An Analysis on How Political Contributions Influence the Voting of Congressmen

Group: The Congressmen
By: Joseph Hudson, Brandon Escobar-Campos, Finn Mokrzycki, Surya Venkatraman

Table of Contents

Introduction/Background

Our original motivation as a group to conduct this analysis was to see the extent political contributions from sectors have in the decision making of members of congress. We decided to focus our project on this question: Between 2006 and 2012, do total sector political contributions towards a congressman influence whether that congressman votes against the majority of their political party on bills and joint resolutions?

The data collected comes from three databases created by Adam Bonica from Stanford University. The first database consists of a dataset that provides demographic information about congressional candidates (recipients of political contributions) from 1979-2022. Such demographic information includes the congressional cycle year they participated in, name, political party, primary election percent (election within their political party), general election percent (election against the opposing political party), and incumbency status at the time of the respective general election. As stated previously, we're only interested in congressional candidates that won their respective general election, which would allow them to vote on bills and joint resolutions in the upcoming congressional session (a two-year period that begins and ends on the 3rd of January following each November election), and for those that are considered in the 2006-2012 time period. So, we filtered the original dataset to fulfill such requirements.

The second database consists of two datasets: congressional bills (which includes both bills and joint resolutions) and congressional votes between 2006-2014. Due to the discrepancy of the time periods between the first and second database, we restricted ourselves with the second time period (2006-2014) and then further conditioned it based on the rationale explained in the previous section - an incomplete congressional session from 2013 to 2014. In congressional bills, it contained information about each bill that was voted on throughout the time period, the congressman who sponsored or cosponsored the bill, and the topic weight values for 25 unique topics. Using these topic weight values, we assigned an attribute to each bill based on its highest topic weight. This is used to categorize each bill with a relevant topic. For congressional votes, it provided data on the voting decision of every congressman for every bill that was voted on throughout our time period.

The third database consists of a dataset that provides information about political contributions from Fortune 500 companies to congressmen throughout our time period. Within the dataset, it contains the company's name, industry, sector, the congressman they donated to, and the amount they contributed to that congressman. More generally, the dataset provides each individual transaction a company made throughout our time period to congressmen.

The final dataset used in our analysis model consisted of merging each of the datasets from each database. To start, we merged the dataset from the first database, demographic information about each congressman, with the dataset from the third database, fortune 500

contributions, by a unique identifier number, that's given to each congressman, and by congressional cycle, since some congressmen can receive donations from multiple cycles. Essentially, it's the same as the dataset from the third database except that we now have values about the demographics of each congressman who received contributions. Next we merge the two datasets from the second database, congressional voting and bills, together by a unique identifier number for the bill. So, this merged dataset is the same as the voting dataset except we have additional columns from the bills dataset. Lastly, we left merge the second merged dataset with the first merged dataset by the unique identifier number of each congressman and by the election cycle. So, the final dataset should contain each vote by a congressman, multiplied by the amount of distinct contributions the congressman received during the relevant election cycle.

In regard to the source of the data and whether it is trustworthy, the data for congressional bills (second database) comes from the Government Publishing Office's Federal Digital System (FDsys) and Congressional Research Service. The data for congressional votes (second database) comes from the website *voteview*, which is currently managed by UCLA, but was originally managed by the University of Georgia. The data for recipients (first database) and company information (third database) comes from Stanford's DIME (Database on Ideology, Money in Politics, and Elections) and DIME+ databases, which comes from the FEC (Federal Elections Commission). Given that all the data sources are collected by government agencies (aside from the voting record, which is still a public source that is easy to attain and verify correctness), we believe it to be trustworthy data. Institutions like Stanford University and the University of Georgia also add to the credibility of it.

One type of data we would have liked to use is one that identifies committee positions for each congressman. In Fowler et al. (850), it reasoned that if political contributions did influence changes in votes, it would be in the incentive of a company, or sector in general, to mainly use their contributions on congressmen with powerful positions like chairs of committees. A congressman that holds any committee chair has a unique influence in that they act as the gatekeepers of bills that are voted on, relevant to the topic of the committee. For example, in the House Committee on Foreign Affairs, the head chair is introduced to any bill, or joint resolution, that's relevant to foreign affairs and they control the process on whether it's brought to the actual floor of the House for a vote.

For this project, we'd like to know **between 2006-2012, if political contributions, by sector, can influence the decision of a congressman in voting against the majority of their political party in bills and joint resolutions**. Our original idea was to determine if campaign contributions from sectors influenced the vote outcome of the congressman, but this led to issues when it came to thinking of how to measure the response variable for our model, and what the predictor variables should be.

Analysis

First, we performed **selective filtering**. For congressmen, we looked at donations between Republicans and Democrats. Since we were going to be predicting voting against the majority of the party, it was difficult to determine what that actual party was for those other candidates, which makes it difficult to determine who that candidate is actually voting against. We also only consider votes of yes or no for determining votes for the party, which does not include candidates that abstain from voting.

For calculating the Vote Against Majority label for a candidate, we first created a temporary dataset to get the majority vote for each party per bill. We grouped by the bill_id and Party to get counts. Then, we joined that temporary table to bills and votes. Then, we calculated the most common vote for each party per bill, and set that as the majority vote. Then, we created our variable Vote.Against.Majority that compared that candidate within his or her party to see if they voted the same way as the majority in their party (0) or not (1).

One very important manipulation was **which donations we use for predictions**. **We use donations from the previous election cycle** to predict the vote in the session of Congress. For example, when the donation cycle says 2006 (donations from 2005-2006), this matches votes for the session of Congress that runs January 2007 to December 2008. This is important because it allows us to use donations in a way that can reveal a more immediate effect, instead of aggregating previous donations from multiple cycles. This also does not work for candidates who enter Congress for their first time.

For the **Total Donations** variable, we noticed that some donations were really large, creating outliers, so we performed a log transformation. By performing a log transformation, this helped stabilize the scale. Additionally, this helps with our linear regression (lasso) model to maintain consistency, shown in the Results section.

These are the bill topics that we evaluate in our analysis, along with information on the number of bills. A bill's topic is determined by its top topic weight, a variable originally included in the dataset. The weight columns are the thresholds that we set for minimum weight. For example, 0.30 for healthcare means that a bill had healthcare for the highest topic weight, and that weight was at least 0.3. Here are the counts for bills with top topic weights greater than or equal to 0.4:

	Top_Topic	`n()`	`min(cycle)`	`max(cycle)`	`min(Top_Topic_Weight)`	`max(Top_Topic_Weight)`
	<chr>	<int>	<int>	<int>	<dbl>	<dbl>
1	tw.banking.and.finance	2	2006	2008	0.401	0.446
2	tw.economy	1	2008	2008	0.531	0.531
3	tw.energy	3	2006	2010	0.428	0.514
4	tw.environment	4	2006	2010	0.414	0.531
5	tw.healthcare	1	2008	2008	0.450	0.450

There are too few bills here, even though these would be the ones that are most connected to the topic. Additionally, not all cycles are reflected in these collections of bills. Here are the counts for topic weight ≥ 0.3:

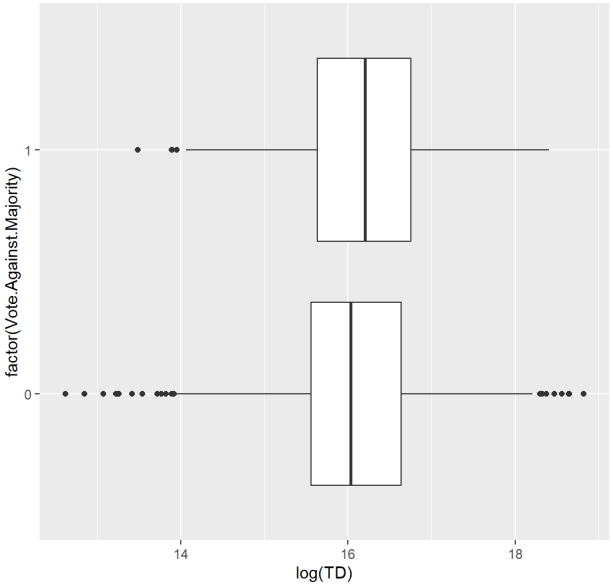
	Top_Topic	`n()`	`min(cycle)`	`max(cycle)`	`min(Top_Topic_Weight)`	`max(Top_Topic_Weight)`
	<chr>	<int>	<int>	<int>	<dbl>	<dbl>
1	tw.banking.and.finance	13	2006	2010	0.303	0.446
2	tw.economy	2	2008	2008	0.370	0.531
3	tw.energy	11	2006	2010	0.307	0.514
4	tw.environment	7	2006	2010	0.345	0.531
5	tw.healthcare	8	2006	2010	0.301	0.450
6	tw.transportation	6	2006	2008	0.302	0.395

