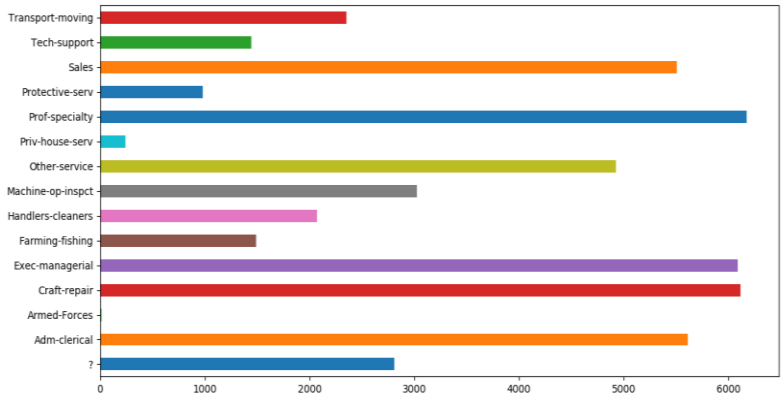
The Data

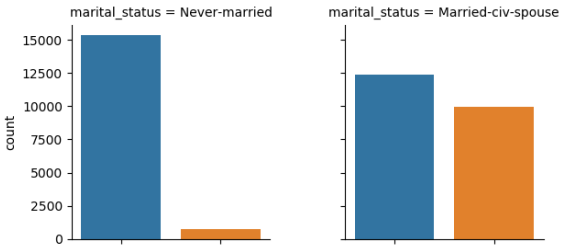


The input data consists of almost 50,000 individuals, each with approximately 20 attributes, including age, occupation, sex, race etc. The sample population is predominantly white, American, privately employed and the majority hold a high-school or college degree. The sample consists of approximately two thirds males, but has an even spread of ages and occupations, the latter of which is shown in the example plot to the right. Importantly, the data includes whether each individual's income is above 50k, with a third of the data meeting this condition. The goal is to build a predictive model for this income using a combination of the other properties of the data.

The Model

The strategy used to model this property is supervised machine learning, where the goal is to find a function that maps the observed properties (X = age, race, occupation etc.) to the labels (Y = Income). Since the labels are non-numerical, this is called a classification task, where we must learn how to classify an individual into an income category from its observed other properties.

A good classification method is a decision tree constructing algorithm to build a flowchart that successively divides the data on its observed attributes, such that each division provides us more clarity about the possible labels (income). At each level of the tree, the data is divided by the attribute that maximizes information gain. The Python programming language has extensive libraries that can achieve this (namely pandas and sklearn). In this method, 80% of the sample was used as training data to build the model, and the remaining 20% was used to test the model's accuracy.



The model built is shown as a decision tree, up to 3 levels, and is shown below. It is important not to go into excessive detail, as the model can fit the data too closely and lose predictive power for new data. Interestingly, the most informative property for income was marital status; individuals who were unmarried were overwhelmingly more likely to have an income below 50k. The graph above shows incomes for the two main marital statuses, where the orange columns represents those with an income above 50k. Other properties that were important were capital gain, age and education level. Some properties such as occupation, nationality and race were found to have little effect, possibly due to either the lack of variation or large number of possible categories.