Predicting Occupation Based on Life Decisions

An agile approach

Danielle Saunders, Galvanize, Nov 2018

What are the strongest hypothetically controllable predictors of an individual's occupation?

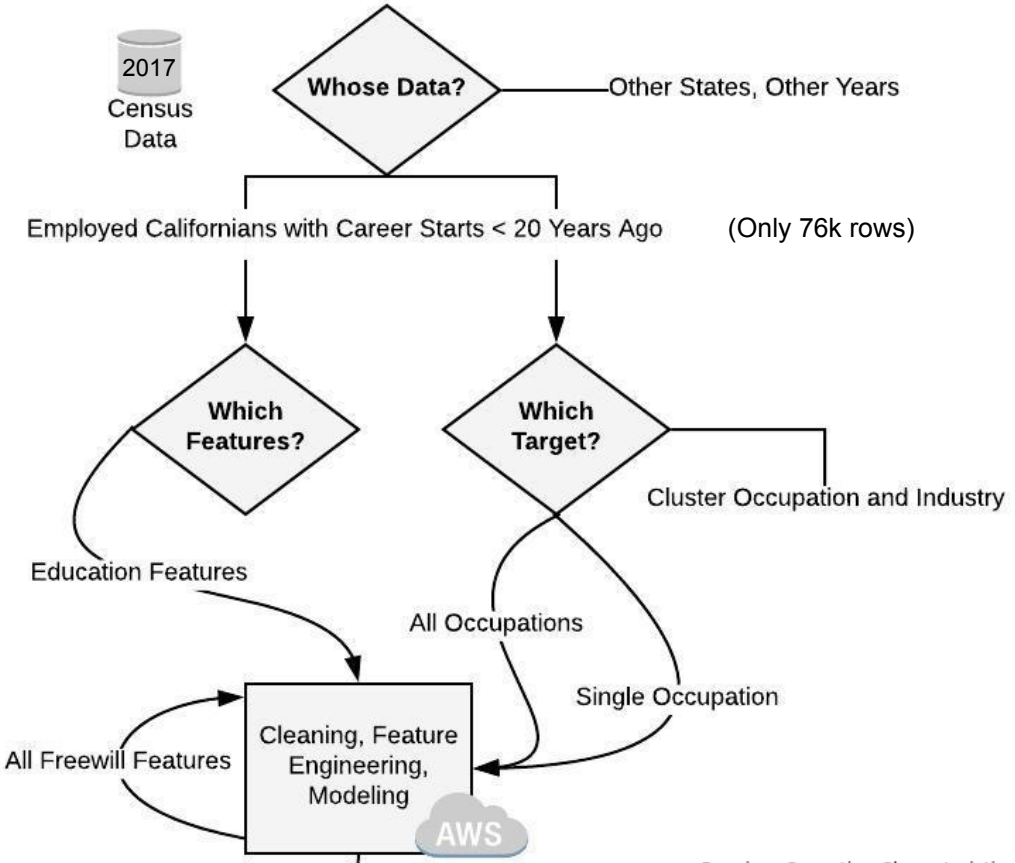
Such as:

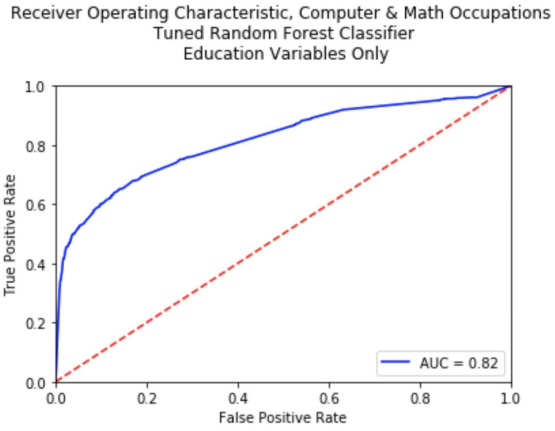
- College Major
- Educational Attainment
- Location
- Military Service
- More complex features