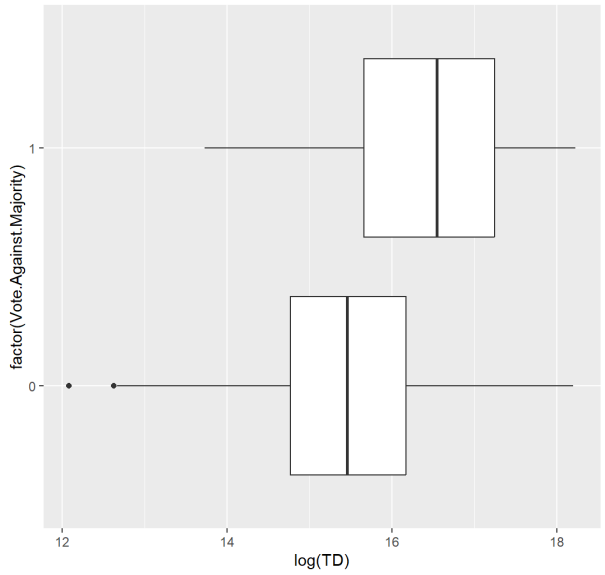
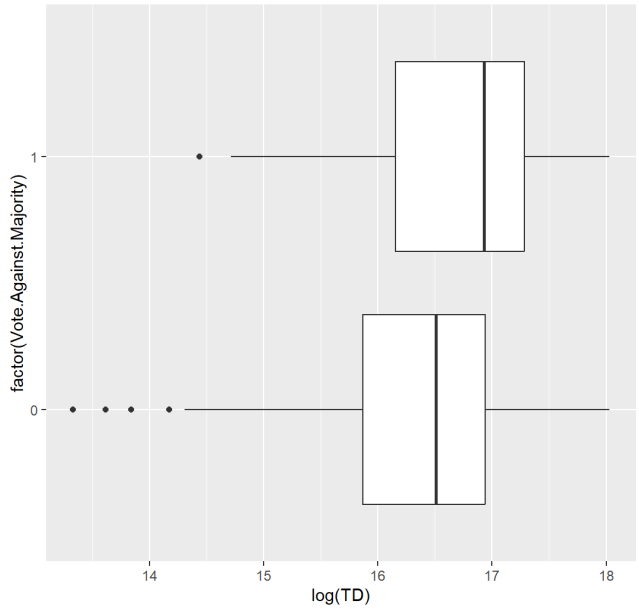
There are still too few bills at 0.3 and there are no bills from cycles 2010 for some categories. After using the minimum topic weight 0.25, we then get this table:

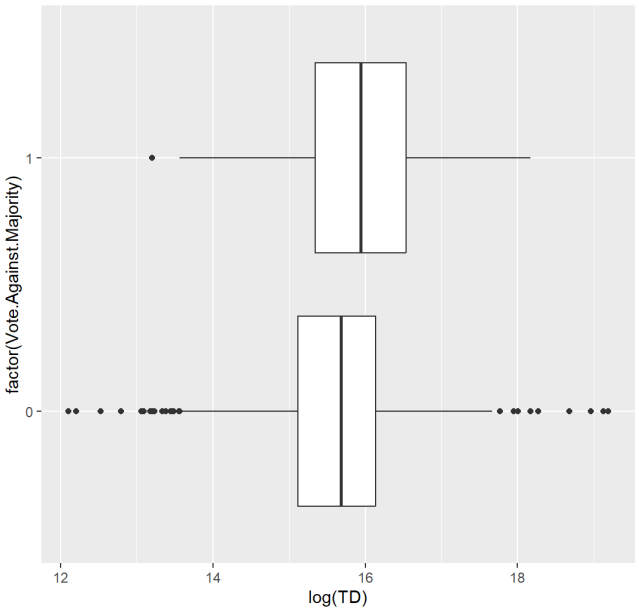
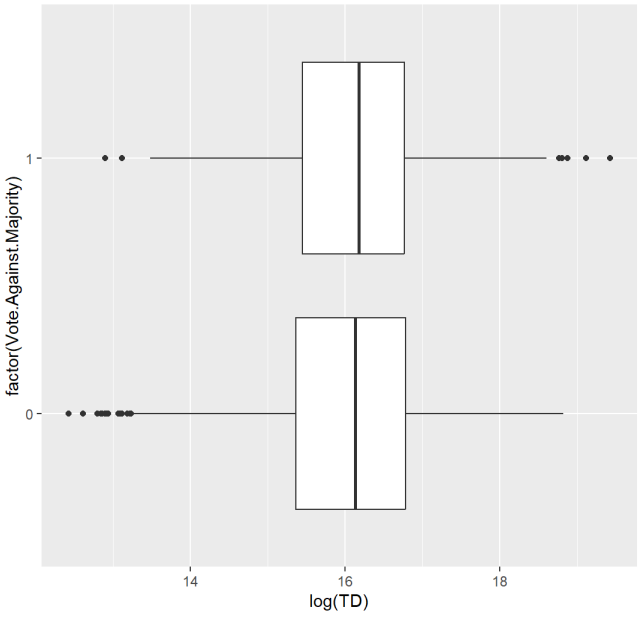
	Top_Topic	`n()`	min(cycle)	max(cycle)	min(Top_Topic_Weight)	max(Top_Topic_Weight)
	<chr>	<int>	<int>	<int>	<dbl>	<dbl>
1	tw.banking.and.finance	17	2006	2010	0.250	0.446
2	tw.economy	9	2006	2010	0.254	0.531
3	tw.energy	17	2006	2010	0.258	0.514
4	tw.environment	10	2006	2010	0.270	0.531
5	tw.healthcare	14	2006	2010	0.272	0.450
6	tw.labor	2	2006	2010	0.266	0.270
7	tw.transportation	12	2006	2010	0.264	0.395

We decided to set the topic weight to be 0.25, since each cycle is included in this time period, and a large number of bills are available for each topic.

For the breakdown between total donations and voting against the majority per topic, here are the breakdowns in boxplots (the donations are transformed with log() function):

Topic	**Model Summary**
Energy	

Healthcare	
Transportation	

Financial	
Environment	

Looking at these breakdowns, it looks like for some sectors, there is some level of difference, such as Healthcare and Transportation.

In terms of data cleaning, we had to extract relevant information from three different databases. One important aspect that we had to clean was the types of congressional bills the original data included. Because we're looking at both the Senate and the House of Representatives, there are some types of bills that aren't important to the context of our project because they're more relevant to the internal procedures of Congress. As stated previously in our

Introduction section, we're looking only at Bills and Joint Resolutions because they both follow the same congressional procedure to become law. Another aspect of data cleaning conducted was aggregating political contributions of individual companies, from the Fortune 500 companies dataset, based on sector. The decision behind this is mainly due to our emphasis on creating a generalized model that is interpretable and more broad in application, rather than providing a model that is more limited in its predictive ability because of its heavy emphasis on individual companies. Additionally, the decision was made to look at a select number of sectors: energy, transportation, healthcare, financial, and environment. The reason for solely looking at these sectors is because they had direct relationship, in terms of relevance, with some of the bill topics that were assigned in the bill dataset. In other words, we didn't want to look at how a sector's political donations influenced congressmen from voting against their party's majority on a bill that isn't relevant, or related, to that sector. Doing this allows us to filter on the bills that are particularly relevant to a sector's interest, which is more likely to allow us to determine the true relationship between political donations and political influence.

We performed a logistic regression and a decision tree in order to determine the influence sector political donations had on the decision of a congressmen to vote against their party's majority vote. The reason behind performing two models is because each provides advantages and disadvantages in their implementation. Statistically, logistic regression is a generalized linear model, which means that the response variable, the log-odds of a congressman's vote against their party's majority vote, ultimately depends on the various inputs and parameters we decide to implement. This could be advantageous since assuming there's a linear relationship between the independent variables and the log-odds of the dependent variable makes it easier to interpret. However, such an advantage in interpretation may lead to an oversimplification of the true, underlying relationship between the variables. Such a problem is generally true for less flexible models. To counteract this, we decided to include a decision tree model that would complement in covering for the disadvantages logistic regression faces. This is due to it being a more flexible model, given that it's a machine learning model. Additionally, it may help with covering the potential complex, non-linear relationship the variables may have that logistic regression cannot demonstrate. We're aware of the potential difficulties a decision tree may have, such as overfitting and instability, but having the results of two, relatively different models can help broaden our understanding in solving the ultimate question of whether sector donations influence whether a congressman votes against the majority vote of their political party on bills proposed throughout 2006-2012.

The chosen analysis tools were sufficient for the task of analyzing how donor contributions influenced the voting of congressmen, because they were some of the best options based on the data we were modeling. With predicting a qualitative response variable in whether a congressmen voted against their political party on a particular bill or resolution, a focus on categorical modeling was needed. In terms of using a logistic regression model, our data satisfies

its assumptions. Our data has a binary response variable in whether a congressman voted on a particular bill the same way the majority of their political party did. Our data has no multicollinearity between variables for our linear model, because we are only using political contributions received as our only explanatory variable. Finally, in the case of assumption of independence between observations for logistic regression, it is safe to assume how individual congressmen are voting on bills to not be independent in the sense that these people are talking with each other as they try to come to a decision with their vote. With that being said, there is no transformation that we can make to the data but reasonably this shouldn't affect our analysis with the logistic regression model. In the case of the decision trees, we wanted to use a nonlinear method not only to see how it performs, but also to find new information offered. While the heart of our project is around donor contributions, there are many other potential explanatory variables we have data on. With the decision trees, we have an easy to interpret way to gauge the relationships these other explanatory variables had with the decisions made by congressmen.

