Data Analysis 3

Assignment 1

Introduction

The cps-earnings dataset contains socio-demographic and wage data on U.S. workers. The goal of the assignment is to create models which predict the wage of a specific occupation (Lawyers, Judges, magistrates, and other judicial workers).

Data Wrangling

The dataset was filtered according to the goal, for the chosen occupation. After the EDA process, to create a more precise prediction, the usual working hours were filtered to be between 20 and 60 hours. After the proper filtering the dataset has 989 observations.

A graph with a line and a line

Description automatically generated with medium confidenceA graph of a number of blue bars

Description automatically generated

Feature engineering

As stated previously the target variable of the prediction models is the wage (w) of the workers. The predictors were age and age squared as continuous variables while dummy variables were created for each of the categories of the following categorical variables: sex, highest education, ethnicity, citizenship, marital status, number of children in the household, employment type, union membership. The four models in order consist of 2, 12, 17 and 28 variables. A sanity check confirms that the inclusion of chosen variables makes sense.

Model Comparison

For the comparison of model performance BIC, full data RMSE, and cross validated RMSE is used. Model 2 has the lowest BIC number; Model 4 has lowest full-data RMSE and cross-validated average RSME value. This shows the trade-off between model complexity and predictive accuracy.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **BIC** | **Full Data RMSE** | **Cross Validated RMSE** |
| Model 1 | 8363.795578 | 16.428402 | 16.312359 |
| Model 2 | 8335.808717 | 15.697199 | 15.743707 |
| Model 3 | 8360.996222 | 15.678170 | 15.719916 |
| Model 4 | 8373.459126 | 15.236640 | 15.241130 |

A graph with green bars

Description automatically generated

Additionally Model 2 also offers a more straightforward interpretation of the coefficients of the variables, for the purposes of predictions this is not needed. Instead, the most complex model, Model 4 has the best ability to predict our target variable. This may be caused by the great variance in the earnings which suggests that a very detailed model is needed to capture this.

Conclusion

Overall Model 4 shows both the lowest full data RMSE (which is logical as it has the most variables) and cross validated RMSE and thus it is the most capable to predict the target variable. This suggest that given the high variance in the target variable the overfitting is not yet true, even with this level of complexity.

github: <https://github.com/mateschieszler/DA3_A1>

Máté Schieszler

Appendix

Descriptive Statistics

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **variable** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| age | 989.0 | 43.161780 | 10.905191 | 21.0 | 34.0 | 43.0 | 53.0 | 64.0 |
| agesq | 989.0 | 1981.742164 | 971.713979 | 441.0 | 1156.0 | 1849.0 | 2809.0 | 4096.0 |
| female | 989.0 | 0.400404 | 0.490228 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| ed\_other | 989.0 | 0.055612 | 0.229286 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| ed\_MA | 989.0 | 0.055612 | 0.229286 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| ed\_Pro | 989.0 | 0.703741 | 0.456838 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| ed\_Phd | 989.0 | 0.185035 | 0.388523 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| white | 989.0 | 0.873610 | 0.332457 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| afram | 989.0 | 0.056623 | 0.231237 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| asian | 989.0 | 0.049545 | 0.217113 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| hisp | 989.0 | 0.045501 | 0.208505 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| othernonw | 989.0 | 0.019211 | 0.137337 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| nonUSborn | 989.0 | 0.049545 | 0.217113 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| married | 989.0 | 0.660263 | 0.473860 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| divorced | 989.0 | 0.077856 | 0.268081 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| wiwowed | 989.0 | 0.007078 | 0.083874 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| nevermar | 989.0 | 0.244692 | 0.430122 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| child0 | 989.0 | 0.580384 | 0.493746 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| child1 | 989.0 | 0.050556 | 0.219200 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| child2 | 989.0 | 0.022245 | 0.147553 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| child3 | 989.0 | 0.116279 | 0.320721 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| child4pl | 989.0 | 0.230536 | 0.421389 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| fedgov | 989.0 | 0.125379 | 0.331316 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| stagov | 989.0 | 0.102123 | 0.302964 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| locgov | 989.0 | 0.085945 | 0.280425 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| nonprof | 989.0 | 0.048534 | 0.215000 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| forprof | 989.0 | 0.638018 | 0.480817 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| union | 989.0 | 0.086957 | 0.281914 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |

Cross Validated RMSE table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fold** | **Model1** | **Model2** | **Model3** | **Model4** |
| Fold1 | 16.708608 | 16.061545 | 16.037045 | 15.601646 |
| Fold2 | 16.439893 | 15.854508 | 15.830486 | 15.366008 |
| Fold3 | 15.925247 | 15.492489 | 15.474431 | 15.074837 |
| Fold4 | 16.139987 | 15.541313 | 15.525615 | 15.046968 |
| Fold5 | 16.348058 | 15.768682 | 15.732001 | 15.116191 |
| Average | 16.312359 | 15.743707 | 15.719916 | 15.241130 |

