DATA 606 Data Project

Nick Oliver

DATA 606 - Final Project

Part 1 - Introduction

Abstract

My question was to find if a person's occupation being in a Science, Technology, Engineering, or Mathematics (STEM) field is a stronger predictor of which political party a person will donate to over simply the state that they live in. I used the Federal Election Commissions's Contributions by individuals campaign finance data[1], along with MIT's Presidential Election data[2] to perform binomial logistic regression using two predictors, employment in a STEM field and living in democratic or republican voting state. My results show the combination of the two predictors, employment in a STEM field and residence in a blue or red state, result in the best fit prediction model for predicting which political party a random donating individual is likely to donate to using the Akaike information criterion (AIC) value as the selection criteria for selecting the best fit. In addition, I was able to show that what state a person lives in provides a better fitting model for predicting political party donations over simply using employment in a STEM occupation again using AIC as the criteria for determining best fit. My results show that while there is predictive value in knowing if a person is employed in a STEM field when it comes to determining their political donation proclivities, it does not appear to be a better predictor than simply knowing which state a person lives in and whether the majority of that state voted for a particular party in the U.S. presidential election.

Background

My project was inspired by the data analysis that FiveThirtyEight published in April of 2017 regarding which careers donate to which political parties titled *When Scientists Donate To Politicians, It's Usually To Democrats*[3]. The conclusion of the analysis was that professionals employed in the science, technology, engineering, and mathematics (STEM) fields are more likely to donate to democratic candidates over republicans.

When looking at the data that FiveThirtyEight used for the analysis I was curious if there was a stronger correlation between one's geographic location and their donation habits or their profession. In order to perform my analysis I had to supplement the FiveThirtyEight dataset with presidential election results by state. In addition, simply using FiveThirtyEight's dataset was not sufficient to perform my analysis because they filtered the data to only include donations made by those employed in STEM occupations.

Part 2 - Data

[1] Contributions by individuals. FEC.gov. (n.d.). Retrieved December 6, 2021, from https://www.fec.gov/campaign-finance-data/contributions-individuals-file-description/.

[2] MIT Election Data and Science Lab, 2017, "U.S. President 1976–2020", https://doi.org/10.7910/DVN/42MVDX, Harvard Dataverse, V6, UNF:6:4KoNz9KgTkXy0ZBxJ9ZkOw== [fileUNF]

[3] Benbwieder. (2017, April 21). When scientists donate to politicians, it's usually to Democrats. FiveThirtyEight. Retrieved December 6, 2021, from https://fivethirtyeight.com/features/when-scientists-donate-to-politicians-its-usually-to-democrats/.

FiveThirtyEight FEC Cleaned Data

For my first data set I will use FiveThirtyEight's cleaned and manipulated version of the Federal Election Commission individual contributions data set[4] used in their article When Scientists Donate To Politicians, It's Usually To Democrats.

For my analysis I am using it soley for the purposes of using their definition of STEM occupation. For that reason I will only be selecting a dataframe containing unique occupations from FiveThirtyEight's original data set.

Below is a glimpse of the data.

```
## Rows: 9,266
## Columns: 1
## $ cleanedoccupation <chr> "ENGINEER", "CIVIL ENGINEER", "ENGINEER (SOFTWARE)",~
```

Federal Election Commision Bulk Data

My second data sources is a subset of the Federal Election Commision's (FEC) individual contributions bulk data set. The source data is publicly available for download on the FEC's website bulk data download page.

In addition I used the FEC's committee data to enrich the individual contributions data with the political affiliation of the organization that the individual contributed to. That data is also publicly available on the FEC's website committees dat page

For my analysis I am using the donating individuals occupation, data of residence, date of donation, and contribution beneficiary from the data set. Due to the size of the data I randomly sampled one million rows from the following years, 2022-2021, 2020-2019, 2018-2017, and 2016-2015. I filtered all contributions to parties that were not democrat or republican.