We didn't consider many other modeling tools for our analysis. We wanted to analyze the data with a generalized linear model and a non-linear model. We could have looked to a linear discriminant analysis (LDA) for another option as a linear model, but given our data met the assumptions for logistic regression we never applied a LDA model to our data. In the case of the non-linear model, we wanted to use something that had easy interpretation and help fill in the potential disadvantages of using a linear model. As a result, we focused only on utilizing a decision tree, given it is one of the most interpretable qualitative response models.

Results

Here are the results of the logistic regression models.

Topic	Model Summary

Energy	`Call:`
	`glm(formula = Vote.Against.Majority ~ log(TD), family = binomial,`
	`  data = filtered.df)`
	
	`Deviance Residuals:`
	`  Min  1Q  Median  3Q  Max`
	`-0.6108  -0.5221  -0.5001  -0.4746  2.2204`
	
	`Coefficients:`
	`  Estimate Std. Error z value Pr(>|z|)`
	`(Intercept)  -4.3769  1.2145  -3.604 0.000314 ***`
	`log(TD)  0.1484  0.0751  1.976 0.048164 *`
	`---`
	`Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1`
	
	`(Dispersion parameter for binomial family taken to be 1)`
	
	`  Null deviance: 1538.8  on 2090  degrees of freedom`
	`Residual deviance: 1534.8  on 2089  degrees of freedom`
	`AIC: 1538.8`

Healthcare	`Call:`
	`glm(formula = Vote.Against.Majority ~ log(TD), family = binomial,`
	`  data = filtered.df)`
	
	`Deviance Residuals:`
	`  Min  1Q  Median  3Q  Max`
	`-1.2964  -0.5031  -0.3515  -0.2157  2.9699`
	
	`Coefficients:`
	`  Estimate Std. Error z value Pr(>|z|)`
	`(Intercept) -18.74362  1.42919  -13.12  <2e-16 ***`
	`log(TD)  1.04499  0.08777  11.90  <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: 1324.4  on 1914  degrees of freedom`
	`Residual deviance: 1149.4  on 1913  degrees of freedom`
	`AIC: 1153.4`

Transportation

```
Call:
glm(formula = Vote.Against.Majority ~ log(TD), family = binomial,
    data = filtered.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9277  -0.6581  -0.5247  -0.4256   2.5179

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.1939     1.9359  -6.815 9.41e-12 ***
log(TD)       0.6975     0.1160   6.015 1.80e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1120.6  on 1263  degrees of freedom
Residual deviance: 1079.7  on 1262  degrees of freedom
AIC: 1083.7
```

Financial

```
Call:
glm(formula = Vote.Against.Majority ~ log(TD), family = binomial,
    data = filtered.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9939  -0.5615  -0.5060  -0.4299   2.4599

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.54476    0.87848  -9.727  < 2e-16 ***
log(TD)      0.42195    0.05525   7.638 2.21e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3281.6  on 4219  degrees of freedom
Residual deviance: 3221.2  on 4218  degrees of freedom
AIC: 3225.2
```

Environment✝

```
Call:
glm(formula = Vote.Against.Majority ~ log(TD), family = binomial,
    data = filtered.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6226  -0.5206  -0.5000  -0.4718   2.2337

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.59741    1.01506  -4.529 5.92e-06 ***
log(TD)      0.16236    0.06305   2.575     0.01 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2423.2  on 3305  degrees of freedom
Residual deviance: 2416.6  on 3304  degrees of freedom
AIC: 2420.6
```

✝:This does not have a specific stock sector, but relies on Energy and Utilities companies. In addition, this also relies on bills with Energy and Environment topics.

Given that these donations are logged, the interpretation of a one unit increase is more complicated for the models. For example, with environment, for each one unit increase in the log(Total Donations) for the previous election cycle, the log odds ratio increases by 0.16236, and the odds ratio of voting against their party versus voting with is multiplied by a factor of 1.1763, increasing 17.63%. The one interpretation that is useful though from the beta coefficients is that they are all positive, indicating an increase in total donations is related to an increase in voting against the party majority.

After looking at the boxplots for exploratory data analysis and seeing the difference by sector, it makes sense that the models that perform best based on AIC are Healthcare and Transportation. These ones had boxplots that showed the largest difference between the medians. Looking at each of these individually too, some of these models seem suspect for donation amount at probability = 0.5 . In order to calculate what the predicted donation value would be at this point, we find out when the equation $\boldsymbol{\alpha} + \boldsymbol{\beta}\log(x_{donations}) = 0$, since the equation for logistic regression is

$$\pi_{Vote.Against.Majority} = \frac{e^{\alpha+\beta log(x)}}{1+e^{\alpha+\beta log(x)}} = \frac{1}{1+1} = 0.5$$

when $e^{\alpha+\beta log(x)} = 1$

when $\boldsymbol{\alpha} + \boldsymbol{\beta}\log(x_{donations}) = 0$

Here are the donation amounts where $\pi = 0.5$ and max donation amount for the model, along with the 50th and 80th percentile donations and their prediction probabilities:

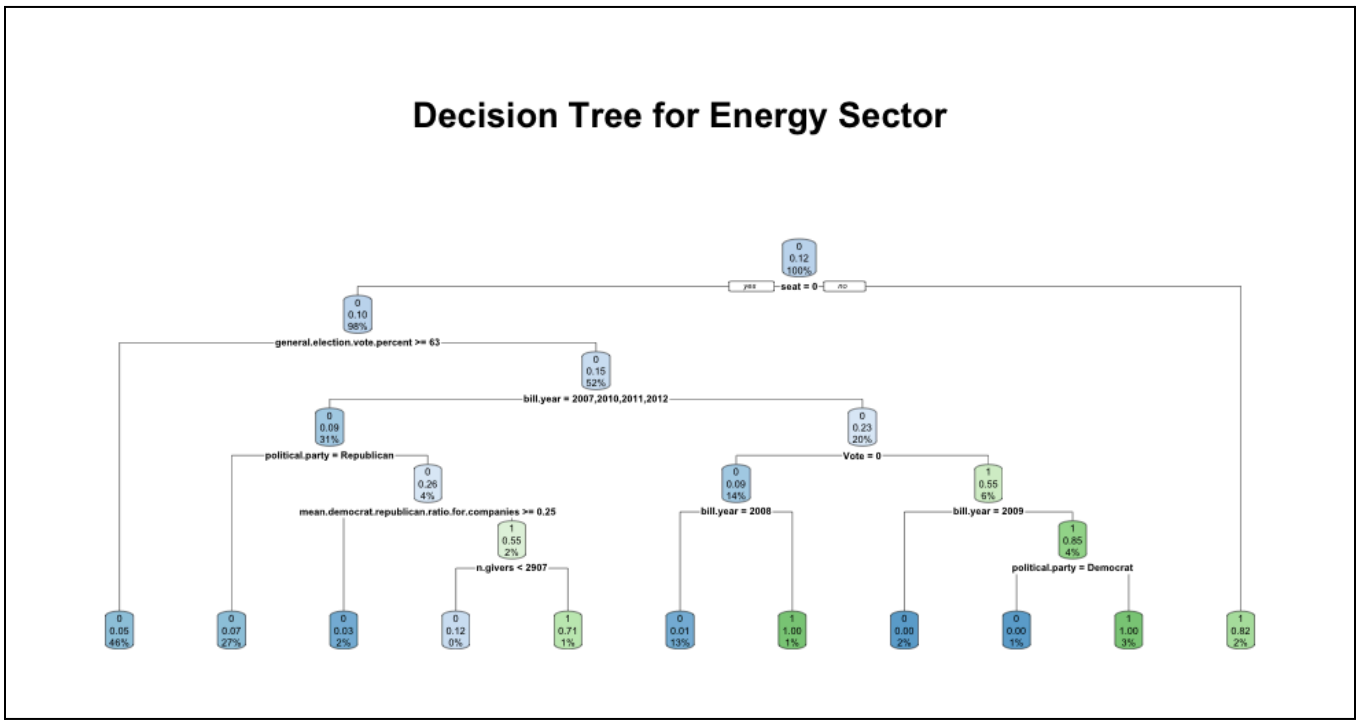
Topic	Donation Amount Where $\pi = 0.5$	Max Donation Amount from Data
Energy	$6.4424841e+12	$148,573,161
Healthcare	$61,629,432.30	$82,308,219
Transportation	$164,099,853.46	$67,466,017
Financial	$623,366,889.50	$215,411,835
Environment	$1.9840261e+12	$272,164,250

Sector	50th Percentile Donation	80th Percentile Donation	Probability with 50th Percentile Donation Amount (Median)	Probability with 80th Percentile Donation
Energy	$9,474,066	$20,351,410	0.1199407	0.1324432
Healthcare	$5,619,272	$14,048,181	0.07567012	0.1757848
Transportation	$16,566,425	$24,849,638	0.1680606	0.2113806
Financial	$6,648,624	$12,177,782	0.1283129	0.1596823
Environment	$8,684,337	$16,974,213	0.1188486	0.1307243

