Methodology

The task at hand was to retrieve, flatten, and analyze data relating to an individual’s income and various demographic data collected for each individual. More specifically, we were asked to explore the data and determine what factors influence whether or not an individual was likely to make over fifty thousand dollars per year.

To do this, I used a SQL query to flatten the data, and R for analysis. I decided to use a logistic regression model for predictive purposes. In order to create better separation and increase significance of certain variables, some binning of variables was required, notably for capital gain and loss, education categories, and countries, which were binned into developing and developed. The model ended up delivering a strong predictive power on training and test data sets, achieving roughly a 78% accuracy for both sets. Improvements can certainly be made to the model, and certain assumptions were not checked.