College Major and Career Exploration

Abhishek Mahajan, Gregory Glatzer,
Hojin Ryoo, Pranay Muthineni

Objective

Using the PUMS recent-grads dataset, we will build a machine learning model that can predict a salary based on one’s major and career.

EDA

Unemployment Rates by Major

10 lowest

EDUCATIONAL ADMINISTRATION AND SUPERVISION	0%
MILITARY TECHNOLOGIES,	0%
BOTANY	0%
MATHEMATICS AND COMPUTER SCIENCE	0 %
SOIL SCIENCE	0 %
ENGINEERING MECHANICS PHYSICS AND SCIENCE	.633%
COURT REPORTING	1.169%
MATHEMATICS TEACHER EDUCATION	1.620%
PETROLEUM ENGINEERING	1.838%
GENERAL AGRICULTURE	1.964%

10 highest

ARCHITECTURE	11.333%
GEOGRAPHY	11.345%
COMPUTER PROGRAMMING AND DATA PROCESSING	11.398%
MINING AND MINERAL ENGINEERING	11.724%
COMMUNICATION TECHNOLOGIES	11.951%
PUBLIC POLICY	12.842%
CLINICAL PSYCHOLOGY	14.904%
COMPUTER NETWORKING AND TELECOMMUNICATIONS	15.184%
PUBLIC ADMINISTRATION	15.949%
NUCLEAR ENGINEERING	17.722%

Median Salary by Major

5 lowest

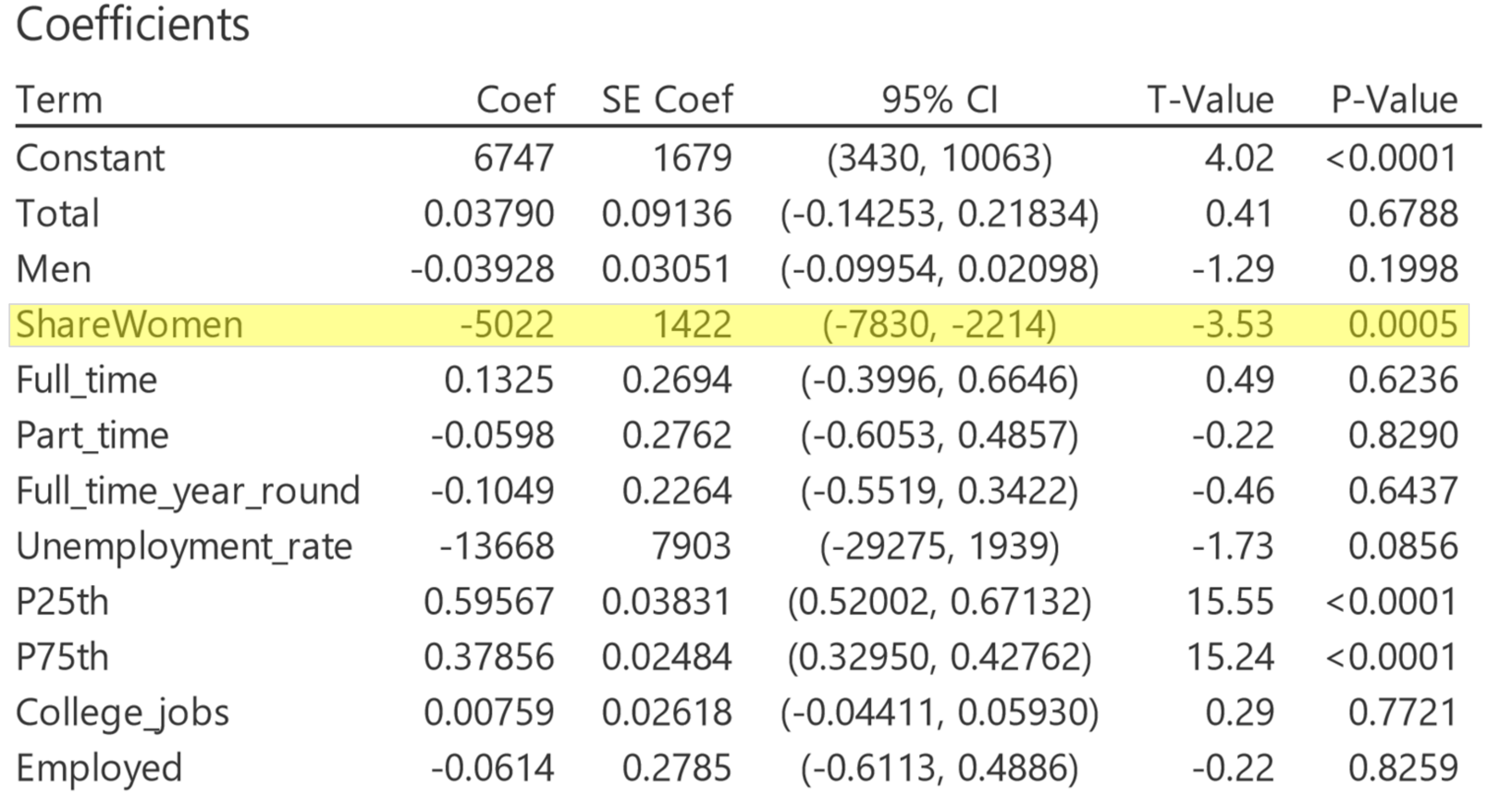
Psychology & Social Work	$30100.00
Humanities & Liberal Arts	$31913.33
Education	$31913.33
Arts	$33062.50
Communications & Journalism	$34500.00

5 highest

Engineering	$57382.75
Business	$43538.46
Computers & Mathematics	$42745.45
Law & Public Policy	$42200.00
Physical Sciences	$41890.00

Given that there are no extremely high IQRs, we don’t need to normalize our data, as there are no features with very high standard deviations.