Demo Use Case: Computer Science & Healthcare Practitioner Occupations

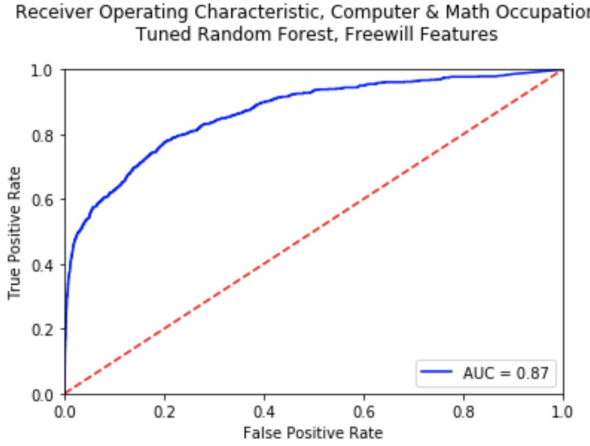
Source: US Census ACS PUMS dataset

Pandas scikit learn matplotlib jupyter ubuntu

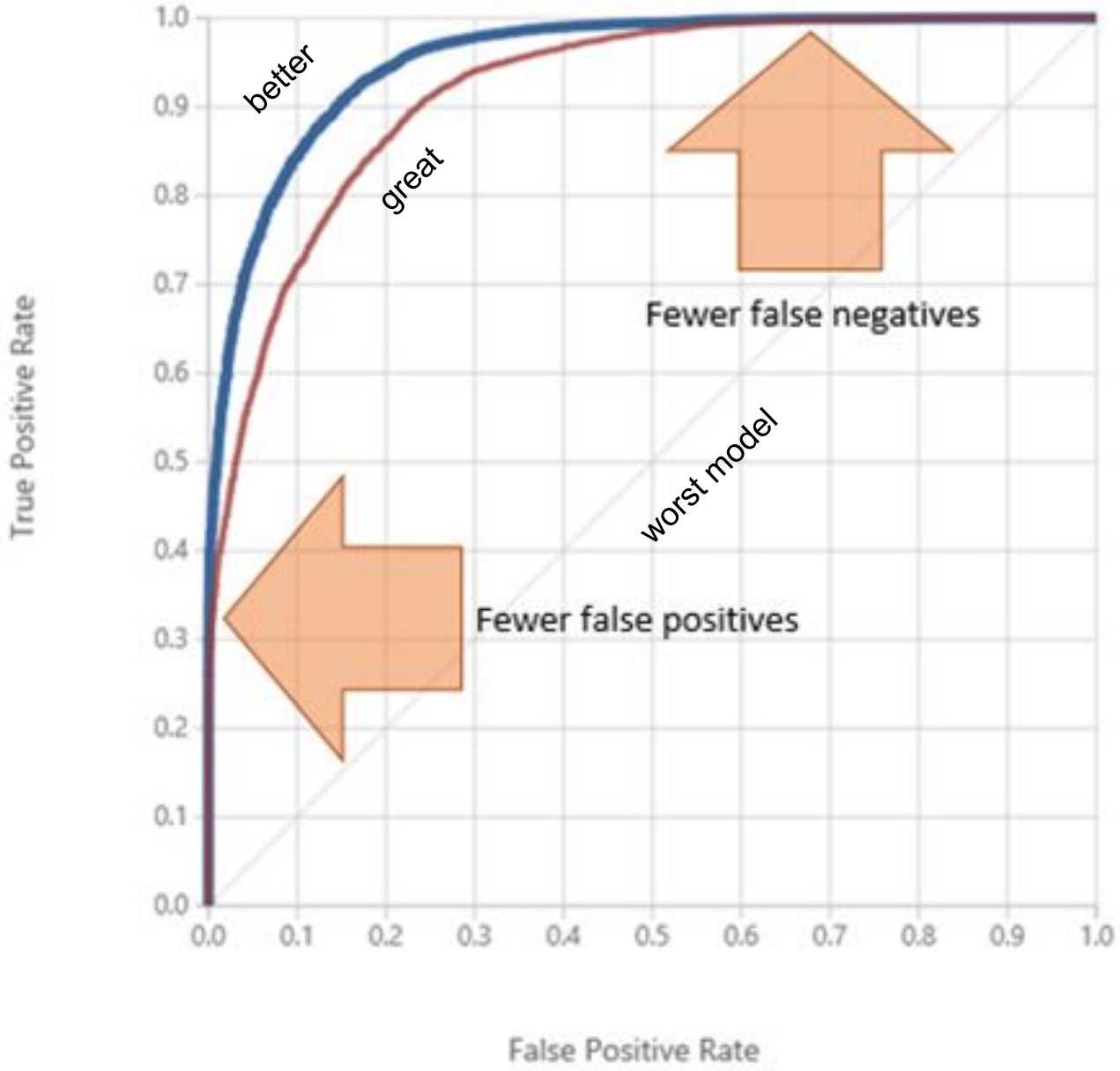




	Multiclass	Binary
Education Features	F1 .21 Acc .21	F1 .42 AUC .82
Freewill Features	/	F1 .37 AUC .87 Acc 96.92%