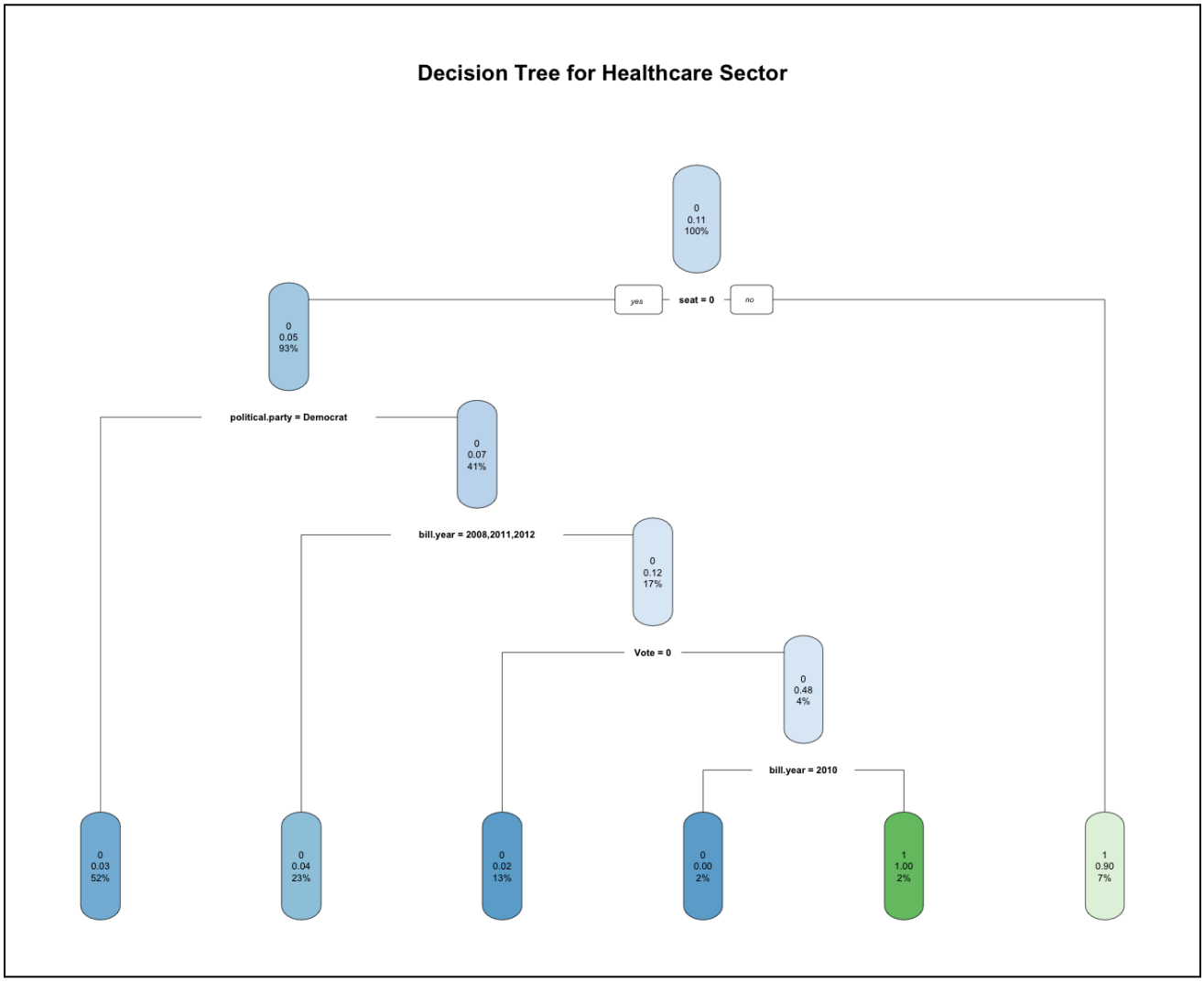
After looking at the donations required for each probability, the models do not seem useful for predicting the actual dollar amount required to get a congressman to vote against his party (especially based on the table showing donations for probability = 0.5); rather, these models indicate how much impact each donation has by topic and allows for comparison. For example, for energy, there is about a $11 million difference between the 50th and 80th percentile donations, but the increase in probability only increases from around 12% to 13.2%. For healthcare, the difference is about $9 million and the probability increases from 7.5% to 17.6%, which is over 10%. This information is important for identifying topics where donations may have more influence, such as healthcare and transportation, where these donations that are similar to the other topics result in larger increases in probability of voting against the party.

Here are the final decision tree models for each sector:

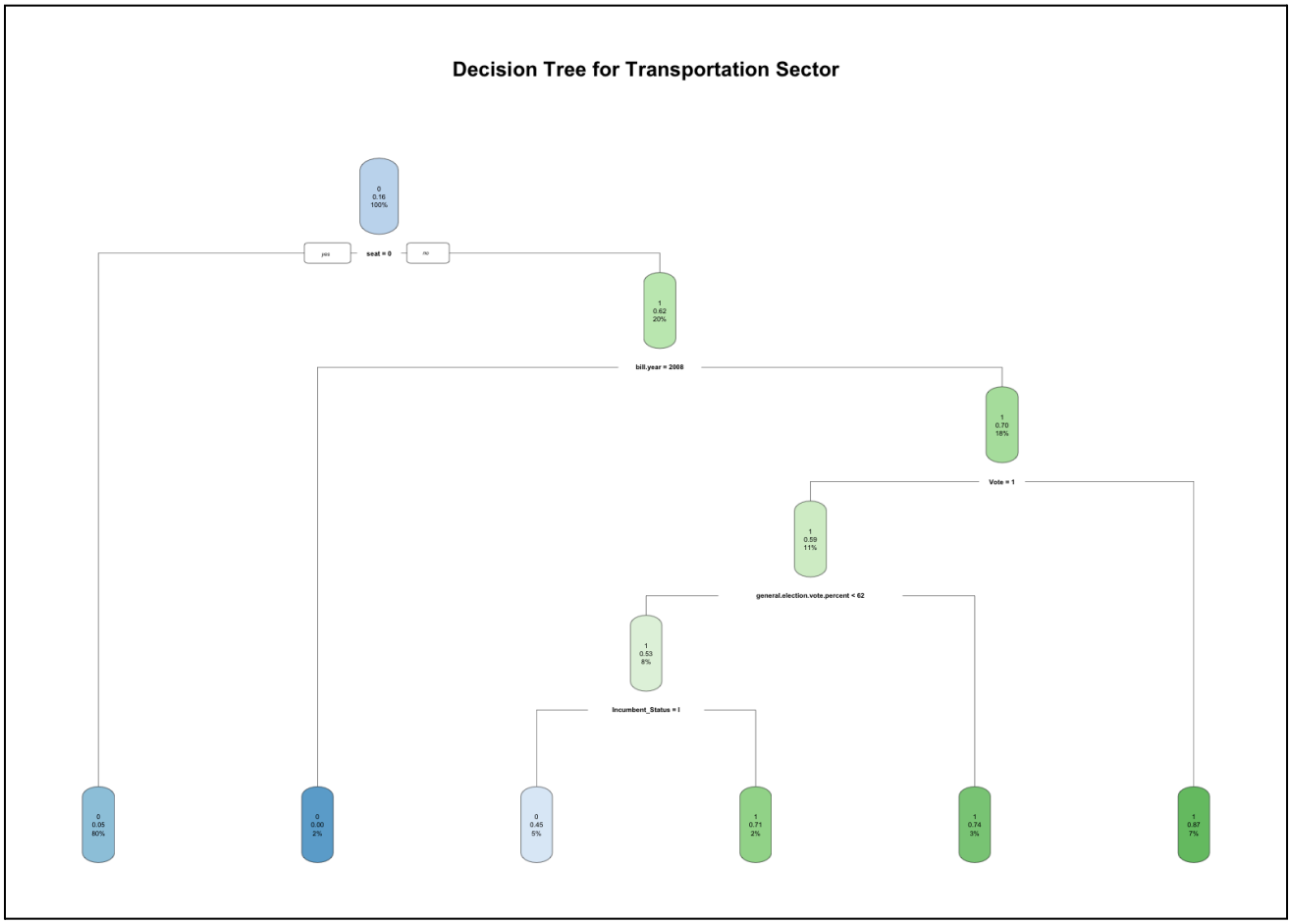
Energy Sector:



