Determining Factors and Patterns in University Faculty Salary Differences with Regressive and Correlative Analysis

Tyler Branscombe & Edmund Chen

Introduction

Goal: Identify and model relationships race, gender, title, and rating have with faculty salaries at research universities

Reason: Salaries vary a lot past different factors, which factors are most determining?

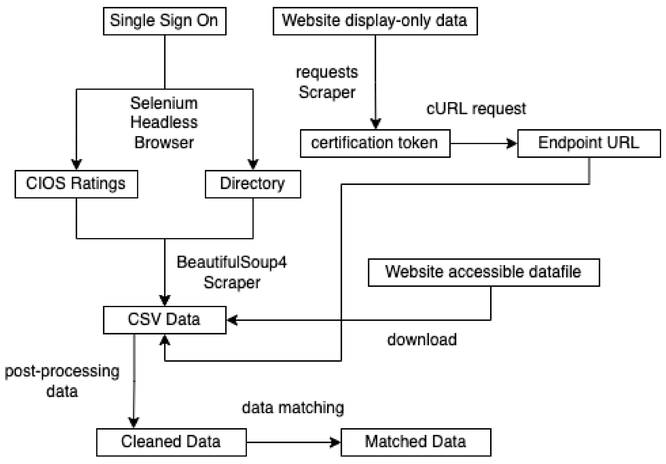
Data Aggregation

Base Data: 7-8 fields for 497,834 observations, public institution data from TN, IL, CA, GA, NC

Macro Data: Analysis across all universities
Micro Data: Analysis across Georgia Tech

Race + Gender: Uses demographic distribution to best guess off last and first name respectively (ethnicolr, gender-guesser)

Rating: Scraped from headless browser into CIOS/smartevals. Systematic gathering of by professor evaluation data

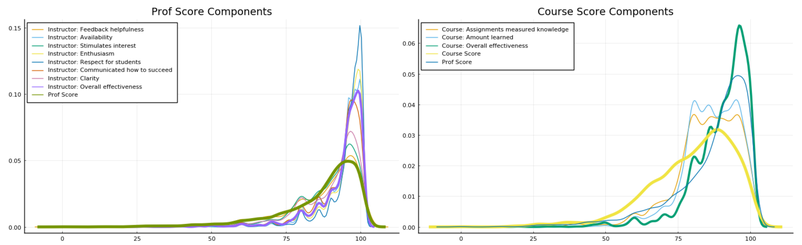


Data Processing

Role Normalization: Fuzzy matching and generalization of professor roles into four main roles (Assistant, Associate, Adjunct, Full)

- ie. VST ADJ PROF → Adjunct

Rating Normalization: Rating data many dimensions, normalized to one overall effectiveness course score, further reduced out of 5 and matched to Georgia Tech professors (roughly ~70% matches)



Research Impact: Challenges matching and gathering enough significant datapoint coverage

Regression Modeling

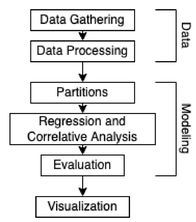
Modeling: (for 3 subsets of our data)
Linear, Lasso, Ridge, Neural Network, Logistic

$$Y_i = f(X_i, \beta) + e_i$$

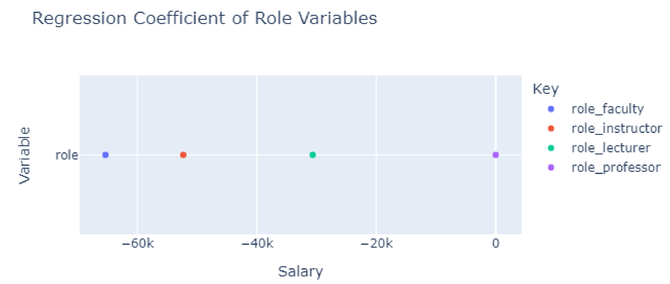
Evaluation of models:
R-Squared, RMSE, MAE
Neural Networks were the most effective model

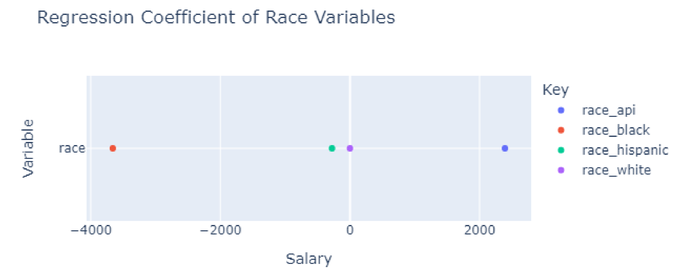
Evaluation of Variables:
P-value of linear model for signifigance
Coefficient of Linear model to see impact

Method	R2	RMSE	MAE
linear model	0.393053478600865	0.0299978822203757	0.023283070
Neural Network	0.404417784088155	0.029715376874596	0.023160592
Ridge Regression	0.393132238841919	0.0299945820200855	0.023288753
Lasso Regression	0.385788195522131	0.0302315097680273	0.023366951
Logistic Regression	0.393015767253654	1.13637920162471	1.132569827



Results and Analysis





Linear Regression Coefficients for Role-Partitioned Salary Analysis

Rating Coefficient for GT Data:
0.027339

```
Residuals:
     Min       1Q   Median       3Q      Max
-382690  -21398   -5371   24223  784702

Coefficients: (4 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.453e+06  4.028e+04 -36.081  < 2e-16 ***
state_ca        -2.661e+04  9.001e+02 -29.568  < 2e-16 ***
state_ga         8.381e+04  1.195e+03  70.128  < 2e-16 ***
state_il         1.063e+04  8.996e+02  11.813  < 2e-16 ***
state_nc         3.556e+04  9.480e+02  37.515  < 2e-16 ***
state_tn                NA         NA      NA       NA
year             7.607e+02  1.992e+01  38.180  < 2e-16 ***
race_api         2.391e+03  2.204e+02  10.848  < 2e-16 ***
race_black      -3.660e+03  1.250e+03  -2.928  0.00341 **
race_hispanic   -2.772e+02  2.076e+02  -1.335  0.18184
race_white              NA         NA      NA       NA
gender_andy      1.426e+03  4.344e+02   3.281  0.00103 **
gender_female   -7.897e+03  2.045e+02 -38.621  < 2e-16 ***
gender_male     -3.023e+03  2.023e+02 -14.941  < 2e-16 ***
gender_mostly_female -8.858e+03  3.394e+02 -26.097  < 2e-16 ***
gender_mostly_male   -4.687e+03  4.256e+02 -11.012  < 2e-16 ***
gender_unknown          NA         NA      NA       NA
role_f          -6.531e+04  6.622e+02 -98.630  < 2e-16 ***
role_i          -5.230e+04  6.380e+02 -81.972  < 2e-16 ***
role_l          -3.063e+04  1.265e+02 -242.214  < 2e-16 ***
role_p                  NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38920 on 448031 degrees of freedom
Multiple R-squared:  0.3432,    Adjusted R-squared:  0.3432
F-statistic: 1.463e+04 on 16 and 448031 DF,  p-value: < 2.2e-16
```

Discussion & Conclusions

Implicates predictive power of the different factors on salary, and how they vary by partition. Results suggest there likely is biases with all investigated variables