Decision Tree to Predict Income Over $50K

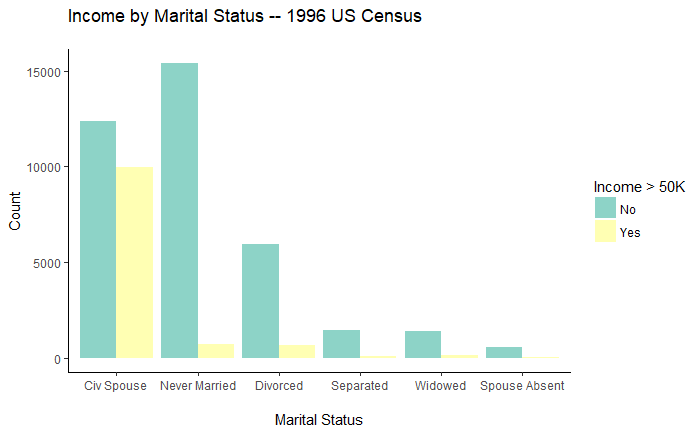
## Results

The final decision tree model correctly predicts income above or below $50K 85.8% of the time. It uses a branching structure to divide observations sequentially into 11 homogenous groups that can be classified accurately by income. The model creates groups by splitting observations on the criteria below (listed in order of importance):

* Marital status
* Capital gain
* Highest education level
* Occupation
* Age
* Capital loss

The first and most important split in the decision tree is civil spouse vs. other marital status. As seen in Figure 1, people with a civil spouse are split almost evenly by income. By contrast, the majority reporting some other marital status make less than $50K.

Figure 1. Income Over $50K by Marital Status



## Methods

First, I explored the distributions of each variable and collapsed categories of large factor variables. For example, I grouped countries into United Nations world regions. After investigating two-way associations with the target variable, I generated four candidate models: two logistic regressions and two decision trees. Methodology for the logistic regression models is described in a separate .rmd file.

The final selected decision tree balances low classification error with simplicity and interpretability. To create it, I first generated a full model using all available variables. I determined optimal pruning using the complexity parameter and cross-validated error. The lowest error corresponded to a tree with 25 branches, but an elbow plot revealed that an 11-branch tree performed almost as well. I tested both models on new data and judged that the tradeoff between simplicity and error favored the 11-branch model.