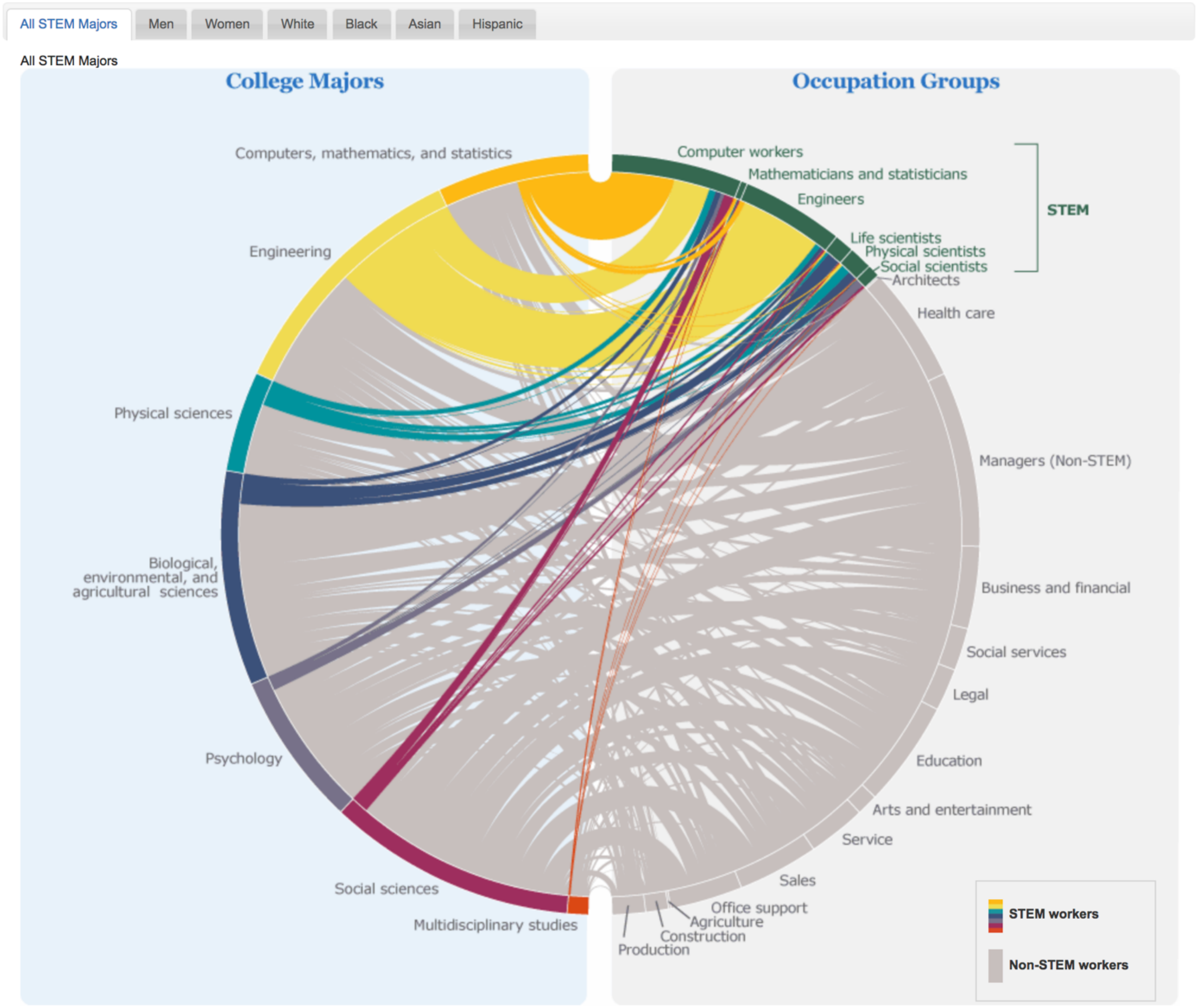
# Predictive Analytics on Achieving an Occupation

Danielle Saunders
Nov 4, 2018

In my preliminary proposal, I wanted to create an alluvial visualization of paths from college majors to occupation. Luckily, I found that the Census Bureau beat me to it in 2014, and I feel this is recent enough that I do not want to repeat it. It does not break down the occupational categories as far as would be useful, but I have still decided not to make this visualization. A similar visualization I thought of but which I will put on hold is showing the path between occupation and industry. That way it will be possible to see the breakdown of industries each profession works in.



https://www.census.gov/dataviz/visualizations/stem/stem-html/?#

Instead I will focus on predicting occupation based on features from the ACS PUMS data that an individual can control. This way, for young adults who have a career in mind, I can tell them which features are predictive of that career, and in which way. I would like to produce actionable recommendations, however, all I can ever say from this analysis which features produce the most signal about each occupation, and in which direction, based on what the sample has done with their lives. I cannot say it is the best way to achieve the occupation, but I do anticipate something useful to come out of this.