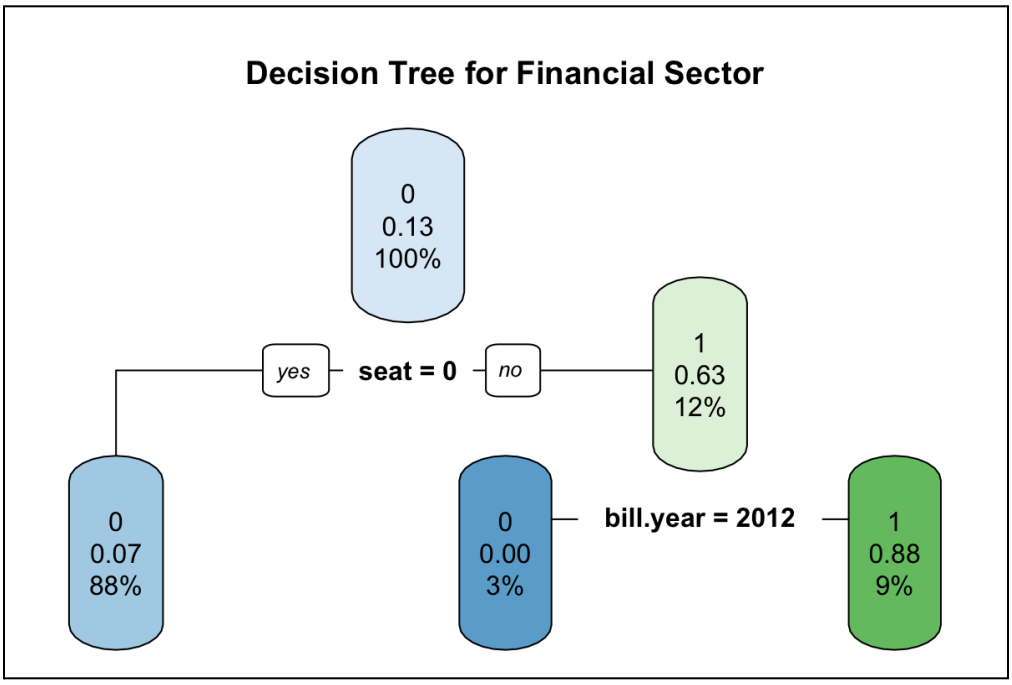
Healthcare Sector:



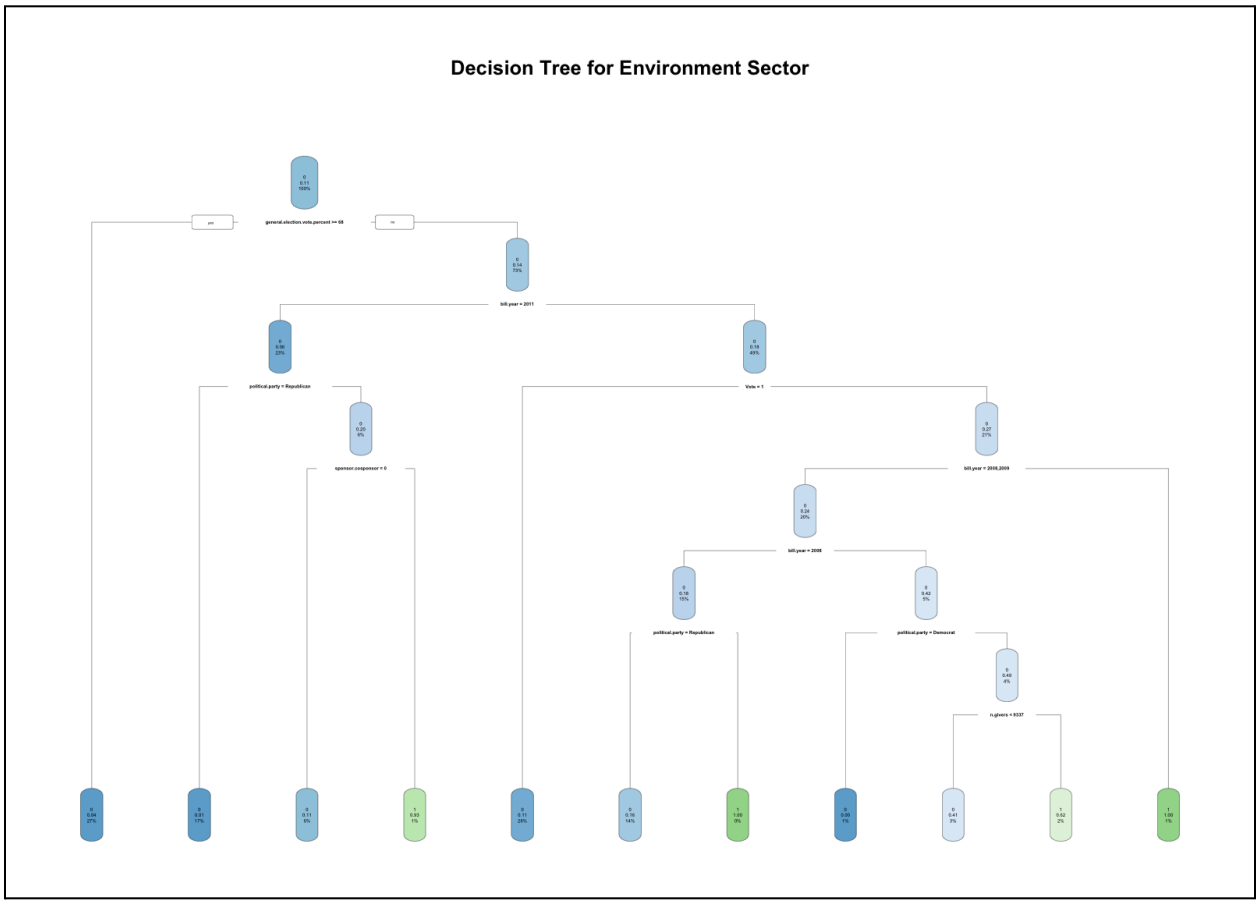
Transportation Sector:



Financial Sector:



Environment Sector:



✝:This does not have a specific stock sector, but relies on Energy and Utilities companies.

From the decision trees shown above, the different coloring and shading, between blue and green, is related to the predicted probability at that leaf of voting against the majority. So, the closer the predicted probability is to 0, not voting against the majority, the darker the blue shade on the leaf. And, the closer the predicted probability is to 1, voting against the majority, the darker the green shade on the leaf. However, as the predicted probability gets closer to 0.5, the more ambiguous the coloring of the leaf is.

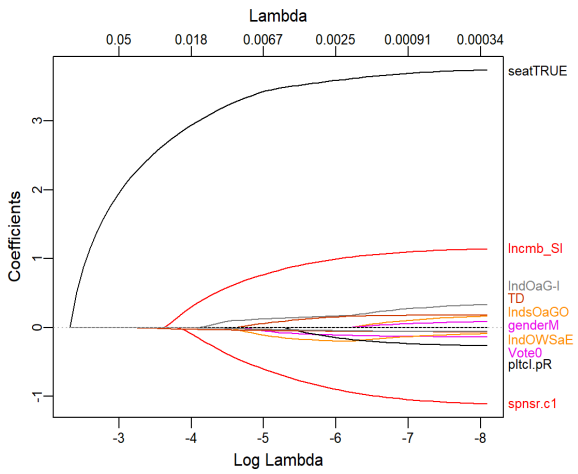
As we can see from above, the decision trees vary a good amount for each sector. We covered the same sector for the decision trees that we did for the logistic regression. For each decision tree, we predicted whether a congressman voted against the majority of their party on a bill with the explanatory variables of whether they voted yes or no on a bill or joint resolution, Industry, Incumbent Status, Total Donations, bill year, political party, gender of the congressman, total number of donors to the congressman, the general election vote percentage of the congressman, whether the congressman was a sponsor or cosponsor on the bill, the mean democrat republican ratio for the companies donating to the congressman, and whether a

congressman is in the Senate or the House where the seat variable being a 1 means they are in the Senate.

We are able to see the variables of seat, bill year, and general election vote percentage be more relevant variables over the many decision trees, as they consistently pop up as nodes on the tree and are nodes closer to the top of the tree, meaning their splits is of more importance than the splits below them.

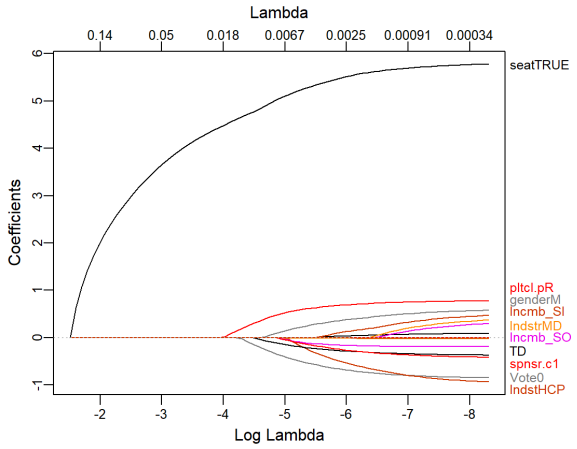
An interesting observation is that, through all the decision tree models, total donations was not determined to be a significant variable in predicting whether a congressman votes against the majority of their political party on a bill (response variable). This is in contrast to the results provided by logistic regression, which showed the log of total donations to be statistically significant. This is important because our main research question was understanding the influence total donations had on a congressman voting against the majority of their party on bills and joint resolutions. Our main variable, in log form, only shows significance through assuming a generalized linear relationship, but it shows no significance under the assumption of a non-linear assumption. The different responses underline the inherent complexity of this topic and makes it difficult to confirm a singular response to our question.

