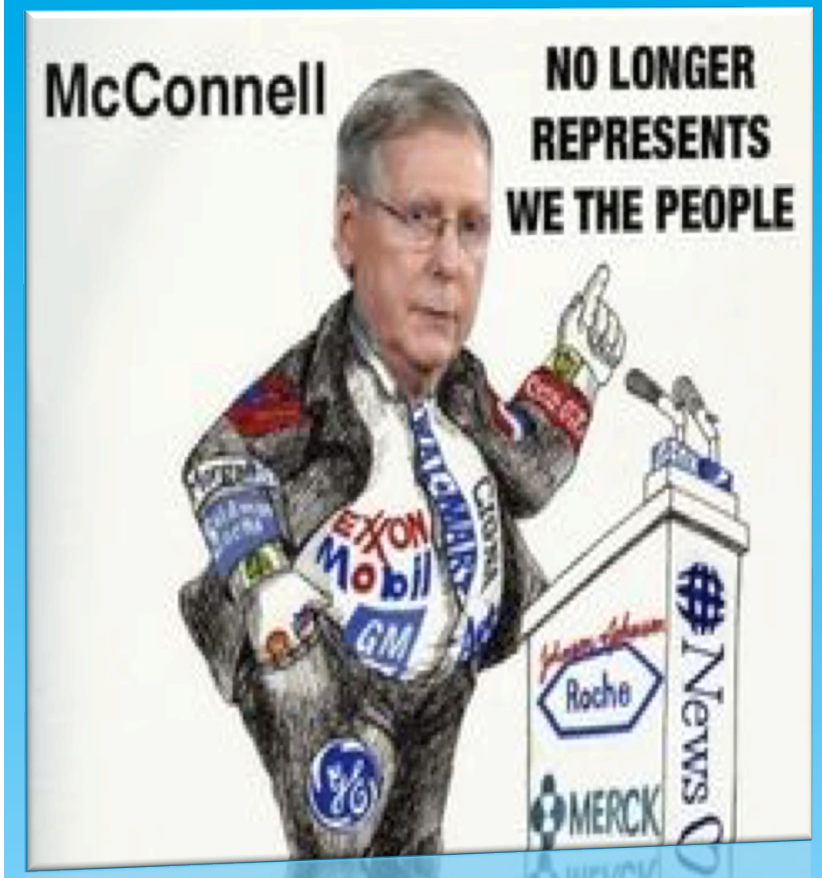
Money in Politics: The relationship between contributions and congressional voting

By: Alex Hubbard



Project Overview:

Research Question:

Is there a relationship between campaign contributions and congressional voting?

Null Hypothesis:

Using campaign contributions to predict congressional voting on armed forces and national security bills will not be more accurate than predicting congressional voting on the same bills using the baseline metric*.

* The baseline metric will be defined subsequently.

Data Set

Collection:

- All bill from the 101st to the 114th congress - 1989 to present (subset for project, Armed Forces and Nation Security) (api: govtrack.us)
- How each official voted on a bill (api: govtrack.us)
- Contributions to each political official since 1981 (bulk export sunlight_labs)
- Total rows: 51,120

Cleaning

Cleaning:

- Create SQL database for all contributions
  - Subset database to only be congressional contributions
  - Reduced where contributions were received from
    - Mapped 400+ contribution category codes to their 114 parent contribution industries
  - Create CSV for each of the 114 contribution industries
    - Because of the quantity of data, it is less expensive to open multiple CSV for industries I want rather than everything all at once.

Cleaning Cont.

Cleaning:

- Connect government id to campaign funding id (opensecrets_id)
- Running Tally of cumulative contributions & voting history for each vote that took place. i.e.:
  - Add columns for each campaign funding industry to show how much money each congressperson has accumulated up the date of each vote.
  - Add columns to track each persons voting history

Cleaning Cont.

