Capstone Project 1 Suggestion

Predicting Income Percentile Based on United States Department of Commerce Current Population Survey

Overview

In this project I will use data obtained from the "2007 Department of Commerce Current Population Survey" (for which the raw un-aggregated dataset is readily available for academic research) to predict a person's income percentile. The dataset consists of approximately 200,000 individual records, from which approximately 100,000 belongs to working adults. The dataset includes about 700 sociological markers (features), but only 46 have less than 30% missing values.

Goals

1. Predict the annual income percentile of an household based on sociological markers
2. Identify a relatively small important set of markers which will be readily available for the clients and which will be able to predict income percentile

Potential Client

The data will be useful for credit companies in the risk assessment stage. While considering a long term loan, being able to predict the the income percentile and not just depend on current wages, which can be transient anomalies, can be very useful. There is also the potential of predicting students future income base on their area of study.

Specific clients included companies like: Ernest (https://www.earnest.com/) and Upstart(https://www.upstart.com/)

Datasets:

I will try to base the project on the 2007 Department of Commerce Current Population Survey" as I have complete access to all of the microdata and the number of records is not overwhelming.

Should this data prove insufficient I will try to use the census bureau Annual Community Survey(ACS) public use microdata sample, which is much larger in scope but less economic oriented.

Supplementary queries of aggregate features ,if needed, will be done using the census bureau API .

Methodology

As the samples per state are relatively small (few thousands) I will first try to aggregate states with similar median income (and checking that the distribution of income is similar for all aggregated states). After a first filtering of attributes which seems highly irrelevant a random forest learning algorithm will be used to make the predictions.

Deliverables

Deliverable will include detailed explanatory statistical analysis of the data, which will include a slide deck with graphical results. The code for the classifier and a detailed report about statistical findings.