Rosalie Day – 2ndProposalCapstone1
Prediction of Income from Education and Sector Status
November, 2019

The Purpose

News is filled with stories about how opportunities for employment are increased by having a college degree. In fact, often articles imply that jobs paying a living wage in the US require a college degree. Some articles may refer to exceptions in which certain vocations, persons skilled at building trades for example, are well compensated, while the norm is asserting that a 4-year college degree is necessary to get a "good job." This study focuses on exploring education and employment sector and status as predictors of whether a person's income is less or greater than $50,000.

The use for this information is for describing the frequency of educational levels, comparison with the employment status and sector, and as a predictor of income for informing public policy. Education levels are explored along with sectors, which includes levels of: government; private; self-employed; and other.

The Data

The Adult Income dataset was extracted from 1994 U.S. Census Bureau database and appears in the Kaggle competition data sets. The income threshold of $50,000, reflected in this data set, is $87,000 in 2019 dollars. However, it must be noted that wages have not tracked the inflation or discount rates this period. It is only a snap shot and the relative importance of the predictor variables will likely change, however, the concept of educating the workforce of the future seems enduring.

The first step in the analysis will provide context in form of descriptive statistics for this sample, including education levels, ages, sector and hours worked per week.

Inferential Statistics

The second step will test the difference in threshold income between having specifically a bachelor's degree, and no graduate degrees, with not having a bachelor's degree. The correlation between education level and work sector ("workclass" variable) and hours worked will be tested.

Prediction Models

The third step will predict using all the variables on a number of supervised machine learning models and compare performance. The top performing model will then be used on a subset variables chosen for privacy to what happens to performance.

Deliverables

Deliverables will include interim reports, at a minimum - statistics, midway milestone and final - and codes and data visualization in an updated markdown document for each of 3 steps. The final deliverable will include a written results summary and a slide deck that captures interesting and significant results for the entire project.

More research

It would be interesting to do the classification modeling on data at 10-12 year increments to explore whether the education levels and sector profile changed for the income threshold similar to this one. However, replication of weighting and treatment of the variable ‘workclass”is not straightforward.