This has the potential to be large in scope, so I am making a concerted effort to keep it simple. I will start with California in 2017 only: there are 377,577 rows corresponding to individuals in a single excel sheet, about 1% of the US population. I want to look at a recent working generation, so I will filter by age under a threshold and employment status as employed (incl self-employed) (using employment status recode).

Here are most of the features available that an individual can generally control:

| ***Feature*** | ***Abbreviation*** |
|---|---|
| Area (PUMA code) | PUMA |
| class of worker (private/govt) | COW |
| ability to speak English | ENG |
| gave birth in last 12 months | FER |
| means of transportation to work | JWTR |
| travel time to work | JWMNP |
| number of times married | MARHT |
| educational attainment (see appendix A) | SCHL |
| usual hours worked per week | WKHP |
| weeks worked during past 12 months | WKW |
| field of undergrad degree 1 | FOD1P |
| field of undergrad degree 2 | FOD2P |
| marital status incl.- spouse present/spouse absent | MSP |
| presence and age of own children (female only) | PAOC |
| subfamily relationship (position in household) | SFR |

The target is SOC occupational code. There are 481 occupational categories in the SOC codes, with 23 major categories, but the PUMS uses the 2010 SOC so these codes are 8 years outdated, unfortunately. https://www.bls.gov/soc/soc_2010_user_guide.pdf

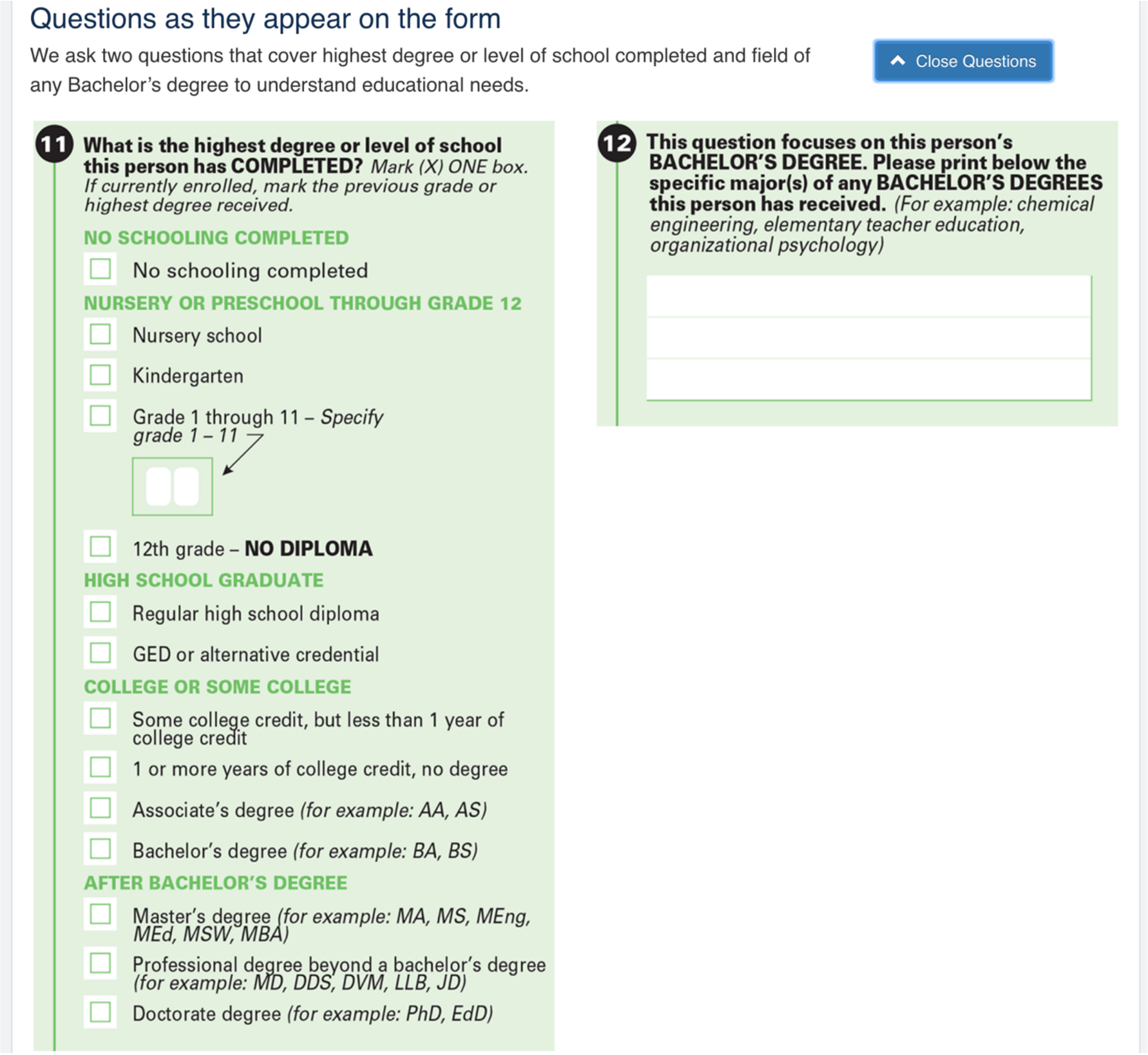
I could in the future explore significance of features of other members of the household on the person in question but it is currently out of scope. In fact, for this project, I will start with only 2 features: educational attainment, which is an ordinal variable, then a single dummy variable for married with children vs married without children. I do not think this will necessarily be predictive but I want to see.