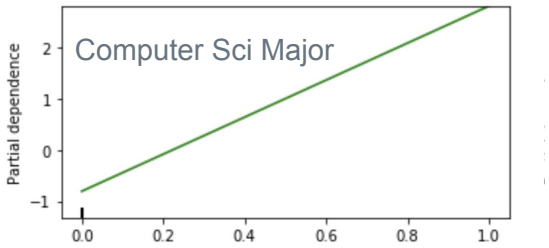


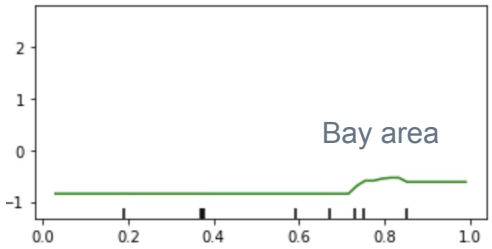
Interpreting ROC Curves

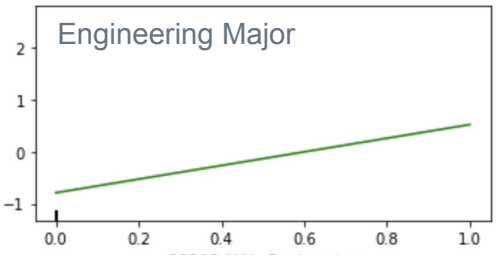


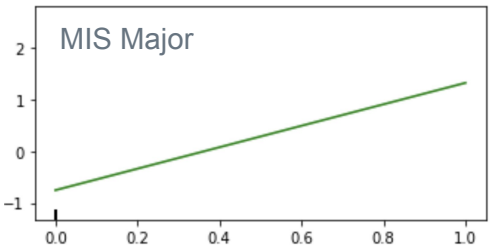
Models and Results: Computer Sci

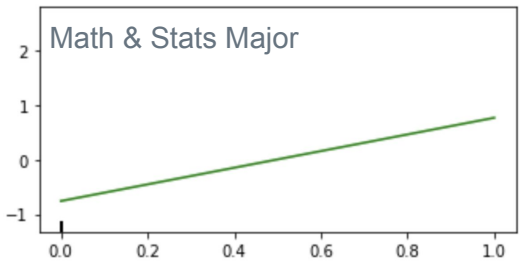
- 8 classification models, 10 CV
  - trees worked best
- focused on interpretability
  - embedded feature selection