We returned to linear models to see if we could use a penalized regression model like Lasso to perform variable selection. These were the results after using the largest plotted lambda for the loss function:

Topic	Coefficients at Minimum Lambda on Graph (λ = 0.0034)	Lambda Plot for Coefficients
Energy	(Intercept) -1.773949e+00 Vote0 -1.347591e-01 Vote1 4.481501e-15 IndustryOil and Gas - Integrated 3.303453e-01 IndustryOil and Gas Operations 1.625913e-01 IndustryOil Well Services and Equipment -9.080936e-02 Incumbent_StatusI 1.138332e+00 Incumbent_StatusO . TD 1.834343e-01 sponsor.cosponsor1 -1.104253e+00 political.partyRepublican -2.636591e-01 seatTRUE 3.735380e+00 genderM 8.307782e-02 n.givers 2.955066e-06 general.election.vote.percent -6.165187e-02 mean.democrat.republican.ratio.for.companies -6.029854e-02	

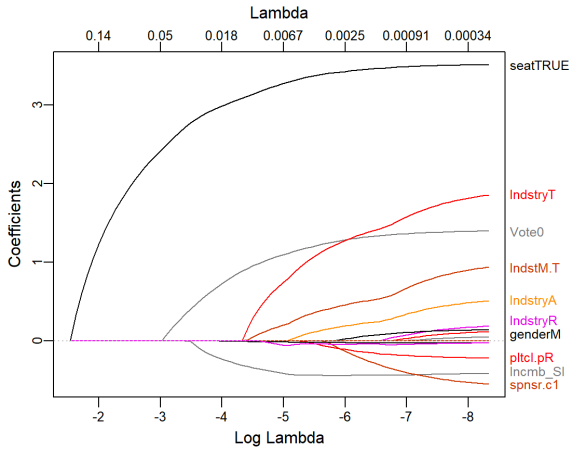
Healthcare

	Coefficient.Value
(Intercept)	2.883920e+00
Vote0	-8.440275e-01
Vote1	5.381493e-14
IndustryHealth Care Plans	-9.153524e-01
IndustryHealthcare Facilities	8.436248e-02
IndustryMajor Drugs	3.456559e-01
IndustryMedical Equipment and Supplies	-1.932914e-01
Incumbent_StatusI	4.473074e-01
Incumbent_StatusO	2.732218e-01
TD	-3.673869e-01
sponsor.cosponsor1	-4.103905e-01
political.partyRepublican	7.711420e-01
seatTRUE	5.768589e+00
genderM	5.637393e-01
n.givers	1.231251e-06
general.election.vote.percent	-2.614651e-02
mean.democrat.republican.ratio.for.companies	-1.556967e-02



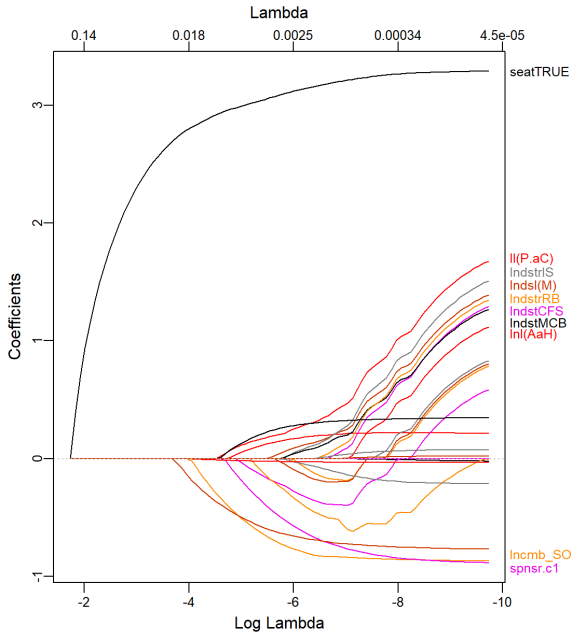
Transportation

	Coefficient.Value
(Intercept)	-3.336526e+00
Vote0	1.390407e+00
Vote1	-3.691252e-13
IndustryAirline	4.823935e-01
IndustryMisc. Transportation	8.981087e-01
IndustryRailroads	1.707796e-01
IndustryTrucking	1.813229e+00
Incumbent_StatusI	-4.194623e-01
Incumbent_StatusO	4.241682e-02
TD	1.039129e-01
sponsor.cosponsor1	-5.283692e-01
political.partyRepublican	-2.159597e-01
seatTRUE	3.507088e+00
genderM	1.350995e-01
n.givers	6.469822e-07
general.election.vote.percent	-2.987554e-02
mean.democrat.republican.ratio.for.companies	-3.068215e-02



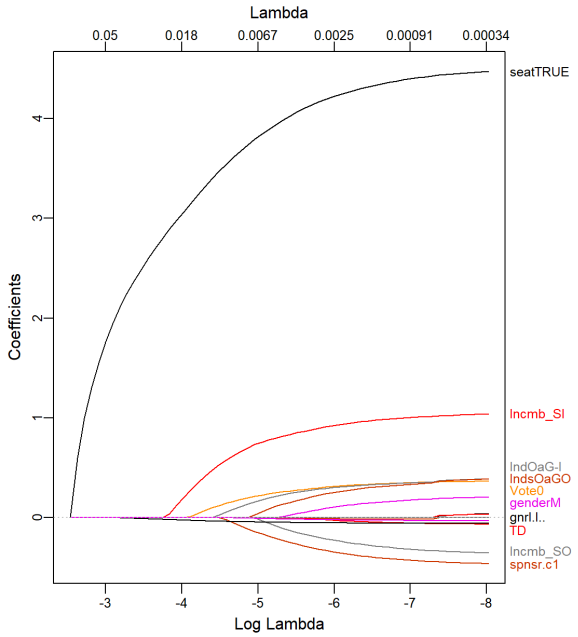
Financial

	Coefficient.Value
(Intercept)	1.462458e+00
Vote0	2.199449e-01
Vote1	-9.705885e-03
IndustryCollectibles	2.046478e-01
IndustryConsumer Financial Services	6.255497e-01
IndustryInsurance	1.308476e-01
IndustryInsurance (Accident and Health)	4.747541e-01
IndustryInsurance (Life)	-9.922005e-03
IndustryInsurance (Miscellaneous)	7.438596e-01
IndustryInsurance (Prop. and Casualty)	1.011523e+00
IndustryInvestment Services	8.342797e-01
IndustryMoney Center Banks	6.498089e-01
IndustryMortgage Investment	-4.624290e-01
IndustryRegional Banks	6.802225e-01
IndustrySandLs/Savings Banks	1.556785e-01
Incumbent_StatusI	6.552464e-02
Incumbent_StatusO	-8.563337e-01
TD	-1.918255e-01
sponsor.cosponsor1	-8.416419e-01
political.partyRepublican	-7.502407e-01
seatTRUE	3.264216e+00
genderM	3.370778e-01
n.givers	-1.046716e-06
general.election.vote.percent	-3.298457e-02
mean.democrat.republican.ratio.for.companies	1.630385e-02



Environment

	Coefficient.Value
(Intercept)	1.130651e+00
Vote0	3.629634e-01
Vote1	-1.128294e-15
IndustryElectric Utilities	.
IndustryNatural Gas Utilities	3.738224e-02
IndustryOil and Gas - Integrated	3.869146e-01
IndustryOil and Gas Operations	3.861479e-01
IndustryOil Well Services and Equipment	3.408259e-02
Incumbent_StatusI	1.036179e+00
Incumbent_StatusO	-3.543940e-01
TD	-6.568265e-02
sponsor.cosponsor1	-4.612372e-01
political.partyRepublican	.
seatTRUE	4.474282e+00
genderM	2.043816e-01
n.givers	1.943226e-06
general.election.vote.percent	-5.920891e-02
mean.democrat.republican.ratio.for.companies	-3.169563e-02



Looking at the results from the Lasso Regression, all the models included the logged version of Total Donations (TD), which means they considered it an important variable at the lambda closest to 0 in the graphs. Comparing the coefficients within each sector, TD is not one of the largest effects by absolute value of the coefficient. It tends to be behind seat = TRUE (this

means the vote is for the U.S. Senate) and Incumbent Status, along with a few other variables. Comparing the logged TD values to each other, it looks like the effects range between -0.367 for healthcare and 0.183 for energy, with the absolute values ranging from 0.0657 (environment) to 0.367 (healthcare). Given that these are penalized linear regression models, the variables are easy to interpret, since each one unit increase in a value *i* results in a beta *i* increase. This means that for each 1 unit increase in the logged total healthcare sector donations for a candidate, the probability of voting against the party decreases by 0.367, which is interesting. Our logistic regression model outputted a positive beta, meaning that an increase in the log of total donations for and from that sector is related to an increase in voting against the party, rather than a decrease.

