Predicting Salaries from Census Data:

An exercise in Data Science for RTI International

Court Haworth
February 27th, 2018

For this project, I was tasked with building a model to predict whether an individual made over or under $50,000 a year based on a number of demographic variables obtained from census data. This report details the approach I took to the project as well as my results. To start with, the data was stored in a database, spread through nine tables, so I transformed the data into a single table in order to be processed all in one place. Next, there were a number of individuals with missing demographic information. The information that was missing was either information about the country they were from, or information related to their occupation. I used a standard procedure to handle this missing information, and replaced the missing data with the most common value of those variables (i.e. the United States, the most common country in the data).

Next, I did some visualizations of the data. This is a good step to try to identify any underlying relationships in the data. As you can see in Figure 1 at the bottom, the education level of an individual tells a lot about the salary of the individual. Individuals with a Doctorate or Masters are more likely to be making more than 50k than they are to be making less, while someone whose highest education was 11th grade is very unlikely to be making more than 50k. This information was not explicitly hardcoded into the models, but is a nice sanity check that there are underlying relationships in the data that might be able to help make accurate predictions, and the models should pick up this relationship implicitly as well.

I built and evaluated a series of models, each of which with a goal to predict whether an individual made more, or less, than $50,000 a year. The three model types are all standard methods when it comes to binary classification tasks, or problems where you are trying to classify an individual into one of two classes (under 50,000 or over 50,000). I evaluated these three types of models, Logistic Regression, Support Vector Machines (SVM), and Random Forests, by comparing their respective prediction accuracy on unseen data. After varying different parameters of the model, in the end I settled on a SVM model that achieved 87% accuracy on unseen data. This means that nearly 9 times out of 10, the model would be able to accurately predict whether an individual model makes more or less than $50,000 a year. This performance is very satisfactory and if this model was implemented it could create a number of effective use cases for our business and help drive our future success.

Figure 1.