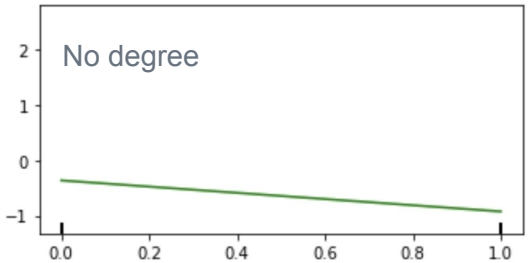


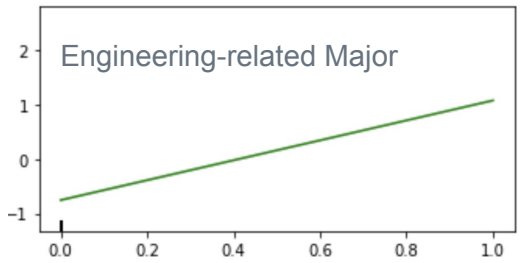










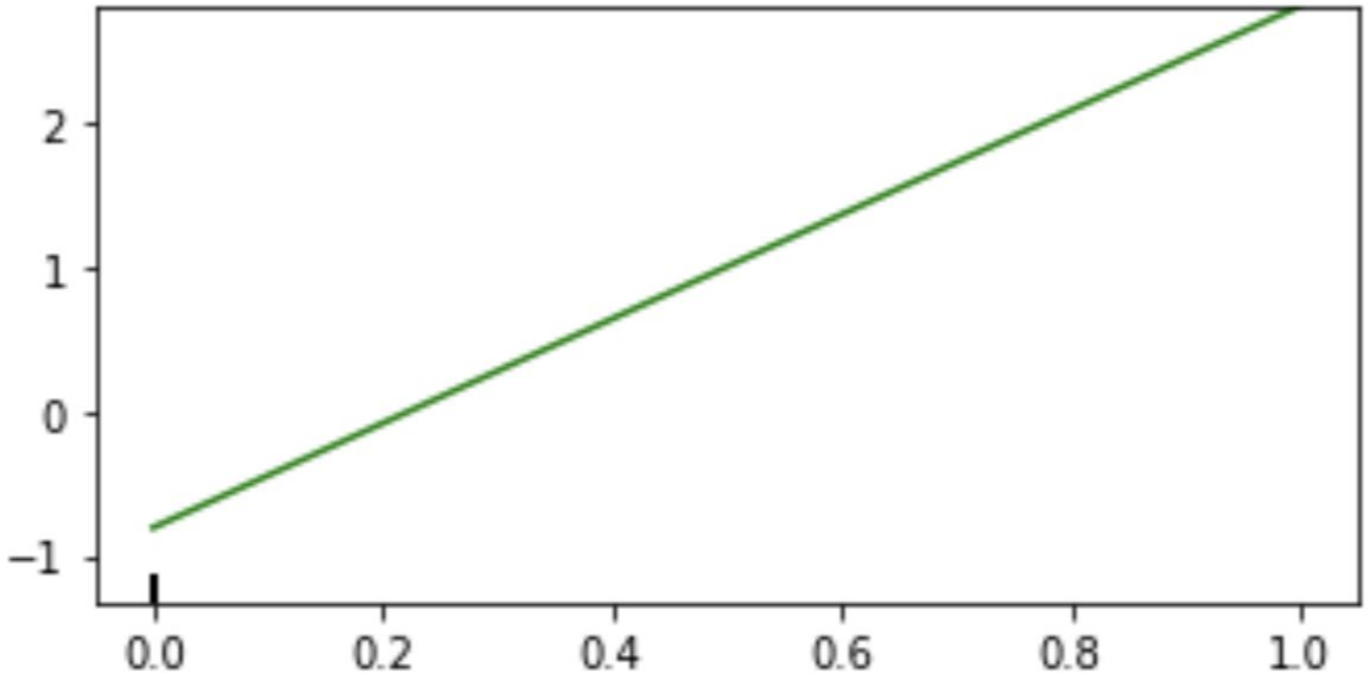


Interpreting Partial Dependence Plots

relative contribution to a positive prediction

At zero there is no average impact on class probability according to the model

Below zero, these values of x make predicting the positive class less likely



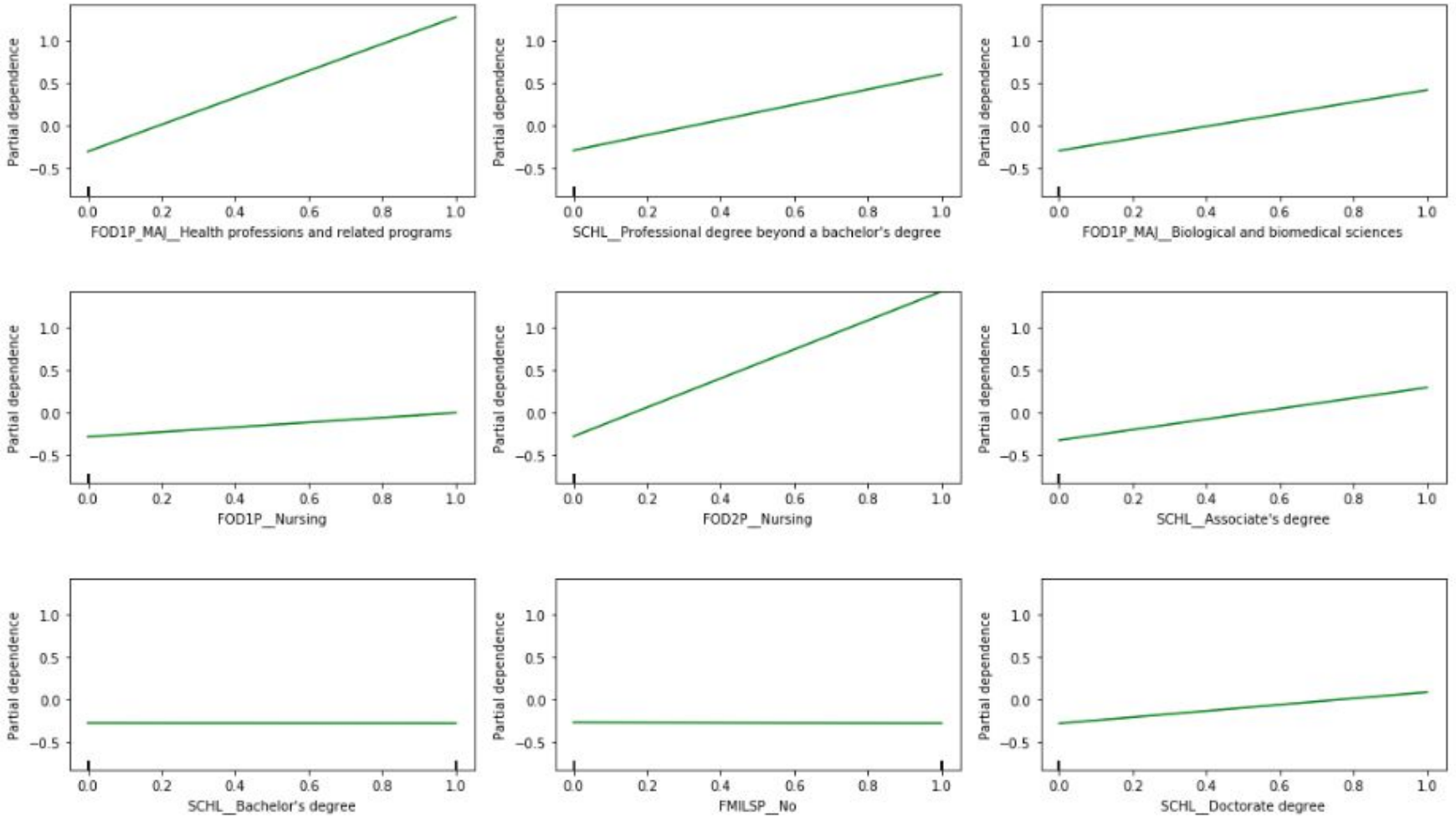
Suggested Paths to Programming

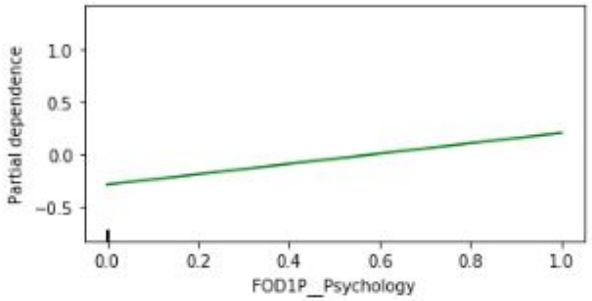
- CS degree
- Live in the bay
- Engineering degree
- Any degree

