**RTI CDS Analytics Exercise 01 – Kelsey Campbell**

Category values for the 1996 US Census data were combined and many of the variables re-binned for inclusion in a predictive model for whether an individual makes over $50,000/year. Three first-order logistic regression models were tested, the first being a simple model only accounting for age and race. The second model included most of the information available, such as age, marital status, education, occupation, whether the individual is from the US or not, hours worked per week, and capital gains or losses. The third model used a feature selection algorithm aiming to reduce the second complete model to only the important measures. Comparisons on both a 70% training and 20% holdout sample concluded that the second complete model was the most accurate on percentage predicted correct as well as the highest on area under the ROC curve, which can be used to compare how well a model accurately predicts 0’s and 1’s. This final model predicted a 10% test dataset with 84.3% accuracy.

Final model estimates were generated from the entire dataset to produce odds ratios, showing the comparative effect of each covariate category to that of the reference group. I chose to focus only on the “changeable” factors in the data, which I determined to be area of occupation and level of education. Results for occupation were largely uninteresting, so I have displayed odds ratio results for education only. You can see from the figure that furthering your education significantly (*p* < .001) increases the odds of earning over $50,000 compared to the odds of completing less than a high school degree. Completing an advanced degree had the largest effect, with that group being 11 times more likely to make over $50,000 than the group with less than a high school diploma, holding all else constant. So stay in school kids!

**Odds Ratios by Education Category, Relative to “Less than a High School Degree”**

