RTI International
Census Data Write-up
Samuel Eure
June 4, 2019

In this investigation, census level data of adults (sample size = 48842) was explored in with the intentions of producing a model capable of predicting whether an individual earns ≥ $50,000 per year based on their demographic information. All complete data entries (sample size = 45222) were taken from this data set and used to train a machine learning model (gradient boosting with logistic loss) capable of making binary classifications decisions similar to those of this investigation. This model works by taking advantages of differences in the demographic features (such as age, years of education, etc.) between the groups of workers who earn ≥ $50,000 per year and those that do not; such differences can be seen in the figures below on the left. This model also works by searching for groups of workers that are difficult to classify, and give special attention to these workers while classifying them. After tuning, this model is capable of predicting whether or not an individual earns ≥ $50,000 per year based on their census level data with an accuracy of 87%.