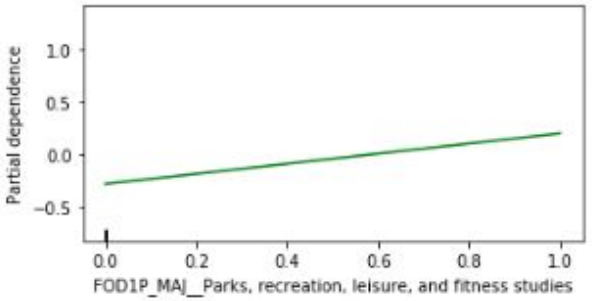
Models and Results: Healthcare

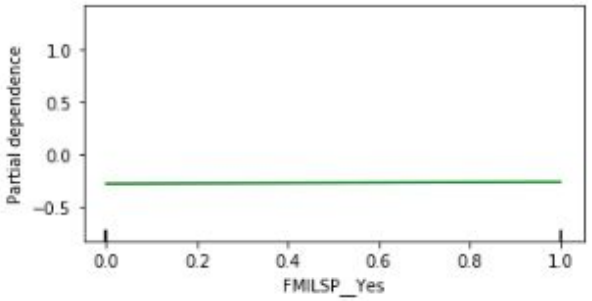
```
Best training f1: 0.965
test accuracy 0.9648178451468435:
test auc 0.6233580746738642:
Test set f1 score for best params: 0.378
```

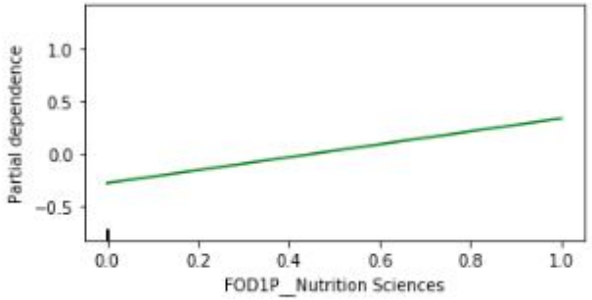
Partial Dependence Plots Using Lightly Tuned Gradient Boost

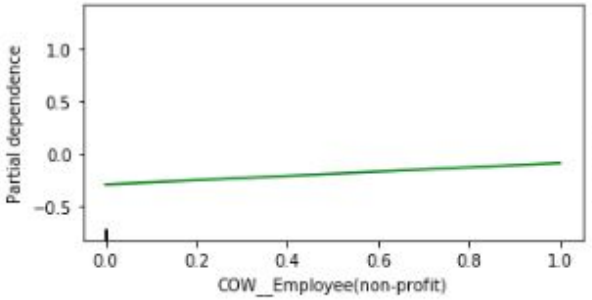


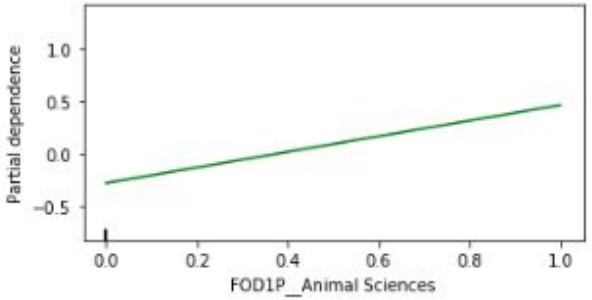

Partial dependence
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0
FOD1P__Psychology

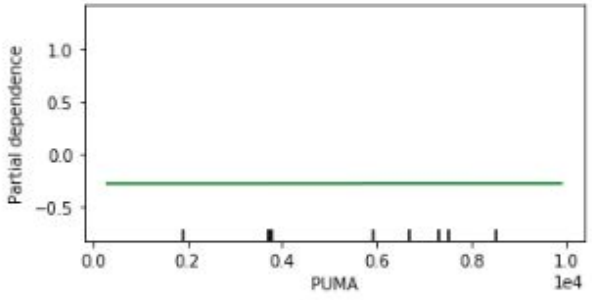

Partial dependence
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0
FOD1P_MAJ__Parks, recreation, leisure, and fitness studies

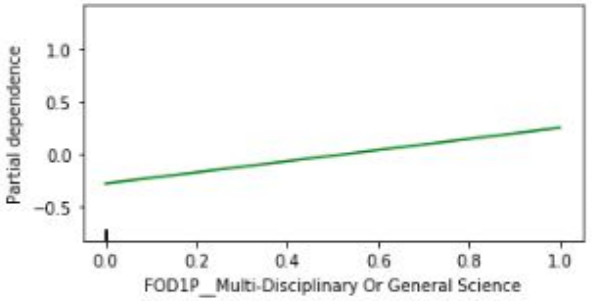

Partial dependence
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0
FMILSP__Yes