Running Tally visualized:

GENERAL CONTRACTORS	LOBBYISTS	PUBLIC SECTOR UNIONS	bill_id	class_variable	date	opensecrets_id	party	prior_total	prior_yea_percent	prior_yea_total
+ 0 → 500		3000	hr2402-101	1	6/22/89	N00000010	R	1	1	1
+ 300 → 500	250	3250	hr3012-101	1	10/26/89	N00000010	R	2	1	2
etc. → 800	250	8850	hr2748-101	1	11/16/89	N00000010	R	3	1	3
800	250	8850	hr3072-101	1	11/17/89	N00000010	R	4	1	4
6100	5300	14150	hr486-101	-1	10/23/90	N00000010	R	5	1	5
6100	5300	18150	hr5803-101	-1	10/24/90	N00000010	R	6	0.833333333	5
6100	5300	18150	hr5313-101	-1	11/5/90	N00000010	R	7	0.714285714	5

Initial Exploration*: 2 Mistakes

1. Holist view of the data set:

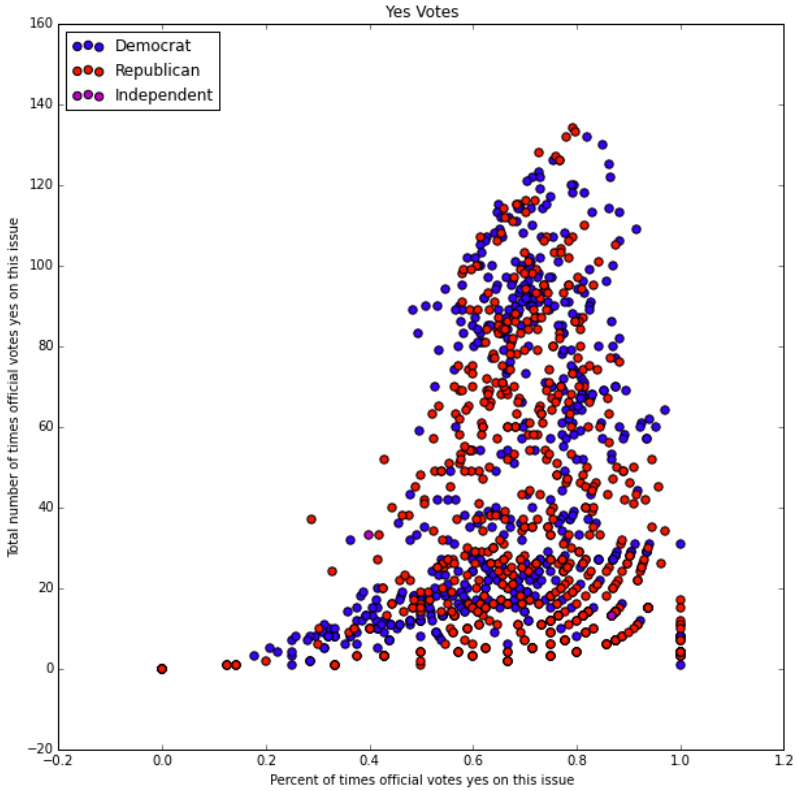
- Could not see the true relationship between contributions and votes.
- Instead, I was looking at two variables that were unlinked:
    1. Total money collected from contributions by industry.
    2. Everyone's voting record.
- How could a vote in 1990 be influence by money collected in 2007? It can't.

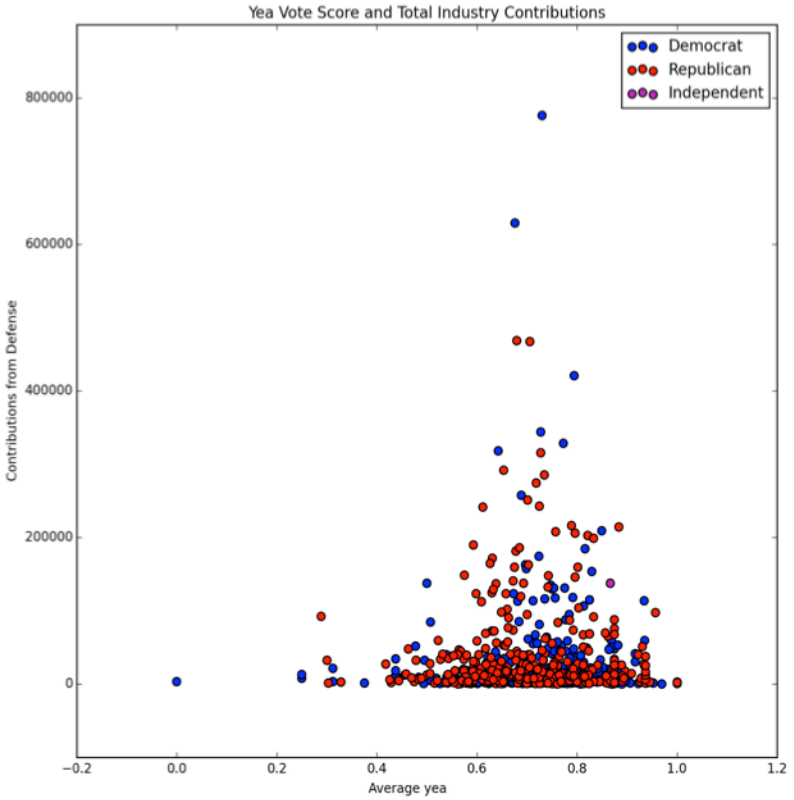
2. Looked at contributions and voting history together