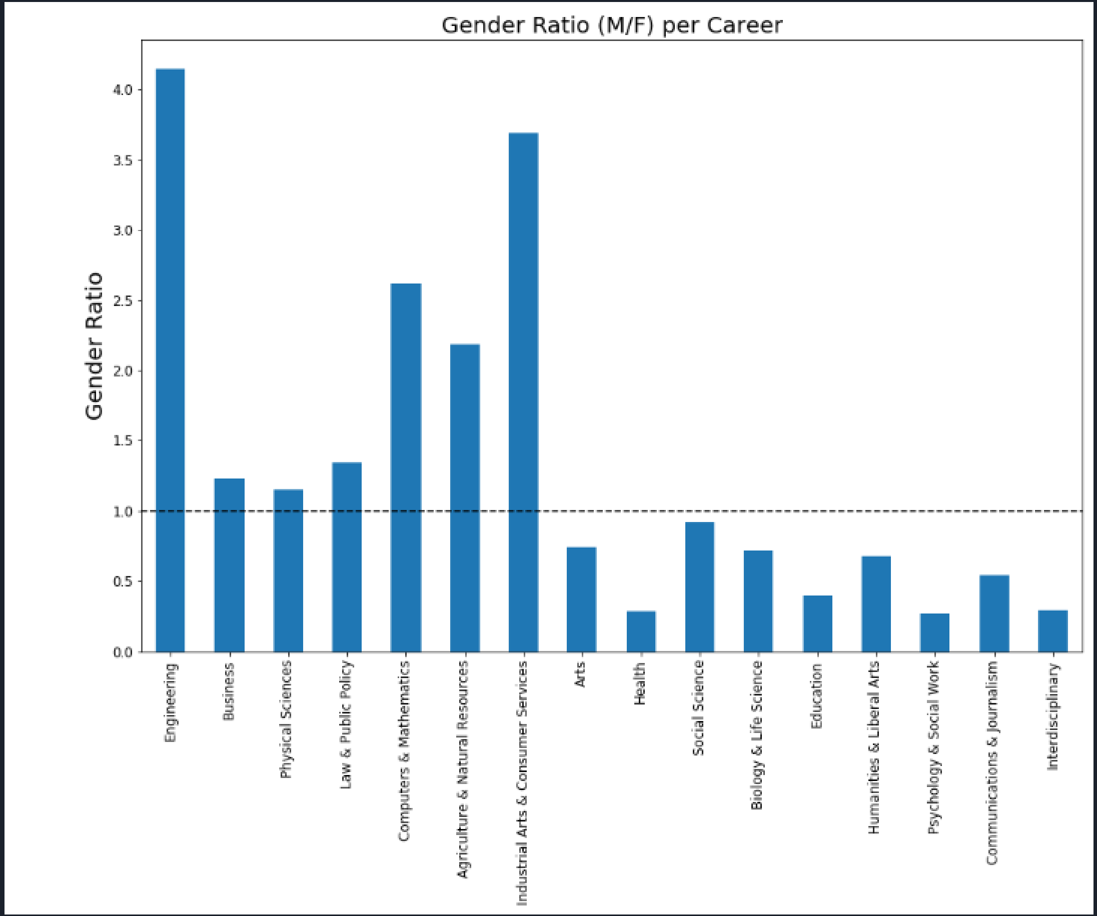
Adding relevant features



Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value
Constant	6747	1679	(3430, 10063)	4.02	<0.0001
Total	0.03790	0.09136	(-0.14253, 0.21834)	0.41	0.6788
Men	-0.03928	0.03051	(-0.09954, 0.02098)	-1.29	0.1998
ShareWomen	-5022	1422	(-7830, -2214)	-3.53	0.0005
Full_time	0.1325	0.2694	(-0.3996, 0.6646)	0.49	0.6236
Part_time	-0.0598	0.2762	(-0.6053, 0.4857)	-0.22	0.8290
Full_time_year_round	-0.1049	0.2264	(-0.5519, 0.3422)	-0.46	0.6437
Unemployment_rate	-13668	7903	(-29275, 1939)	-1.73	0.0856
P25th	0.59567	0.03831	(0.52002, 0.67132)	15.55	<0.0001
P75th	0.37856	0.02484	(0.32950, 0.42762)	15.24	<0.0001
College_jobs	0.00759	0.02618	(-0.04411, 0.05930)	0.29	0.7721
Employed	-0.0614	0.2785	(-0.6113, 0.4886)	-0.22	0.8259

Adding relevant features



Linear Regression Model

Logistic Regression Model

Model Evaluation

Using RMSE as our metric of choice, a Linear Regression model outperformed a Logistic Regression model.

Most important features to predict Median salary

- Employed
- College_jobs
- Gender_Ratio
- Career

RMSE: 5879.62.

Challenges