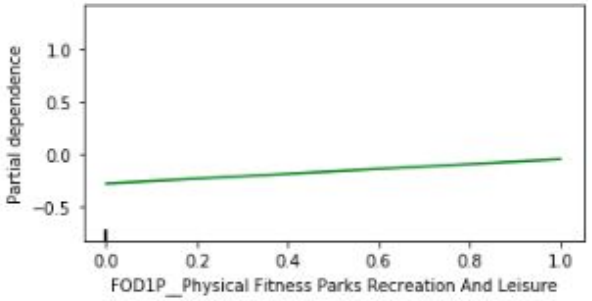

Partial dependence
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0
FOD1P__Nutrition Sciences


Partial dependence
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0
COW__Employee(non-profit)


Partial dependence
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0
FOD1P__Animal Sciences

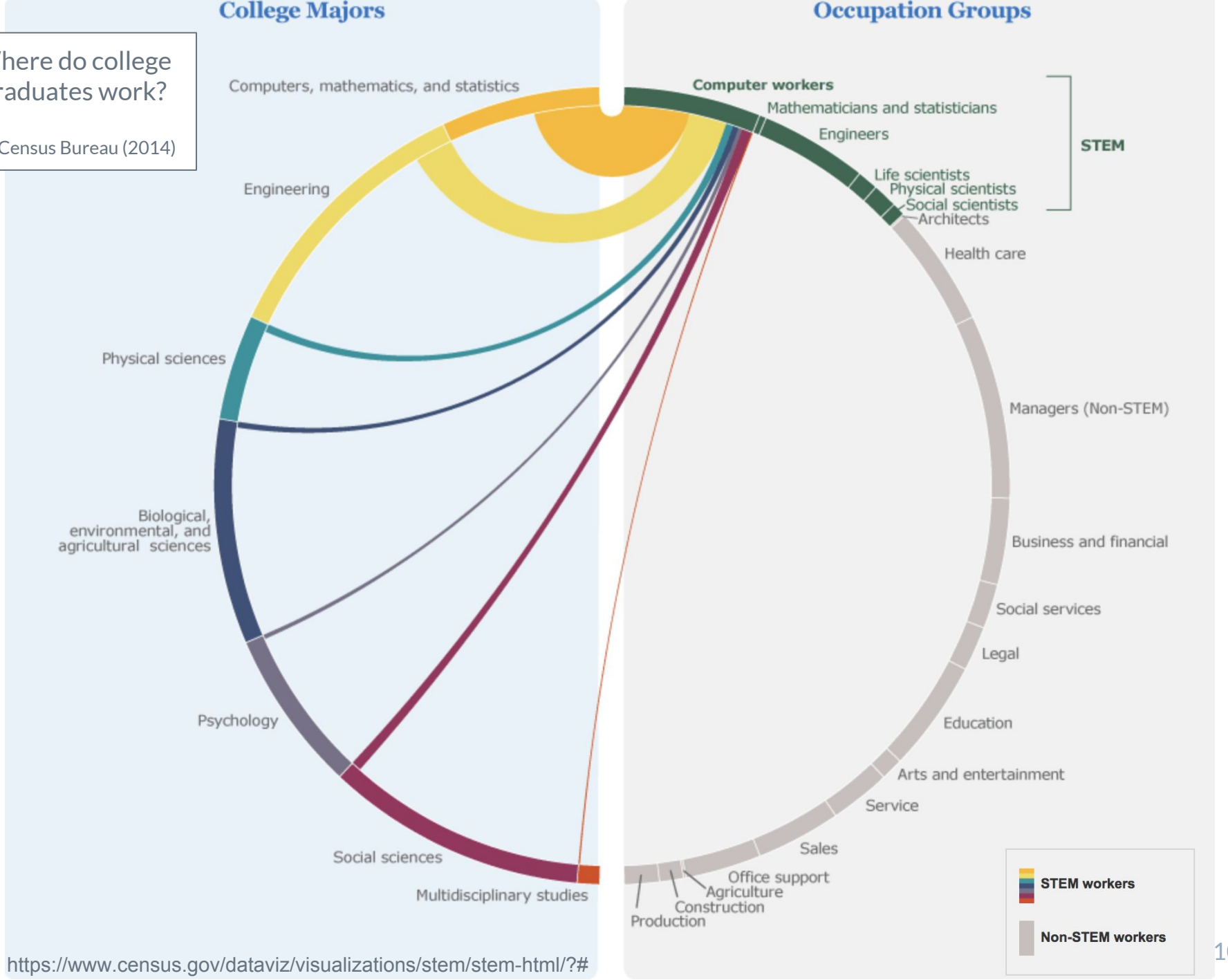

Partial dependence
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0
1e4
PUMA


