**Assignment-1**

**By.: Szabados Dávid**

**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:

1. RMSE in the full sample
2. cross validated RMSE
3. 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. 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, 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 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.

a.: RMSE in the full sample

A képen szöveg, Betűtípus, tipográfia látható

Automatikusan generált leírás

b.: cross-validated RMSE

A képen szöveg, képernyőkép, Betűtípus, szám látható

Automatikusan generált leírás

c.: BIC in the full sample



**Appendix1**