A copy of this data is located in comrpess format for download in this projects GitHub Repository here

Here is a glimpse of the resulting data set.

```
## Rows: 6,674,211
## Columns: 5
## $ X          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
## $ party      <chr> "REP", "REP", "REP", "REP", "REP", "REP", "REP", "REP", "RE~
## $ state      <chr> "WA", "WA", "WA", "WA", "WA", "WA", "WA", "WA", "WA", "WA",~
## $ OCCUPATION <chr> "CHAIRMAN", "PRESIDENT OF BERNTSON PORTER WEALTH MA", "RETI~
## $ year       <int> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015,~
```

MIT Presidental Election Data

My last data set is MIT's publicly available presidential election data set. This data is publicly available for download on the Harvard Dataverse website

```
## Rows: 4,287
## Columns: 15
## $ year             <int> 1976, 1976, 1976, 1976, 1976, 1976, 1976, 1976, 1976,~
## $ state            <I<chr>> ALABAMA, ALABAMA, ALABAMA, ALABAMA, ALABAMA, ALABA~
## $ state_po         <I<chr>> AL, AL, AL, AL, AL, AL, AL, AK, AK, AK, AK, AZ, AZ~
## $ state_fips       <int> 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4, 4,~
## $ state_cen        <int> 63, 63, 63, 63, 63, 63, 63, 94, 94, 94, 94, 86, 86, 8~
## $ state_ic         <int> 41, 41, 41, 41, 41, 41, 41, 81, 81, 81, 81, 61, 61, 6~
## $ office           <I<chr>> US PRESIDENT, US PRESIDENT, US PRESIDENT, US PRESI~
## $ candidate        <I<chr>> CARTER, JIMMY, FORD, GERALD, MADDOX, LESTER, BUBAR~
```

[4] Benbwieder. (2017, April 21). When scientists donate to politicians, it's usually to Democrats. FiveThirtyEight. Retrieved December 6, 2021, from https://fivethirtyeight.com/features/when-scientists-donate-to-politicians-its-usually-to-democrats/.

```
## $ party_detailed   <I<chr>> DEMOCRAT, REPUBLICAN, AMERICAN INDEPENDENT PARTY, ~
## $ writein          <I<chr>> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FA~
## $ candidatevotes   <int> 659170, 504070, 9198, 6669, 1954, 1481, 308, 71555, 4~
## $ totalvotes       <int> 1182850, 1182850, 1182850, 1182850, 1182850, 1182850,~
## $ version          <int> 20210113, 20210113, 20210113, 20210113, 20210113, 202~
## $ notes            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ party_simplified <I<chr>> DEMOCRAT, REPUBLICAN, OTHER, OTHER, OTHER, LIBERTA~
```

After manipulating the data to fit my analysis I am left with a simple data set which contains true/false values for if the individual donated to a democratic organization, if their occupation was in a STEM field, and if they resided in a state where the majority of the population voted for a democratic presidential candidate in the year of the donation.

Here is a glimpse of the data set.

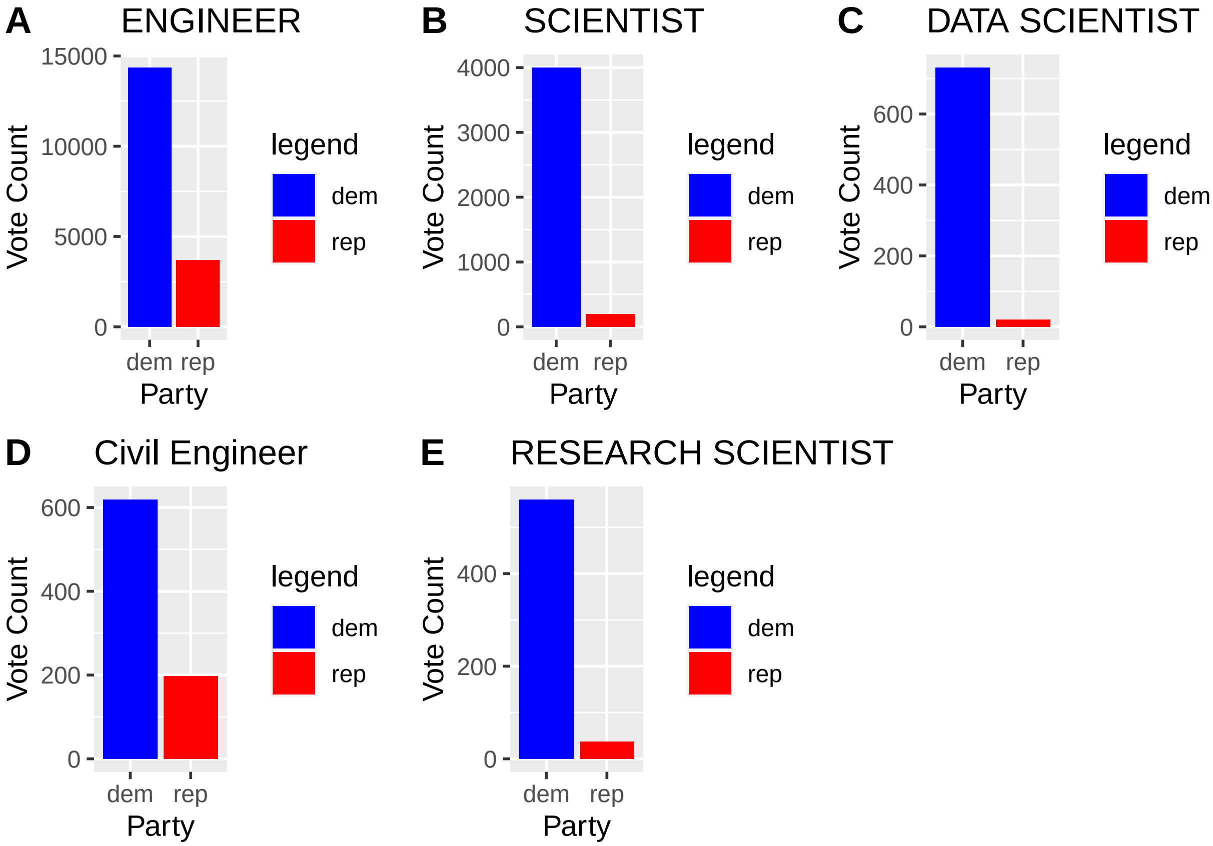
```
## Rows: 1,411,520
## Columns: 4
## $ donateDem <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE,~
## $ stem      <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS~
## $ blueState <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, ~
## $ year      <int> 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, ~
```

