Rosalie Day – Data Wrangling, Capstone 1
Prediction of Income from Education and Sector Status
November, 2019

Problem statement

For the past decade, the conventional wisdom has been jobs that pay a living wage require a college degree. The claim that having a bachelor's degree is necessary for getting a "good" job has been around for decades. In the current discourse, the bachelor degree is the symbolic threshold of being able to adapt to workforce needs in the digital age going forward.

This study explores the predictors of whether a person's income was less or greater than $50,000, a selected "good job" threshold corresponding to the 1994 dataset. The focus is on education levels and employment sector and status as predictors. This study does not explore race, gender and marital information, although these will be used for classification of income outcomes in the full predictor set of attributes.

Data Wrangling

The Adult Income data set is an extraction of the 1994 US Census. The data was downloaded in csv file format from the Kaggle website https://www.kaggle.com/uciml/adult-census-income into a Panda's dataframe.

The data set included 48842 samples, composed of mainly categorical variables and no missing values. Notable exceptions to the categorical class were age and hours-per-week. The variables, capital gains, capital loss and final weight variables were excluded as they did not contribute to the intended analysis. The working data set included 10 predictor variables: age, workclass, education/educational-num, marital-status, occupation, relationship (to family), race, gender, and hours-per-week, native-country; and the income target variable.

There were few modifications for analysis. The income target variable, which was an object, was converted into the integers 0 and 1 for equal or less than and greater than $50, respectively. There were two education variables containing the same information: explanatory labeling in string objects; and integers which corresponded to the string values. These levels did not correspond to education years completed. The choice of string or integer was made on a case-by-case basis depending on the use.

The variables were largely self-explanatory except for income (previously explained) and work class. The work class variable includes sector information, public, private, self-employed and an 'other' level which may be non-profit. Race, gender, relationship (to family) and marital-information were not explored, alt hough these will be used for classification of income outcomes.

Descriptive statistics, distribution and count plots for this sample, included education levels, ages, employment sector and status, hours worked per week and income. The sample ages ranged from 17 to 99. Other variables, for example, the distribution of country of origin were highly skewed, where 43,832 were listed as from the US.