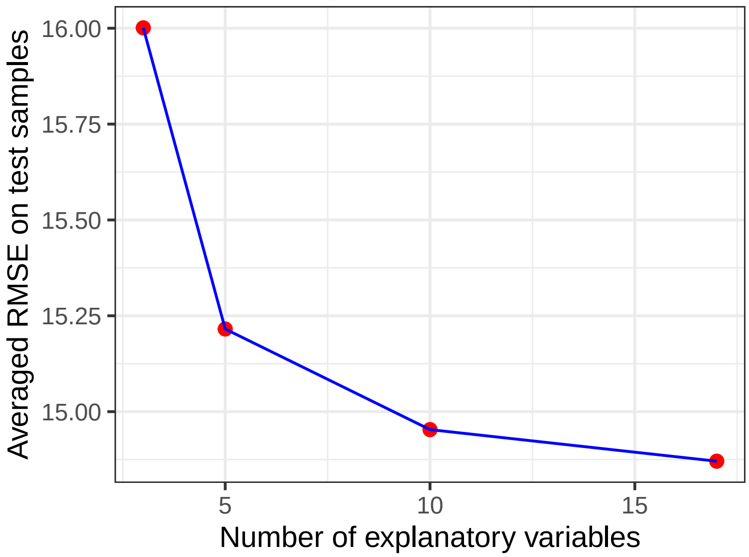

Figure 2: Prediction performance and model compexity