Part 3 - Exploratory data analysis

Take top five STEM occupations

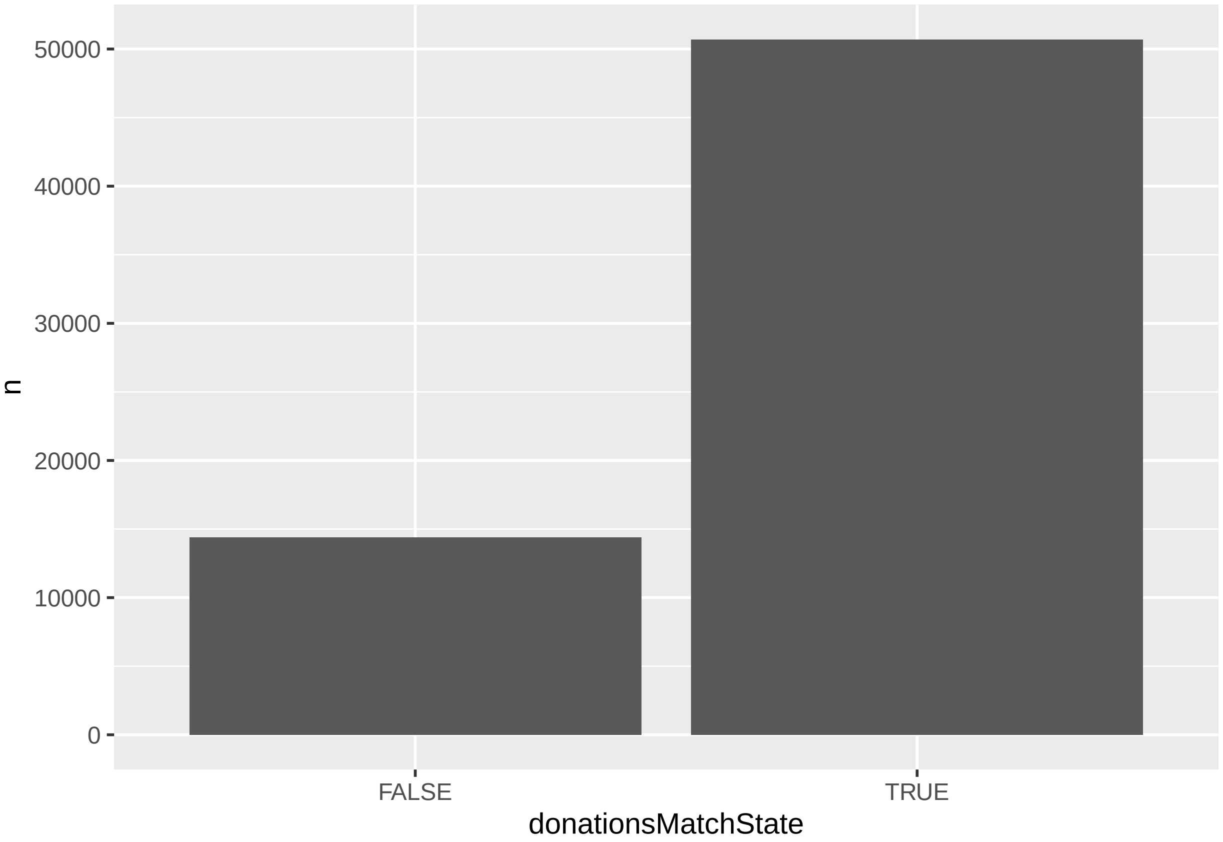
```
## # A tibble: 5 x 2
## # Groups:   occupation [5]
##   occupation             n
##   <chr>              <int>
## 1 ENGINEER           18059
## 2 SCIENTIST           4200
## 3 CIVIL ENGINEER       817
## 4 DATA SCIENTIST       752
## 5 RESEARCH SCIENTIST   597
```