Overall, this project highlighted the immense complexity determining the relationship between variables in the real world. There's no model that will ever be able to do it justice in terms of trying to grasp what is really taking place. In the case of this project, even though we were able to find significant connections between donations and the factor of congressmen voting against their parties on bills, the true cause and effect as to why these congressmen make this decision is much more complicated and can't be represented solely by a statistical model. With that in mind, the efforts of this project did produce results that showcase a significant relationship between political contributions and congressmen going against the majority of their political party. Again, it isn't as simple as pay this much to get this congressmen to vote in favor of what you want on a particular bill in Congress, but large companies and organizations that want political influence could make it a strategy to use their financial resources in hope of persuading those in power to side with them when it comes to legislation. Another idea to consider is that with there potentially being the option for these larger corporations, within the sectors discussed in this paper, to use their wealth to influence American politics, maybe more regulation should be put into place to limit individuals of Fortune 500 and other resource abundant companies on how much they can legally contribute to politicians in hope of controlling their influence on the political world as a whole.

Work Cited

Fowler, Anthony, et al. “Quid pro quo? Corporate returns to campaign contributions.” *The Journal of Politics,* vol. 82, no. 3, July 2020, pp. 844-858, https://doi.org/10.1086/707307

Appendix 1: Data Dictionary

Variable	Variable Name	Description	Example
Bill ID	bill.id	The unique identifier for each bill/joint resolution that's introduced in the U.S. House.	110_hr434
Bill Year	bill.year	The year the bill was introduced to the floor of Congress	2008
Recipient ID	boinca.rid	The unique identifier for a congressman that's won a general election.	cand1000
Seat	seat	Indicator variable for whether the politician is a member of the House of Representative or the Senate.	0 (0 = House of Representative, 1 = Senate)
Name	name	The name of the congressman.	ARCURI, MICHAEL ANGELO
Vote Choice	vote_choice	The vote choice of the congressmen.	Yes
Sponsor/Cosponsor	sponsor.cosponsor	An indicator variable that determines if a congressman was either a sponsor or cosponsor of a bill, or neither.	0 (0 = Neither a sponsor or cosponsor)
Election Cycle	cycle	The year of the general election and the two-year election cycle (starting two years before the stated "cycle" variable and up to it) during which political	2006

		contributions were recorded.	
Party	political.party	The political party of the congressman.	100 (100 = Democrat, 200 = Republican)
Swing State	swing_state	An indicator variable that identifies if a congressman represents a district in a state that's considered a swing state (+/- 2 congressional seat difference between both parties)	1 (0 = No, 1 = Yes)
Incumbency Status	Incumbent_Status	The incumbency status of the congressman.	O (O = Open Seat Candidate, I = Incumbent, C = Challenger)
Gender	cand.gender	The gender of the congressman.	M (M = Male, F = Female)
Number of Givers	n.givers	Number of distinct donors that gave to the congressman during a specific election cycle.	1172
Primary Election Percent	prim.vote.pct	FEC reported vote share in the primary election.	1
General Election Percent	general.election.vote.percent	FEC reported vote share in the general election.	.5395
Topic	top_topic	The topic most relevant to the bill that is being voted on.	tw.healthcare
Company Name	corp.name	The name of the corporation that contributed to the	Apple Inc

		congressman.	
Ticker	ticker	Stock symbol of the corporation.	AAPL
Sector	sector	Sector of the corporation.	Technology
Industry	industry	Industry of the corporation. A subset of a sector.	Computer Hardware
Average Company's Political Contribution Ratio	mean.democrat.republican.ratio.for.companies	Average donation the company has given to Democrats within our time period / Average donation the company has given to Republicans within our time period.	15.823
Contribution Amount	total_cont	A single transaction of the contribution amount from the corporation to the congressman.	4200
Vote Against Majority	vote.against.majority	An indicator variable that identifies if a congressman's vote was against the majority decision of their political party ("party")	0 (0 = No, 1 = Yes)

Appendix 2: Code

Loading and Filtering

```
library(dplyr)
library(tidyr)
library(stringr)


companies <- read.csv("dataverse_files/bod_fortune_500_DIME.csv")

companies <- select(
  companies,
  -c("corp.person.id", "middle.name", "age", "gender", "ceo", "privatefirm",
     "chairman", "dime.cfscore", "self.funded", "pct.to.dems", "to.incumbs",
     "to.open.seat", "to.challs", "to.winner", "to.losers"))

companies <- companies %>%
  mutate(
    total2 = ifelse(total == "NA", total.dem + total.rep, as.numeric(total))) %>%
  select(-total) %>%
  rename(total = total2) %>%
  filter(total > 0) %>%
  mutate(num.conts = ifelse(is.na(num.conts), 1, num.conts))


companies.ind <- read.csv("dataverse_files/bod_fortune_500_DIME_cont_records.csv")

companies.ind <- companies.ind %>%
  filter(cycle >= 2006, cycle <= 2014) %>%
  select(-c(latitude, longitude, cfscore, date, contributor.mname,
            contributor.suffix, contributor.title, contributor.occupation,
            contributor.employer, contributor.address, contributor.city,
            contributor.state, contributor.zipcode, recipient.name, recipient.party)) %>%
  filter(!str_starts(dime.rid, "comm")) %>%
  select(-c(contributor.lname, contributor.fname, corpname))


recipients <- read.csv("fixed_recipients.csv")

recipients <- recipients %>%
  select(-fecyear) %>%
  filter(
    cycle %in% c(2006, 2008, 2010, 2012, 2014),
```

```
    str_starts(seat, "federal"),
    gwinner == "W"
    )


votes <- read.csv("vote_db.csv/vote_db.csv")

bills <- read.csv("vote_db.csv/bills_db.csv")

bills$date <- as.Date(bills$date)

bills <- bills %>%
  filter(
    date >= as.Date("2005-01-03") & date <= as.Date("2013-01-02"),
    congno > 109,
    tw.latent1 <= 0.95
    ) %>%
  select(-c(date, bill.str, sponsors, cosponsors))

bills.long <- bills %>%
  pivot_longer(
    cols = -c(bill.id, year, bill.desc, congno), # Columns to keep fixed
    names_to = "Attribute", # Name of the new column for the attribute names
    values_to = "Value" # Name of the new column for the values
  )

bills <- bills.long %>%
  filter(
    !Attribute %in% c("tw.abortion.and.social.conservatism", "tw.civil.rights",
                      "tw.congress.and.procedural", "tw.crime", "tw.education",
                      "tw.fair.elections", "tw.higher.education", "tw.indian.affairs",
                      "tw.latent1", "tw.veterans.affairs", "tw.womens.issues",
                      "tw.law.courts.and.judges"),
    Value != 0
        )


write.csv(companies, "companies.csv")
write.csv(companies.ind, "companies_ind.csv")
write.csv(recipients, "recipients.csv")
write.csv(votes, "votes.csv")
write.csv(bills, "bills.csv")
```

