Assignment-1

By.: Szabados Dávid (ID:2302806)

Introduction

While building models for prediction analysis, individuals may unconsciously use too many predictor variables, which may lead to overestimations in the actual data. Using too few predictor variables may lead to underestimations, then which one is better? During this assignment, the *cps-earnings* dataset will be used, to build four basic prediction models, from very simple ones to more complex ones. The models will be compared based on their performance and their complexity, to get a general understanding about the pros and cons of simpler and more complex models. The significance of the variables will not be discussed, as the main goal of the assignment is to show the differences of the models, and which one should be used.

Data & Data Cleaning

Personal Care and Service occupations were selected from the dataset as the base of the models. Most variables, while being represented as numbers, are qualitative variables, with many possible values. These will be transformed into dummy, True/False variables:

- grade92: True if the value is 41 or more *(has an associate degree or higher)*
- unionmme: True if the person is a union member.
- Private: True if the person works at a private firm.
- race: True if the person's race is white.
- marital: True if the person has been married at least once.
- prcitshp: True if the person is a native USA citizen.
- sex: True if the person's gender is male.
- lfsr94: Ture if the person is currently at work (not absent).

Other quantitative variables that have been used for the models:

- age: The age of the person.
- ownchild: The numbers of own child in the primary family.

Using these variables, the four models have been constructed, which can be seen on **Appendix1**. For the four models: 1, 5, 9, 15 predictor variables were used from the data, to be able to compare their performances from the least to the most complex one.

Model Comparison

The performances of the models will be compared based on:

a) RMSE in the full sample
b) cross validated RMSE
c) BIC in the full sample

RMSE is the square Root of the Mean Squared Error, in other words, the root of the numerator of the R-squared. for our models, we want this value to be as low as possible, but not too low

for the live data. A low RMSE means a good performance of the model. If the model has too few predictors in it, the RMSE will be large. Adding more variables will lower the value of the RMSE, but after a point, the amount of these variables will be too much and the value of the RMSE will start rise in the live data, thus we want to find a model in the middle. The model with few predictors is underfitted, it is unable to capture the relationship between the input and output variables accurately, while the model with too many predictors is overfitted, as the model can only capture very accurately the initial training data, but not the possible new, live data, which is the model should be made for.

The RMSE is calculated first with the full sample, and after that the sample is sliced into four equal parts and will be cross validated.

The BIC, or the Bayesian Information Criterion, also helps for the model selection. Just like at RMSE, a lower BIC value means a better model. It is mostly used for a finite set of models, as four models will be compared, using BIC should help the selection, as it can be calculated relatively very easily. BIC also includes a penalty term for more variables in the model, thus it helps more to properly select the model and avert to select one, which would be overfitted for the live data.

a.: RMSE in the full sample

```
Model 1: 10.209360676188265
Model 2: 9.990962087286785
Model 3: 9.985220830785064
Model 4: 9.978600826062083
```

From the RMSE values in the full sample, we can see, that the first model's value is the highest amongst the four and the fourth model has the lowest, which shouldn't be surprising, as using only the original data, more complex models will output a lower RMSE value.

b.: cross validated RMSE

	Model1	Model2	Model3	Model4
Fold1	9.994657	10.376396	10.201266	10.289894
Fold2	10.811281	10.340119	10.181047	9.861922
Fold3	9.314395	9.608586	10.174534	9.456408
Fold4	10.643370	9.600842	9.349573	10.260475
Average	10.190926	9.981486	9.976605	9.967175

For the cross validation, on average the first model bares the highest RMSE and the fourth model has the lowest, just like in the full sample.

c.: BIC in the full sample

BIC	32574.27	32419.66	32448.17	32492.67

From left to right, model one trough four's BIC values can be seen. Contrary to RMSE, model two bares the lowest value amongst every model. As model three and four has much more variables, the penalty term from BIC is also bigger, thus the reason for a higher value, while model one does not have enough predictors to be able to predict the data accurately.

Choosing the Model

Overall model four brought back the lowest RMSE value all the time, so logically, that one should be chosen. Before we rush into some hasty conclusion, we need to remember, that the low RMSE values are from the fact, that only the original data has been used and the model may not yield the same results on the live data, but higher values only. The BIC value should also made into account, as it penalizes models with too many variables and helps us to avoid choosing overfitting models. Model two's value is the lowest, it does not have too many variables, but is also able to catch the relationships somewhat accurately. In my opinion model two should be used for the live data. It has a much better RMSE value than model one, which failed to accurately catch the relationship. It also has a lower BIC than model three and four, those models have a higher chance to be overfitted if used on a live data.

Appendix1

				Dependent variable: w
	(1)	(2)	(3)	(4)
Intercept	10.769***	9.456***	9.284***	10.132***
	(0.466)	(0.537)	(0.951)	(1.494)
Private[T.True]			-0.455	-0.292
			(0.608)	(0.636)
age	0.076***	0.044***	0.049***	0.049***
	(0.011)	(0.014)	(0.015)	(0.016)
age:grade92[T.True]		-0.008	-0.008	-0.006
		(0.026)	(0.026)	(0.025)
age:ownchild				-0.009
				(0.013)
age:unionmme[T.True]				0.007
				(0.042)
grade92[T.True]		3.712***	3.713***	3.668***
		(1.049)	(1.051)	(1.051)
lfsr94[T.True]				-1.047
				(0.995)
marital[T.True]		1.343***	1.156***	1.131***
		(0.333)	(0.339)	(0.348)
ownchild			0.177	0.261
			(0.158)	(0.572)
ownchild:race[T.True]				0.392
				(0.356)
age:unionmme[T.True]				0.007
				(0.042)
grade92[T.True]		3.712***	3.713***	3.668***
		(1.049)	(1.051)	(1.051)
lfsr94[T.True]				-1.047
				(0.995)
marital[T.True]		1.343***	1.156***	1.131***
		(0.333)	(0.339)	(0.348)
ownchild			0.177	0.261
			(0.158)	(0.572)
ownchild:race[T.True]				0.392
				(0.356)
ownchild:unionmme[T.True]			0.398	-0.048
			(0.508)	(0.575)
prcitshp[T.True]				0.236
				(0.368)
race[T.True]			0.525	0.233
			(0.324)	(0.391)
sex[T.True]		3.006***	3.042***	3.019***
		(0.404)	(0.405)	(0.405)
unionmme[T.True]				0.861
				(1.893)
Observations	4350	4350	4350	4350
R^2	0.010	0.052	0.053	0.054
Adjusted R^2	0.010	0.051	0.051	0.051
Residual Std. Error	10.212 (df=4348)	9.998 (df=4344)	9.997 (df=4340)	9.997 (df=4334)
F Statistic	45.260*** (df=1; 4348)	42.141*** (df=5; 4344)	26.308*** (df=9; 4340)	16.461*** (df=15; 4334)
BIC	32574.27	32419.66	32448.17	32492.67
Note:				*p<0.1; **p<0.05; ***p<0.01