Do they uniformly donate to democratic organizations? These plots show that the conclusions drawn by the author of FiverThirtyEight's article have some validity given that an overwhelming majority of the donations made by these occupations are to democratic organizations.



By including the election data by state we can look in aggregate and see how often the donation data diverges from the presidential election results. Meaning how often does an individual donate to a political party that did not win in the general presidential election in their state.

Grouping by state and occupation, summing up the number of donations for each party, then adding a column which indicates if that majority of that occupation per state donated to democratic or republican organizations.



Here we can see that out of our the majority of the donations matched the state that the political party of the candidate that received the most votes in the state the person resided in for a given year.

Part 4 - Inference

I will be using binomial logistic regression with the two predictors, residence in a blue state and employment in a STEM field, as my predictors for donating to a democratic candidate. In addition I will use both as predictors in a single model. I will then compare the fit using Akaike information criterion (AIC).

STEM As Single Predictor

First we I fit the model using employment in a STEM occupation as our single predictor.

```
stemResult <- glm(donateDem ~ stem, data = filteredPartyDf, family = binomial)
summary(stemResult)
```

```
##
## Call:
## glm(formula = donateDem ~ stem, family = binomial, data = filteredPartyDf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9201   0.5870   0.7282   0.7282   0.7282
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.191939   0.002016  591.34   <2e-16 ***
## stemTRUE    0.479193   0.015013   31.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1526206  on 1411519  degrees of freedom
## Residual deviance: 1525077  on 1411518  degrees of freedom
## AIC: 1525081
##
## Number of Fisher Scoring iterations: 4
```

$$logit(p_i) = 1.191939 + 0.479193 \times stemoccupation$$

What this means according to our model is that if a randomly selected person who donates to a political party is not employeed in a STEM field their probability of donating to a democrat is

We set stem_occupation = 0 and solve for p

$$\frac{e^{1.191939}}{1 + e^{1.191939}} = 0.7670877$$

$\hat{p}_i$=0.7671

subsequently the probability if they are employed in a STEM field is equal to

$$\frac{e^{1.191939+0.479193}}{1 + e^{1.191939+0.479193}} = 0.84172668778$$

$\hat{p}_i$=0.8417

Blue State

```r
demResult <- glm(donateDem ~ blueState, data = filteredPartyDf, family = binomial)
summary(demResult)
```

```
##
## Call:
## glm(formula = donateDem ~ blueState, family = binomial, data = filteredPartyDf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7864   0.6732   0.6732   0.6732   1.3857
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.477424   0.006622   -72.1   <2e-16 ***
## blueStateTRUE  1.846412   0.006968   265.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1526206  on 1411519  degrees of freedom
## Residual deviance: 1454603  on 1411518  degrees of freedom
## AIC: 1454607
##
## Number of Fisher Scoring iterations: 4
```

$$logit(p_i) = -0.477424 + 1.846412 \times blueState$$

Which means the probability of a random donating person donating democrat if they do not live in a blue state is equal to

$$\frac{e^{-0.477424}}{1 + e^{-0.477424}} = 0.38286059434$$

$\hat{p}_i$=0.3829

And the probability of a random donating person donating to a democrat if they do live in a blue state is

$$\frac{e^{-0.477424+1.846412}}{1 + e^{-0.477424+1.846412}} = 0.79721660053$$

$\hat{p}_i$=0.7972

STEM Occupation + Blue State

```
##
## Call:
## glm(formula = donateDem ~ stem + blueState, family = binomial,
##     data = filteredPartyDf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9967   0.5412   0.6763   0.6763   1.3905
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.488228   0.006635  -73.58   <2e-16 ***
## stemTRUE       0.488241   0.015458   31.59   <2e-16 ***
## blueStateTRUE  1.847023   0.006972  264.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1526206  on 1411519  degrees of freedom
## Residual deviance: 1453502  on 1411517  degrees of freedom
## AIC: 1453508
##
## Number of Fisher Scoring iterations: 4
```