Once I've looked at that, I will add 'field of undergrad degree 1', a categorical variable which will need to be split into 174 dummies.

I’ve already launched into curse of dimensionality territory for KNN so I will need to prepare for that with dimensionality reduction techniques. I believe I will use decision trees, boosting, logistic regression, and SVM, time permitting. These are all multiclass classifiers. If I need to pare down to binary classifiers for some reason I can look at a single occupation. I think I will owe my thanks primarily to this page of sklearn: http://scikit-learn.org/stable/modules/multiclass.html There is also lots of documentation for using the PUMS, which I will cite.

I want to focus on my strengths in this project so I will endeavor to keep it well documented, commented, and visualized, and emphasize data storytelling. I will also talk about the potential for this project as a product if I could make predictive analytics given a person’s specific demographics, and pending the usefulness of the findings. This would entail a different model for each combination of demographics I think. Even in the simplest form, this would be great as a web app so I’ll add that to the list of potential features if time permits.

Appendix A



## Questions as they appear on the form

We ask two questions that cover highest degree or level of school completed and field of any Bachelor's degree to understand educational needs.

**11** **What is the highest degree or level of school this person has COMPLETED?** *Mark (X) ONE box. If currently enrolled, mark the previous grade or highest degree received.*

**NO SCHOOLING COMPLETED**

□ No schooling completed

**NURSERY OR PRESCHOOL THROUGH GRADE 12**

□ Nursery school

□ Kindergarten

□ Grade 1 through 11 – *Specify grade 1 – 11* □□

□ 12th grade – **NO DIPLOMA**

**HIGH SCHOOL GRADUATE**

□ Regular high school diploma

□ GED or alternative credential

**COLLEGE OR SOME COLLEGE**

□ Some college credit, but less than 1 year of college credit

□ 1 or more years of college credit, no degree

□ Associate's degree *(for example: AA, AS)*

□ Bachelor's degree *(for example: BA, BS)*

**AFTER BACHELOR'S DEGREE**

□ Master's degree *(for example: MA, MS, MEng, MEd, MSW, MBA)*

□ Professional degree beyond a bachelor's degree *(for example: MD, DDS, DVM, LLB, JD)*

□ Doctorate degree *(for example: PhD, EdD)*

**12** **This question focuses on this person's BACHELOR'S DEGREE. Please print below the specific major(s) of any BACHELOR'S DEGREES this person has received.** *(For example: chemical engineering, elementary teacher education, organizational psychology)*