Partial dependence
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0
FOD1P__Multi-Disciplinary Or General Science

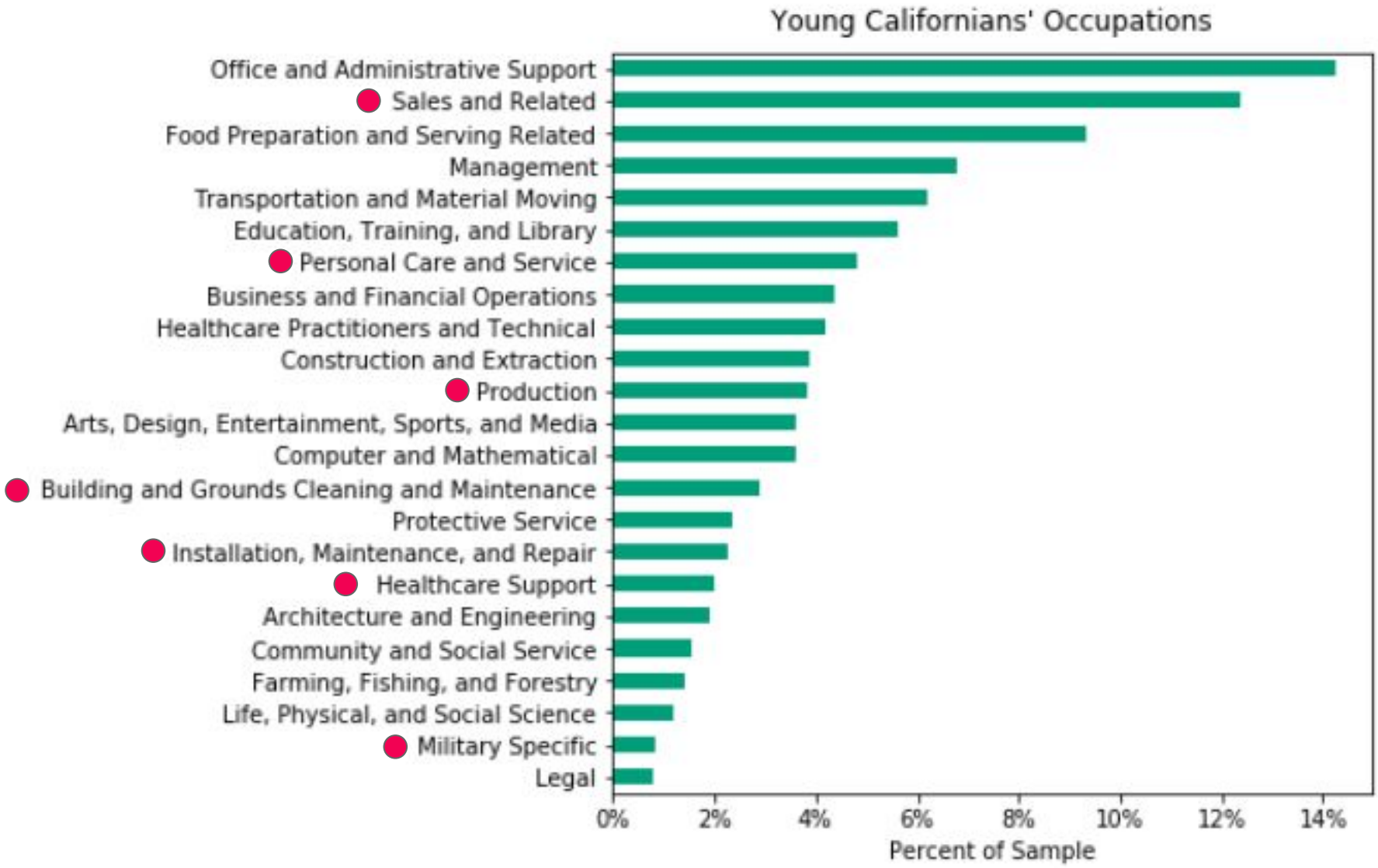

Partial dependence
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0
FOD1P__Physical Fitness Parks Recreation And Leisure

Where do college graduates work?

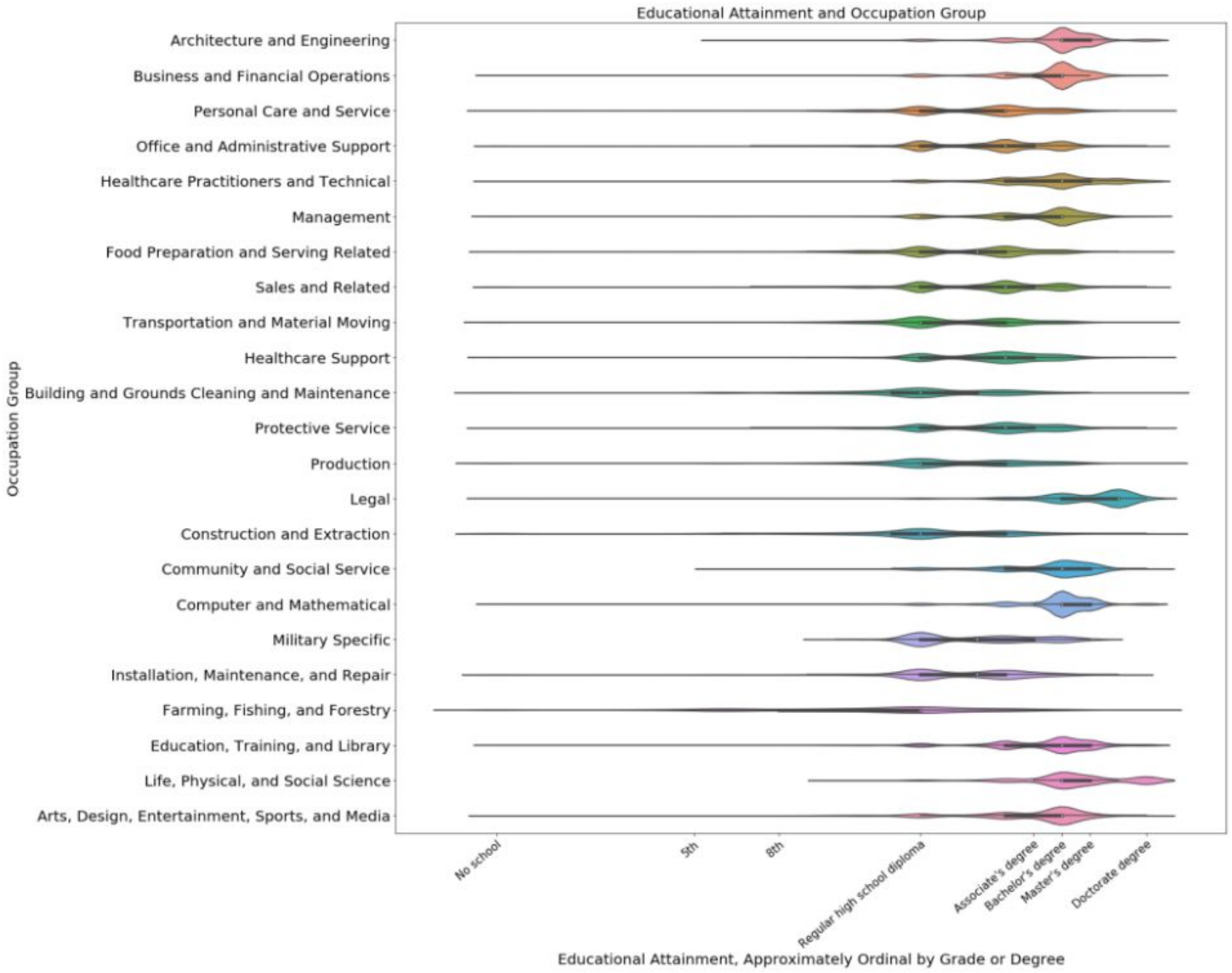
US Census Bureau (2014)



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



- Occupations for which education was not predictive, either low attainment or high


Educational Attainment and Occupation Group
Occupation Group
Architecture and Engineering
Business and Financial Operations
Personal Care and Service
Office and Administrative Support
Healthcare Practitioners and Technical
Management
Food Preparation and Serving Related
Sales and Related
Transportation and Material Moving
Healthcare Support
Building and Grounds Cleaning and Maintenance
Protective Service
Production
Legal
Construction and Extraction
Community and Social Service
Computer and Mathematical
Military Specific
Installation, Maintenance, and Repair
Farming, Fishing, and Forestry
Education, Training, and Library
Life, Physical, and Social Science
Arts, Design, Entertainment, Sports, and Media
No school
5th
8th
Regular high school diploma
Associate's degree
Bachelor's degree
Master's degree
Doctorate degree
Educational Attainment, Approximately Ordinal by Grade or Degree

Occupation Predictor

Based on Paths Others Have Taken to Your Goal

Your Desired Profession: Select

About You:

Complete for highest accuracy.

Highest degree or level of school completed: Select

Major of your Bachelor's Degree if you have one: Select

Major of your second Bachelor's Degree if you have one: Select

State of Residence: Select

City: Select

Ability to Speak English: Select

Demographics Questions:

Predict!

Your Predicted Likelihood of Being a _____ : 62%

You're Most Likely to Be a:

- Item 1
- Item 2
- Item 3
- Item 4

Not what you want? We believe you can achieve your goal no matter what you are likely to be. You will be all the more incredible for your achievement. To help you get there, we recommend:

What you can change:

Your Likelihood Will Increase to: 75%

Next Steps

- expand to useable dataset + 3 to 8 million rows
  - models for very specific occupations
- more feature engineering, class balancing
- look at effect of having children on occupation
- run on all Census features
  - address correlations with demographics

Thanks!

Any questions?

http://bit.ly/DMS-Predict-Occp

dmariesaunders@gmail.com