PQHS 471 Midterm Write Up

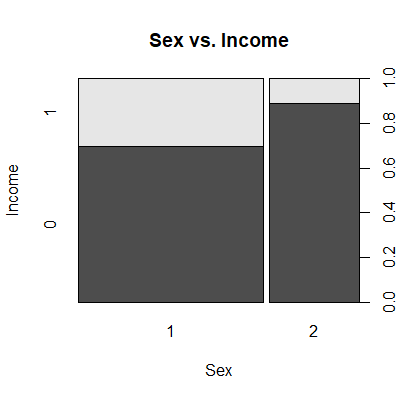
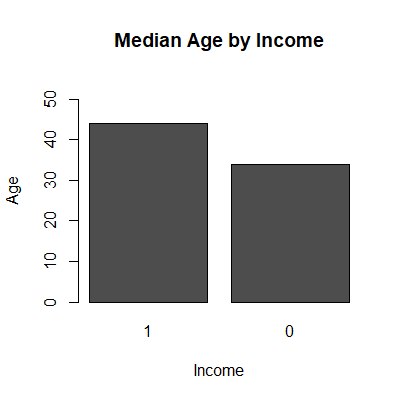
March 8, 2018

Rachel Sabol

**Exploratory analysis**

A brief summary of some measures taken to perform exploratory analysis on the training data (census\_train\_val) are given in the attached code under section header “Sample code from exploratory analysis”. A summary table of variables and some relevant variable information is displayed below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable name** | **Variable type** | **Levels (if applicable)** | **Missing observations** |
| age | continuous |  |  |
| workclass | categorical | 8 | 1404 |
| fnlwgt | continous |  |  |
| education | categorical | 16 |  |
| education.num | continuous |  |  |
| marital.status | categorical | 7 |  |
| occupation | categorical | 14 | 1411 |
| relationship | categorical | 6 |  |
| race | categorical | 5 |  |
| sex | categorical | 2 |  |
| capital.gain | continuous |  |  |
| capital.loss | continuous |  |  |
| hours.per.week | continuous |  |  |
| native.country | categorical | 41 | 437 |
| income | categorical | 2 |  |



**Figure 1: Sample exploratory plots**. The right plot displays a representation of the categorical variable sex and the outcome, income. The plot on the left displays a representation of the numeric variable, age, and the outcome, income. The sex value of 1 indicates male and the sex value of 2 indicates female. An income value of 0 denotes income <50K and a value of 1 denotes >50K.

An analysis of correlations between numeric variables was performed, with no strong associations detected (no |correlation coefficient|>0.2). Each variable was plotted against the outcome variable, income. Sample plots for a categorical variable (sex) and a continuous variable (age) are displayed in Figure 1. Plots were utilized for visualizing the data and for preliminary assessment of potential relationships between variables and the outcome, income.

In order to make the categorical variables easier to interpret, variables were recoded numerically. In cases where an order to recoding was intuitive, for example in completion of education, levels were ordered. Entries coded as “?” were recoded as NA to be easily managed in downstream analysis. In the case of the variable native.country, categories were collapsed on a regional basis in order to reduce the number of levels and increase interpretability. Regions chosen included “United States and Canada”, “Latin America”, “South America”, “Western Europe”, “Eastern Europe”, “South Asian”, “East Asian”, and “Southeast Asian”. The regional assignments were determined geographically.

Missing values comprised less than 10% of the data, and variables were not omitted based on missing data. Individuals with missing data were excluded from analysis where models could not tolerate missing data. Since missing values were only observed in categorical variables, imputation was not considered as a reasonable option.

**Model development**

In order to develop and train a model, the entire set of training observations was split into a training and validation set. The data was randomly split using the sample function, with 60% of the observations denoted as training data and 40% denoted as validation data. This resulted in a total of 15,000 observations in the training data and 10,000 observations in the validation set.

The following models were tested, but are not all included in the presented code (if the original development code is desired, it can be provided): simple logistic regression, multiple logistic regression, forward and backward stepwise variable selection for logistic regression, logistic ridge regression, logistic lasso regression, logistic elastic net regression, linear discriminant analysis, generalized additive models (including polynomial variable modeling, splines, and local regression), simple decision trees, and random forest analysis including boosting.

The final model that was chosen was a generalized additive model for classification. The generalized additive model performed better than multiple logistic regression in predicting income in the validation set and allowed for the inclusion of non-linear modeling of some numeric variables. The variables age and hours.per.week were modeled with smoothing splines, after testing the effects of non-linear modeling on numeric variables. The variable capital.gain was modeled with local regression. All variables were included in the final model, as stepwise backwards and forwards regression did not result in any improvement of model predictions on the validation set.

The model correctly predicted 85.83% of the observations in the validation set and 78.39% in the testing set (provided by the file census\_test and denoted by the variable census\_test). The prediction tables for each set are provided below.