- To truly determine if there is a relationship between money received by a campaign and how a congressperson votes contribution industries need to be the only predictor variables.

*Initial Exploration did not contain Running Tally

Initial Exploration: Plots



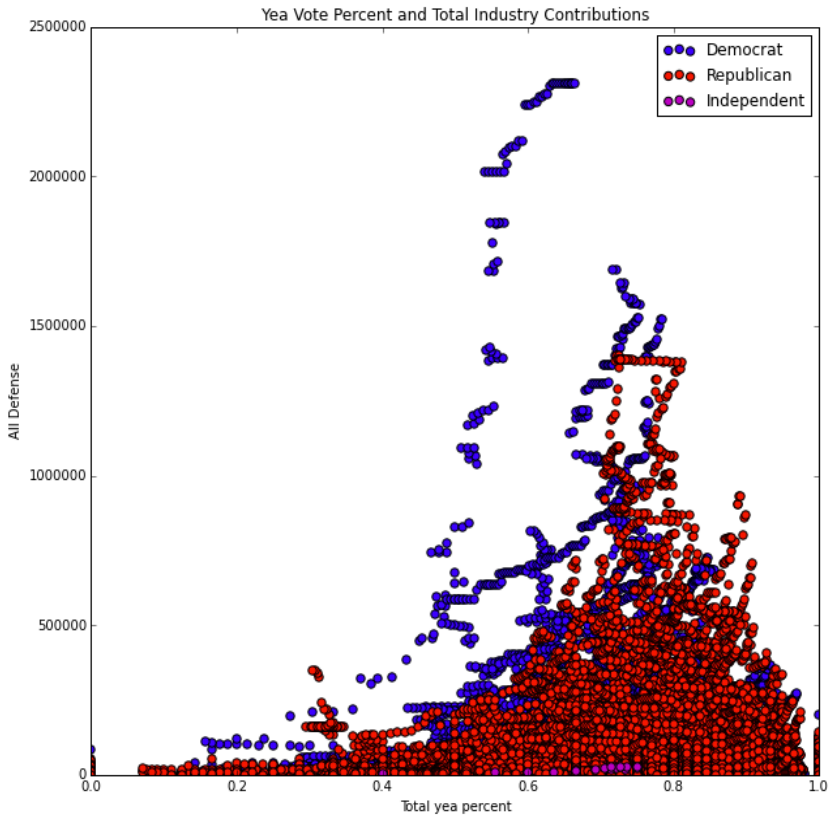


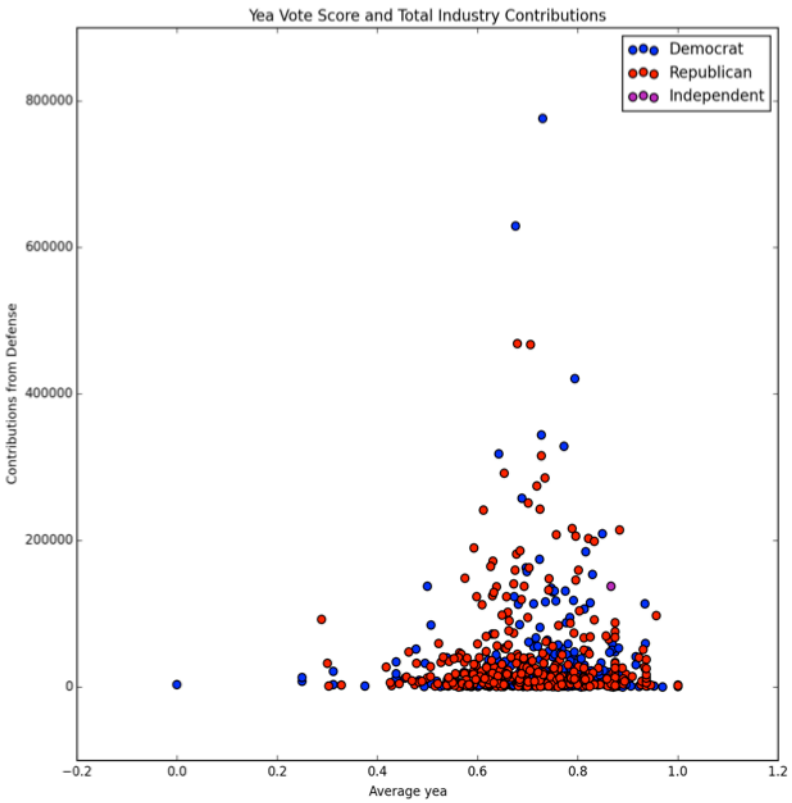
Re-cleaning

- Re-cleaning was only performed to add Running Tally statistics.
- Data set went from being sparse to being rich.
- Data is now ready for modeling!

Re-cleaning: Visualized

With Running Tally vs. Aggregated Contributions





Choosing Variables

- 13 industries were used for the final analysis.
  - ‘Abortion Policy Pro-Choice', ’Accountants', 'Business Services', ’Defense Aerospace', 'Defense Electronics', 'Democratic Liberal', 'Environment', 'Foreign & Defense Policy', 'General Contractors', 'Lobbyists', ’Misc Defense', 'Public Sector Unions', 'Women’s Issues'
- Since the bills analyzed are on “armed forces and national security”, every contribution industry with the word “defense” is included.
  - Four in total.

Choosing Variables Cont.

- The other nine industries were included by analyzing a subset of the data. Analysis was performed as follows:
  - Subset 56 congress-people who participated in the most votes.
  - Subset of 32 industries that were rich in data.
  - Ran logistic regression to find coefficient scores.
  - Ran random forest to find importance scores.

Modeling

Modeling was done examining two sets of predictor variables:

1. Running Tally of contributions by all 13 industries.
2. Running Tally of each congressperson's voting history.

Modeling was split in order to compare how contributions compared to voting history.

Class Variable:

- Voting result
- 1 = yea, 0 = present, but abstained from vote, -1 = no

Modeling

Setting the baseline:

- The baseline accuracy was determined by the most common class variable ('yea' votes).
  - Total instances of yea votes ÷ total votes in data set

If the model is unable to beat the baseline then the null hypothesis cannot be rejected.

- i.e. using campaign contributions is insufficient for predicting how congress people will vote.

Modeling

Because of the size of the data set the modeling was done on a sample of the data. Additionally, a random state condition was included to ensure replicability.

To test the performance of the analysis, 2 baseline metrics were created. One for the entire data set, and on the sample:

1. Baseline for the overall data set: 70.81%
2. Baseline for the sample data set: 70.7%

Modeling

Algorithms used:

- PCA for dimensionality reduction
- SVM to predict vote

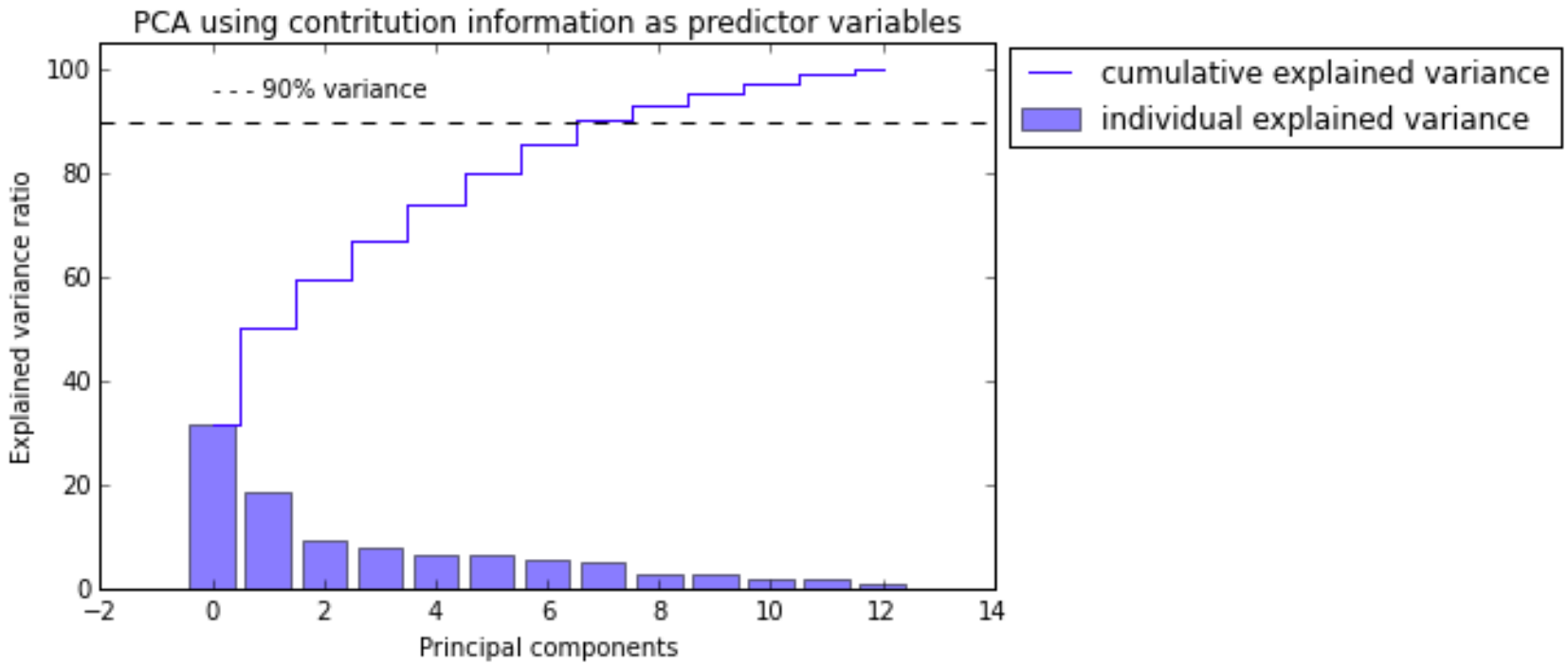
Both models* were subset by 10,000 rows**

* One for contributions, one for vote history

** The same random state was used to ensure replicability

Model #1:
Contributions as Predictor

Number of components used for PCA was when variance reached 90%



Model #1:
Contributions as Predictor

Parameters to use for SVM were determined by performing GridSearch on a subset of 500 data points.

Parameters tested:

- 'C': np.linspace(.001, 10, 10), 'kernel': ['poly', 'rbf', 'linear'], 'degree': range(1,4), 'gamma': np.linspace(.001, 10, 10)

Parameters used:

- C=0.001, degree=3, gamma=1.1119999999999999, kernel='poly'

Model #1:
Contributions as Predictor

Findings:

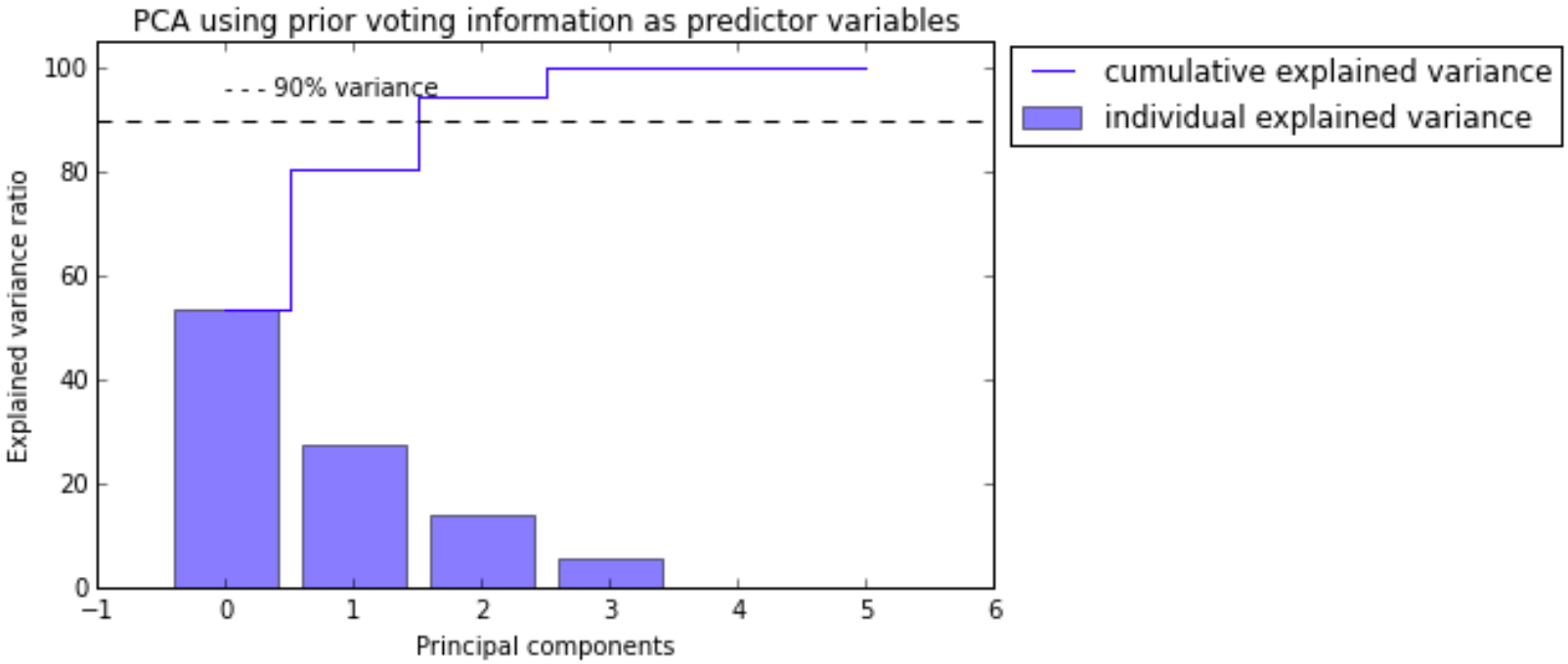
- The accuracy of the model was 82%
- This is ≈ 11% higher than the baseline
- Rejection of null hypothesis!

Below are the results of using contritution information as predictor variables:

	precision	recall	f1-score	support
-1	0.00	0.50	0.01	4
0	0.00	0.00	0.00	0
1	1.00	0.70	0.83	1996
avg / total	1.00	0.70	0.82	2000

Model #2:
Voting History as Predictor

Number of components used for PCA was when variance reached 90%



Model #2:
Voting History as Predictor

Parameters to use for SVM were determined by performing GridSearch on a subset of 100 data points.

Parameters tested:

- 'C': np.linspace(.001, 10, 10), 'kernel': ['poly', 'rbf', 'linear'], 'degree': range(1,4), 'gamma': np.linspace(.001, 10, 10)

Parameters used:

- C=0.001, degree=3, gamma=3.3340000000000001, kernel='poly'

Model #2:
Voting History as Predictor

Findings:

- The accuracy of the model was 83%
- This is ≈ 12% higher than the baseline
- This is only 1% higher than using campaign contributions as the predictor

Below are the results of using prior voting information as predictor variables:

	precision	recall	f1-score	support
-1	0.00	0.00	0.00	0
0	0.00	0.00	0.00	0
1	1.00	0.70	0.83	2000
avg / total	1.00	0.70	0.83	2000

Key Learnings

1. Using contribution data as the predictor variables beat the baseline accuracy. In other words, there is a relationship between how much money a congressperson gets from an industry, and how they vote on armed forces and national security bills. It should be stated that it is unclear if the contribution causes the voting pattern, or if the voting pattern causes the contribution.

- Accuracy of campaign contributions: 82%
- Baseline accuracy: ≈71%

Key Learnings

2. There was an observed relationship between voting history and votes on armed forces and national security bills. When using voting history as the predictor variables the model beat the baseline accuracy.

- Accuracy of voting history: 83%

Key Learnings

3. Using contributions to predict how congressional officials vote on armed forces and national security bills is just about as accurate as using an individual congressperson's voting history.

Extensions

- NLP to create feature that determines if the bill is helpful or harmful the its subject, and retest models.
  - E.g. not all bills where the top_subject as "Armed forces and national security" will be helpful to armed forces and national security.
  - Some of the bills will call for budget cuts, reduction in personnel, etc.
- Perform analysis on other bill_subjects
- Predict congressional voting with > 90% accuracy.
  - Use campaign contributions as a section of the pipeline