Finally running the model on both

$$logit(p_i) = -0.488228 + 0.488241 \times stemOccupation + 1.847023 \times blueState$$

live in blue state	STEM profession	$\hat{p}_i$
yes	yes	0.8638
yes	no	0.7956
no	yes	0.5000
no	no	0.3803

Best Fit Using AIC

When looking at the AIC for the three different models, using just STEM we have 1525081, using Blue State we have 1454607, and using both we have 1453508.

According to OpenIntro Statistics[5]

> Just like multiple regression, we could trim some variables from the model. Here we'll use a statistic called Akaike information criterion (AIC), which is an analog to how we used adjusted R-squared in multiple regression, and we look for models with a lower AIC through a backward elimination strategy.

Given our AIC values this would seem to indicate that the best fit model is the model that includes both the state of residence and if the person is employed in a STEM occupation.

Verifying Logistic Regression Conditions

There are two key conditions for fitting a logistic regression model:

1. Each outcome Y_i is independent of the other outcomes.
2. Each predictor x_i is linearly related to $\text{logit}(p_i)$ if all other predictors are held constant

For condition one we can safely assume that whether an individual donates to a democrat or republican is generally independent across the entire donating population of the United States.

Part 5 - Conclusion

My results show the combination of the two predictors, employment in a STEM field and residence in a blue or red state, result in the best fit prediction model for predicting which political party a random donating individual is likely to donate to using the Akaike information criterion (AIC) value as the selection criteria for selecting the best fit.

In addition, I was able to show that what state a person lives in provides a better fitting model for predicting political party donations over simply using employment in a STEM occupation again using AIC as the criteria for determining best fit. My results show that while there is predictive value in knowing if a person is employed in a STEM field when it comes to determining their political donation proclivities, it does not appear to be a better predictor than simply knowing which state a person lives in and whether the majority of that state voted for a particular party in the U.S. presidential election.

Improving & Expanding

I believe this research could be expanded upon and improved in a few ways.

For one I only used state of residence but the contribution data contained more detailed geographic information about where a contributed was located. Often certain cities or counties will vote overwhelming for a party that does not match the majority of the votes cast in the state. Including this data in the analysis could provide even stronger predicting powers of simply state of residence.

In addition, due to the size of the data I chose to use only a subset of the contribution data available. It is possible different trends could arise given the totality of the data that the FEC provides.

[5] Diez, D. M., Barr, C. D., & Cetinkaya-Rundel Mine. (2019). 9.5.3 Building the logistic model with many variables. In OpenIntro statistics (pp. 374–374). essay, OpenIntro, Inc.

When it comes to the data itself individuals who donate to political parties are a distinct subset of the population. Donating to a political party likely indicates a high level of political engagement which may bias towards certain political inclinations and could be impacting the data in some way. A more ideal data set may be simply individual surveys or other data collection methods.

Finally, it would be interesting to categorize other occupation types such as social services, unemployed or retirees, to see if there is any discernible donation patterns from those groups of individuals. Also, the definition of STEM is fuzzy and open to interpretation especially given the large amount of variability observed in the occupation data that the FEC collected so there may be more optimal ways of defining what constitutes a STEM occupation.

References

1. Diez, D. M., Barr, C. D., & Cetinkaya-Rundel Mine. (2019). 9.5.3 Building the logistic model with many variables. In OpenIntro statistics (pp. 374–374). essay, OpenIntro, Inc.
2. Contributions by individuals. FEC.gov. (n.d.). Retrieved December 6, 2021, from https://www.fec.gov/campaign-finance-data/contributions-individuals-file-description/.
3. MIT Election Data and Science Lab, 2017, "U.S. President 1976–2020", https://doi.org/10.7910/DVN/42MVDX, Harvard Dataverse, V6, UNF:6:4KoNz9KgTkXy0ZBxJ9ZkOw== [fileUNF]
4. Benbwieder. (2017, April 21). When scientists donate to politicians, it's usually to Democrats. FiveThirtyEight. Retrieved December 6, 2021, from https://fivethirtyeight.com/features/when-scientists-donate-to-politicians-its-usually-to-democrats/.