Regression Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Dependent variable: w* | | | |
|  | Model-1 | Model-2 | Model-3 | Model-4 |
|  | | | | |
| Intercept | -2.781 | -24.389\*\*\* | -19.834\*\*\* | -10.633 |
|  | (8.851) | (6.980) | (7.072) | (6.769) |
| afram |  | 26.719\*\*\* | 25.902\*\*\* | 25.239\*\*\* |
|  |  | (3.348) | (3.937) | (3.752) |
| age | 1.849\*\*\* | 1.916\*\*\* | 1.842\*\*\* | 1.688\*\*\* |
|  | (0.424) | (0.402) | (0.468) | (0.459) |
| agesq | -0.018\*\*\* | -0.019\*\*\* | -0.018\*\*\* | -0.016\*\*\* |
|  | (0.005) | (0.005) | (0.005) | (0.005) |
| asian |  | 20.485\*\*\* | 19.761\*\*\* | 20.576\*\*\* |
|  |  | (3.206) | (3.901) | (3.839) |
| child0 |  |  | -3.186\*\* | -2.439 |
|  |  |  | (1.473) | (1.490) |
| child1 |  |  | -3.875\* | -0.577 |
|  |  |  | (1.997) | (1.961) |
| child2 |  |  | -7.644\*\* | -6.217\*\* |
|  |  |  | (3.093) | (3.002) |
| child3 |  |  | -2.069 | -0.338 |
|  |  |  | (2.301) | (2.167) |
| child4pl |  |  | -3.060 | -1.062 |
|  |  |  | (1.970) | (1.884) |
| divorced |  |  |  | -11.193\*\*\* |
|  |  |  |  | (4.157) |
| ed\_MA |  | -5.356\*\* | -4.287\* | -2.323 |
|  |  | (2.384) | (2.415) | (2.316) |
| ed\_Phd |  | -1.695 | -0.521 | 1.810 |
|  |  | (2.075) | (2.110) | (2.063) |
| ed\_Pro |  | 0.797 | 1.956 | 3.329\* |
|  |  | (1.831) | (1.856) | (1.842) |
| ed\_other |  | -18.135\*\*\* | -16.982\*\*\* | -13.449\*\*\* |
|  |  | (2.245) | (2.235) | (2.133) |
| fedgov |  |  |  | 6.069\*\*\* |
|  |  |  |  | (1.875) |
| female |  | -2.906\*\*\* | -2.932\*\*\* | -2.534\*\* |
|  |  | (1.050) | (1.053) | (1.055) |
| forprof |  |  |  | -0.230 |
|  |  |  |  | (1.476) |
| hisp |  | 2.962 | 2.751 | 2.753 |
|  |  | (2.198) | (2.204) | (2.038) |
| locgov |  |  |  | -5.357\*\*\* |
|  |  |  |  | (1.913) |
| married |  |  |  | -11.730\*\*\* |
|  |  |  |  | (3.893) |
| nevermar |  |  |  | -8.576\*\* |
|  |  |  |  | (3.982) |
| nonUSborn |  |  |  | -0.716 |
|  |  |  |  | (2.503) |
| nonprof |  |  |  | -5.823\*\* |
|  |  |  |  | (2.585) |
| othernonw |  | 24.293\*\*\* | 23.383\*\*\* | 24.497\*\*\* |
|  |  | (4.317) | (4.909) | (4.916) |
| stagov |  |  |  | -5.292\*\*\* |
|  |  |  |  | (1.876) |
| union |  |  |  | -0.600 |
|  |  |  |  | (1.675) |
| white |  | 22.167\*\*\* | 21.343\*\*\* | 22.251\*\*\* |
|  |  | (2.309) | (3.164) | (3.010) |
| wiwowed |  |  |  | -17.187\*\* |
|  |  |  |  | (6.755) |
|  | | | | |
| Observations | 989 | 989 | 989 | 989 |
| R2 | 0.040 | 0.123 | 0.126 | 0.174 |
| Adjusted R2 | 0.038 | 0.114 | 0.112 | 0.153 |
| Residual Std. Error | 16.453 (df=986) | 15.793 (df=977) | 15.807 (df=973) | 15.441 (df=963) |
| F Statistic | 23.528\*\*\* (df=2; 986) | 113.348\*\*\* (df=11; 977) | 83.500\*\*\* (df=15; 973) | 57.724\*\*\* (df=25; 963) |
| BIC | 8363.8 | 8335.81 | 8361.0 | 8373.46 |
|  | | | | |
| Note: | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | |