Gender and Racial Disparities in STEM Workforce: A Machine Learning Approach

Luxin Tian

Division of the Social Sciences, The University of Chicago, luxintian@uchicago.edu

Research Question

- While plenty of evidence shows that women racial minorities have been underrepresented in the STEM workforce in the US, there lacks consensus on the classification standards of STEM occupations.
- Traditional Classification approaches have limitations:
    - Official approach: US Bureau of Labor Statistics SOC STEM occupations list, official top-down classification.
    - Brookings approach: scoring occupations on STEM features, top-down classification created by researchers and experts.
    - RAND ALP approach: micro-level survey data, bottom-up response from workers.
- Construct a machine learning classifier
    - Integrate multiple data sources: RAND ALP bottom-up response and O*NET occupation attributes.
    - Feature-based classifier, providing feature-based label prediction and probability, which makes the STEM labels un-biased, continuous and quantifiable.
- Re-examine Gender and Racial Disparities in STEM Workforce
    - Use the STEM labels predicted by the feature-based classifier.
    - Evaluate the result based on previous classification approaches.

Data and Methods

- Training a feature-based classifier:
    - Labels: RAND American Life Panel Survey (ALP) interviews a panel of more than 6,000 nation-wide representatives through the internet each year (Pollard and Baird, 2017). Occupation-relative questions are included in MS480 and MS436, which contains self-reported STEM labels and SOC occupation codes.
    - Features: O*NET dataset contains a comprehensive set of descriptive variables that characterizes all aspects of features of the nearly 1,000 occupations in the US economy.
- Identify Gender and Racial Disparities:
    - American Community Survey (ACS) 2018 provides micro data containing 1,247,722 samples (after dropping records with missing data) across the US.
    - Weighted OLS regression and Logistic Regression

Classification Model Selection

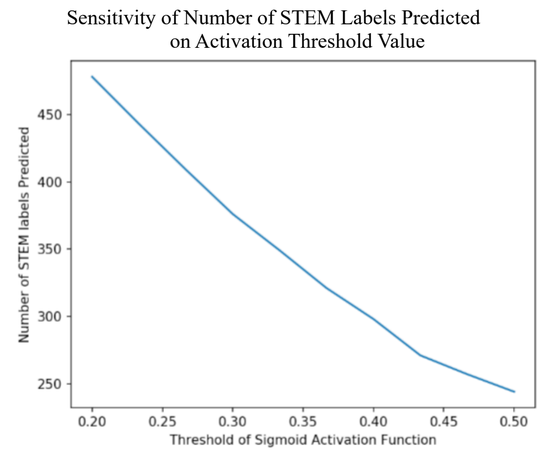
- Feature selection:
    - LASSO, Ridge regression with CV
    - Recursive Feature Elimination (RFE) with SV
    - Support Vector Machine with linear kernel
    - Logistic Regression with CV
- Classifier:
    - Logistic Regression
    - Tree-based models

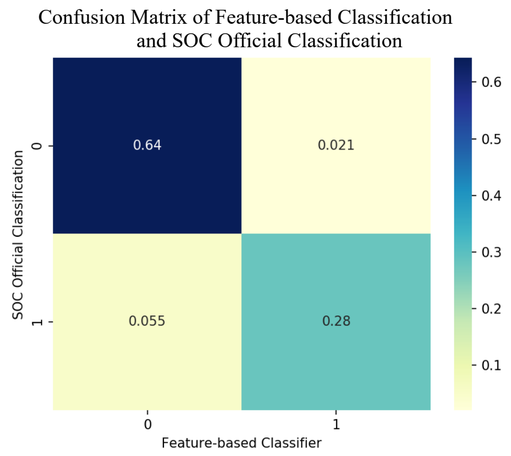
Performance of Experimental Feature Selection Methods and Classifiers

Pipeline	Feature Selection	Classifier	Best CV Score on Training/Test Data	Precision by Labels	Recall by Labels	F1 by Labels
1	None	Random Forest	0.86706, 0.80859	0.88, 0.58	0.86, 0.63	0.87, 0.61
2	LASSO with CV	Random Forest	0.86706, 0.79685	0.87, 0.57	0.87, 0.57	0.87, 0.57
3	Ridge with CV	Random Forest	0.86706, 0.79297	0.87, 0.56	0.86, 0.57	0.86, 0.56
4	Logistic with CV	Random Forest	0.86706, 0.79688	0.87, 0.57	0.87, 0.57	0.87, 0.57
5	RFECV with Linear SVC	Random Forest	0.86706, 0.8125	0.88, 0.63	0.87, 0.57	0.87, 0.60
6	None	Logistic with CV	0.82111, 0.84375	0.86, 0.75	0.95, 0.50	0.90, 0.60
7	RFECV with Linear SVC	Logistic with CV	0.84066, 0.84376	0.87, 0.72	0.93, 0.55	0.90, 0.62

Classification Results: Defining STEM Workforce

- With 0.43 as the activation threshold of the Logistic sigmoid function, the classifier achieves 84% accuracy on the test set.
- Confusion matrix suggests that 7.6% of occupations are mislabeled by the official STEM classification.





Empirical Results: Gender and Racial Gaps

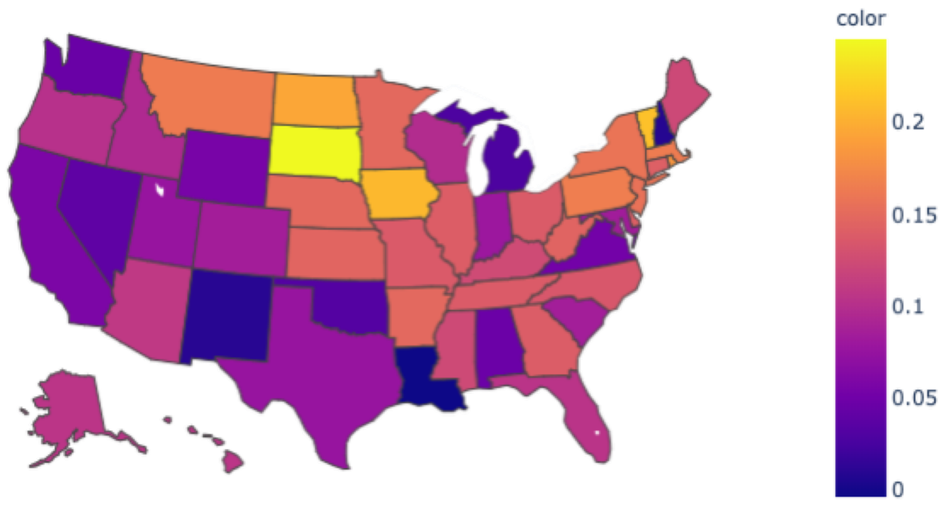
- Country-level Regressions
    - Females remain significantly underrepresented in STEM workforce
    - While Asians have higher probability of engaging in STEM occupations, other racial minorities are significantly underrepresented.
    - Previous research based on traditional classifications standards overestimated the gender and racial gaps in STEM engagement.

Gender and Racial Disparities in STEM Workforce Engagement in the US (2018)

VARIABLES	Logistic: Z = SOC Official STEM Labels I	II	III	IV	V	Logistic: Z = Featured-based Predicted Label I	II	III	IV	V	log(Pr{STEM=1}) I	II	III	IV	V
Male	1.0634***	1.0647***	1.0623***	0.9741***	1.0245***	0.9616***	0.9604***	0.9578***	0.8433***	0.8896***	0.2399***	0.2323***	0.2263***	0.1151***	0.1075***
	(0.0046)	(0.0047)	(0.0047)	(0.0051)	(0.0057)	(0.0048)	(0.0048)	(0.0048)	(0.0053)	(0.0059)	(0.0016)	(0.0016)	(0.0016)	(0.0014)	(0.0016)
Age		0.1046***	0.1036***	0.0529***	0.0529***		0.1028***	0.1018***	0.0474***	0.0474***		0.0499***	0.0497***	0.0258***	0.0257***
		(0.0009)	(0.0009)	(0.0010)	(0.0010)		(0.0009)	(0.0009)	(0.0011)	(0.0011)		(0.0003)	(0.0003)	(0.0003)	(0.0003)
Age-squared		-0.0010***	-0.0010***	-0.0005***	-0.0005***		-0.0010***	-0.0010***	-0.0004***	-0.0004***		-0.0005***	-0.0005***	-0.0002***	-0.0002***
		(0.0000)	(0.0000)	(0.0000)	(0.0000)		(0.0000)	(0.0000)	(0.0000)	(0.0000)		(0.0000)	(0.0000)	(0.0000)	(0.0000)
African American			-0.1776***	0.0050	-0.0902***			-0.2552***	-0.0759***	-0.1722***			-0.1299***	-0.0636***	-0.1522***
			(0.0081)	(0.0090)	(0.0181)			(0.0085)	(0.0095)	(0.0192)			(0.0025)	(0.0022)	(0.0034)
American Indian			-0.6622***	-0.2725***	-0.0960			-0.7168***	-0.3044***	-0.0560			-0.2356***	-0.0811***	-0.0326**
			(0.0322)	(0.0358)	(0.0635)			(0.0343)	(0.0382)	(0.0659)			(0.0105)	(0.0092)	(0.0133)
Alaska Native			-1.1766***	-0.5372***	-0.5636			-1.0376***	-0.3001*	0.0914			-0.3062***	-0.0608	-0.0153
			(0.1617)	(0.1783)	(0.3763)			(0.1615)	(0.1815)	(0.3216)			(0.0492)	(0.0433)	(0.0603)
American Indian or Alaska Native			-0.7308***	-0.2433***	0.1750			-0.6945***	-0.1662*	0.2721**			-0.3250***	-0.1392***	-0.0929***
			(0.0763)	(0.0850)	(0.1349)			(0.0789)	(0.0887)	(0.1378)			(0.0215)	(0.0190)	(0.0270)
Asian			0.5535***	0.2870***	0.6425***			0.5710***	0.2894***	0.5885***			0.1850***	0.0234***	0.0396***
			(0.0083)	(0.0092)	(0.0148)			(0.0084)	(0.0093)	(0.0151)			(0.0033)	(0.0029)	(0.0045)
Hawaiian/Pacific			-0.7973***	-0.4503***	-0.1780			-0.7825***	-0.4007***	-0.0575			-0.2702***	-0.1256***	-0.1001***
			(0.0699)	(0.0766)	(0.1292)			(0.0730)	(0.0803)	(0.1301)			(0.0190)	(0.0167)	(0.0247)
Other Races			-1.1044***	-0.4970***	-0.3140***			-1.1592***	-0.5305***	-0.3212***			-0.4396***	-0.1604***	-0.0964***
			(0.0173)	(0.0191)	(0.0332)			(0.0187)	(0.0206)	(0.0349)			(0.0038)	(0.0035)	(0.0047)
Two or More Races			-0.2013***	-0.1577***	0.0665**			-0.2006***	-0.1502***	0.0881***			-0.0775***	-0.0556***	-0.0470***
			(0.0152)	(0.0167)	(0.0286)			(0.0157)	(0.0173)	(0.0294)			(0.0049)	(0.0044)	(0.0065)
Female African American					0.1233***					0.1240***					0.1513***
					(0.0209)					(0.0222)					(0.0045)
Female American Indian					-0.2542***					-0.3620***					-0.0934***
					(0.0766)					(0.0806)					(0.0184)
Female Alaska Native					0.0285					-0.5457					-0.0961
					(0.4276)					(0.3871)					(0.0866)
Female American Indian					-0.6428***					-0.6896***					-0.0917**
					(0.1722)					(0.1784)					(0.0379)
Female Asian					-0.5575***					-0.4674***					-0.0265***
					(0.0186)					(0.0189)					(0.0059)
Female Hawaiian/Pacific					-0.4020**					-0.5218***					-0.0474
					(0.1596)					(0.1644)					(0.0336)
Female of Other Races					-0.2669***					-0.3093***					-0.1372***
					(0.0403)					(0.0429)					(0.0068)
Female of Two or More Races					-0.3287***					-0.3529***					-0.0158*
					(0.0350)					(0.0362)					(0.0087)
const	-1.7948***	-4.1593***	-4.1092***	-5.0293***	-5.0611***	-1.8634***	-4.2024***	-4.1476***	-5.0016***	-5.0329***	-1.9956***	-3.1434***	-3.0968***	-3.1684***	-3.1644***
	(0.0039)	(0.0192)	(0.0193)	(0.0564)	(0.0564)	(0.0040)	(0.0199)	(0.0200)	(0.0605)	(0.0605)	(0.0012)	(0.0061)	(0.0061)	(0.0096)	(0.0096)
Controls	N	N	N	Y	Y	N	N	N	Y	Y	N	N	N	Y	Y
Num of observations	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722	1,247,722
											0.0175	0.0501	0.0653	0.2757	0.2767

- State-level Regressions: Gender Gaps
    - Gender disparities are heterogeneous across the states in the US
    - States with the most significant gender disparities are South Dakota, Vermont, Iowa, and North Dakota.

State-level Gender Disparities in STEM Workforce Engagement in the US (2018)



- The results are consistent with Anderson et al (2018) who employ the smaller ALP survey dataset and use bottom-up responses as STEM labels.

Reference

Pollard, Michael and Matthew D Baird, "The RAND American Life Panel: Technical Description," 2017.10

Anderson, Drew, Matthew Baird, and Robert Bozick, "Who Gets Counted as Part of America's STEM Workforce? The Implications of Different Classification Approaches for Understanding the Gender Gap in STEM," 2018