Manipulations and Joins

```
library(tidyr)
library(dplyr)

####READ IN INITIAL FILES#########################
company_donations <- read.csv("companies_ind.csv")
company_info <- read.csv("companies.csv")
candidate_info <- read.csv("recipients.csv")
bills <- read.csv("bills.csv")
votes <- read.csv("votes.csv")

####ADJUST THE COLUMN NAMES#################################
names(company_donations) <- gsub("FINAL_COMPANY_DONATIONS_", "",
names(company_donations))
names(company_info) <- gsub("FINAL_COMPANY_INFO_", "", names(company_info))
names(candidate_info) <- gsub("FINAL_RECIPIENTS_", "", names(candidate_info))
names(bills) <- gsub("FINAL_BILLS_", "", names(bills))
names(votes) <- gsub("FINAL_VOTE_", "", names(votes))

names(company_donations) <- gsub("^ï..", "", names(company_donations))
names(company_info) <- gsub("^ï..", "", names(company_info))
names(candidate_info) <- gsub("^ï..", "", names(candidate_info))
names(bills) <- gsub("^ï..", "", names(bills))
names(votes) <- gsub("^ï..", "", names(votes))

#####INITIAL MANIPULATIONS#################################

#Company Party Alignment. Change variable to name
company_info$Company_Party <- ifelse(company_info$Total_Democrat >
company_info$Total_Republican, "Democrat", "Republican")

#Fix the column types for donations to be numeric
company_info <- company_info %>% mutate(
  Total_Donation = as.numeric(gsub(",", "", gsub("\\$", "", Total_Donation))),
  Donation_Standard_Deviation = as.numeric(gsub(",", "", gsub("\\$", "",
Donation_Standard_Deviation))),
  Mean_Donor_Amount = as.numeric(gsub(",", "", gsub("\\$", "", Mean_Donor_Amount))),
  )


#Candidate Party Alignment
candidate_info$Parties <- ifelse(candidate_info$Party == 100, 1, ifelse(candidate_info$Party == 200,
0, 2))
candidate_info$Party <- ifelse(candidate_info$Party == 100, "Democrat", ifelse(candidate_info$Party
== 200, "Republican", "Other"))

#Vote is Yay (originally 6, now 1) or Nay (0)
votes$Vote <- ifelse(votes$Vote == 6, 0, 1)
```

```
#Only include hr/hjres
bills <- bills %>% filter(
  grepl("(hr|hjres)\\d+",bill_id)
)

#######STARTING JOINS#########################################

#Bills, Votes
bills.votes <- left_join(x=bills,y=votes,by = "bill_id")

#Bills and Votes, and Recipients
bills.votes.recipients <- left_join(x = bills.votes, y = candidate_info, by = c("cycle","bonica_rid"))


#Breakdown the majority vote by party to get the party dissenters variable

#Groupby bill_id, Party
#Sum Votes to get total yesses
#Count rows to get total votes
#Find the majority vote by party

temp.3.18.24.grouping <- bills.votes.recipients %>%
  filter(Party == "Democrat" | Party == "Republican") %>%
  group_by(bill_id,Party,Attribute) %>%
  summarize(sum(Vote),n())

#Whether the majority is yes or no
temp.3.18.24.grouping$Maj.Vote <- ifelse(
  temp.3.18.24.grouping$`sum(Vote)`/temp.3.18.24.grouping$`n()` > 0.5,
  1,
  0
)

temp.3.18.24.grouping.lookup.table <- temp.3.18.24.grouping %>%
  pivot_wider(
    names_from = Party,
    values_from = c(`sum(Vote)`,`n()`,Maj.Vote)
  ) %>% select(
    -Attribute
  ) %>%
  distinct(bill_id,.keep_all = TRUE)

# Get majority vote by party
lookup.majority.votes <- temp.3.18.24.grouping.lookup.table %>% select(
  bill_id,Maj.Vote_Democrat,Maj.Vote_Republican)



#Bills, Party Majority Votes
```

```
bills.majority.votes <- left_join(x=bills,y=temp.3.18.24.grouping.lookup.table,by="bill_id")
#Bills and Party Majority Votes, Votes
bills.votes <- left_join(x=bills.majority.votes,y=votes,by = "bill_id")

#Bills and All Votes, and Recipients
bills.votes.recipients <- left_join(
  x = bills.votes, y = candidate_info, by = c("cycle","bonica_rid"))

#Vote_Differently
#temp.bills.votes.recipients <- left_join(
#  bills.votes.recipients,lookup.majority.votes,by='bill_id')

#Calculate the votes against their party's majority vote
#bills.votes.recipients <- bills.votes.recipients %>% mutate(
#  Vote.Against.Majority = case_when(Party == 'Democrat' ~ ifelse(Vote == Maj.Vote_Democrat,0,1),
#            Party == 'Republican' ~ ifelse(Vote == Maj.Vote_Republican,0,1))
#)

bills.votes.recipients <- bills.votes.recipients %>% mutate(
    Vote.Against.Majority = case_when(
      Party == 'Democrat' ~ ifelse(Vote == Maj.Vote_Democrat,0,1),
      Party == 'Republican' ~ ifelse(Vote == Maj.Vote_Republican,0,1)
      )
  )

all.company.info <- left_join(x = company_donations, y = company_info, by = "ticker")
all.company.info$Democrat_Republican_Ratio <- all.company.info$Total_Democrat /
all.company.info$Total_Republican
names(all.company.info) <- gsub("dime_rid", "bonica_rid", names(all.company.info))

## VERY IMPORTANT
## SINCE WE JOIN ON THE CYCLE COLUMN, THIS MEANS THAT OUR DONATION
COLUMNS ACTUALLY ONLY MEAN
## DONATIONS FROM THE PREVIOUS ELECTION FOR THAT CONGRESSIONAL SESSION
## FOR EXAMPLE:
## IF A CYCLE SAYS 2006, THEN IT WILL MATCH TO THE BILLS THAT HAPPEN
IMMEDIATELY AFTER THE 2006
## ELECTION CYCLE, WHICH MEANS THE SESSION OF CONGRESS THAT WENT FROM
JANUARY 2007 TO DECEMBER 2008
## IN THIS EXAMPLE, TOTAL DONATIONS MEANS DONATIONS FROM ONLY THE 2006
ELECTION CYCLE (NO PREVIOUS ONES)
df <- left_join(x = bills.votes.recipients, y = all.company.info, by = c("cycle","bonica_rid"))
df <- na.omit(df)

dim(df)
# [1] 7231383      46
```

Review Topic Alignment

```
#Look at the individual sectors and check which ones align with the bill topics
df$Sector %>% unique()

#The sectors that line up to a topic weight
#Group Energy and Utilities
#Try Technology and Services Grouped, then without
c("Technology", "Services",
  "Financial",
  "Healthcare","Energy","Utilities","Transportation")


df %>% select(Top_Topic) %>% unique()

# The bill topics that align with a sector
#Group economy and banking/finance
#Group energy and environment
c(
  'tw.labor','tw.economy','tw.banking.and.finance',
  'tw.healthcare','tw.energy','tw.environment','tw.transportation')
```

Linear Modeling Function

```
library(ggplot2)

get.model <- function(topic,sector,log.donations = FALSE,top.weight=0.4,include.industries=FALSE){

# topic = bill topic
# sector = company sector
# log.donations = apply log() transformation on donations
# top.weight = minimum threshold for bill top topic weight (default at 0.4 for bills highly related to
#          topic, but some sectors have no bills at weight = 0.4, so there's a need to lower this
# include.industries = look at interactions with industries instead of only sector in model

  if(length(topic) == 1){
    filtered.df <- df %>%
      filter(
        Top_Topic == topic,
        Top_Topic_Weight >=top.weight,Sector==sector,
        Total_Donation > 0) %>% group_by(
          bill_id,bonica_rid,Vote,Vote.Against.Majority,Industry,Incumbent_Status) %>% summarize(
            TD = sum(Total_Donation))
  } else {
    filtered.df <- df %>%
      filter(
        Top_Topic %in% topic,
        Top_Topic_Weight >=top.weight,Sector==sector,
        Total_Donation > 0) %>% group_by(
          bill_id,bonica_rid,Vote,Vote.Against.Majority,Industry,Incumbent_Status) %>% summarize(
            TD = sum(Total_Donation))
  }

# Three summary statistics to understand input data
  cat("Number of Bills: ", filtered.df$bill_id %>% unique() %>% length(),"\n")
  cat("Median Donation Amount: ",filtered.df$TD %>% median(),"\n")
  cat("Donations at 0th,50th,80th percentiles: ",quantile(filtered.df$TD,c(0,0.5,0.8)))

  if(include.industries == FALSE){
    if(log.donations == FALSE){
      glm_result <- glm(Vote.Against.Majority ~ TD,
                        family = binomial, data = filtered.df)
      print(ggplot(data=filtered.df,aes(x=TD,y=factor(Vote.Against.Majority))) + geom_boxplot())
    }else{
      glm_result <- glm(Vote.Against.Majority ~ log(TD),
                        family = binomial, data = filtered.df)
      print(ggplot(data=filtered.df,aes(x=log(TD),y=factor(Vote.Against.Majority))) + geom_boxplot())
    }
  } else {
    if(log.donations == FALSE){
      glm_result <- glm(Vote.Against.Majority ~ TD*Industry,
                        family = binomial, data = filtered.df)
      print(ggplot(data=filtered.df,aes(x=TD,y=factor(Vote.Against.Majority))) + geom_boxplot())
```

```
    } else{
      glm_result <- glm(Vote.Against.Majority ~ log(TD)*Industry,
                        family = binomial, data = filtered.df)
      print(ggplot(data=filtered.df,aes(x=log(TD),y=factor(Vote.Against.Majority))) + geom_boxplot())

    }
  }
# Return summary of logistic regression model
  return(summary(glm_result))
}
```

Linear Model Breakdown by Topic (Equations represent example probability predictions for either 0th,50th, or 80th percentile for the best model for that topic)

```
#Energy
get.model("tw.energy","Energy",top.weight=0.25,log.donations = TRUE) #Better model

exp(-4.3769 + log(20351410)*0.1484) / (1+exp(-4.3769 + log(20351410)*0.1484))

#Healthcare
get.model("tw.healthcare","Healthcare",TRUE,top.weight = 0.25) #Better model

exp(-18.74362 + log(14048181)*1.04499) / (1+exp(-18.74362 + log(14048181)*1.04499))

#Transportation
get.model("tw.transportation","Transportation",TRUE,top.weight=0.25)

exp(-13.1939 + log(24849638)*0.6975) / (1+exp(-13.1939 + log(24849638)*0.6975))

#Financial
get.model('tw.banking.and.finance',"Financial",TRUE,top.weight=0.25) #Best model

exp(-8.54476 + log(12177782)*0.42195) / (1+exp(-8.54476 + log(12177782)*0.42195))

#Environment
get.model(c('tw.environment',"tw.energy"),"Energy",TRUE,top.weight=0.25)

exp(-4.59741 + log(16974213)*0.16236) / (1+exp(-4.59741 + log(16974213)*0.16236))
```

Lasso Regression and Decision Tree Models

```
#Code for the decision trees

df <- df %>%
  select(bill_id, year, cycle, Attribute, Value,Top_Topic,
         Top_Topic_Weight, `sum(Vote)_Democrat`, `sum(Vote)_Republican`,
         `n()_Democrat`, `n()_Republican`, Maj.Vote_Democrat, Maj.Vote_Republican,
         bonica_rid, Sponsor, Cosponsor, Vote, Party, State, Seat,
         Incumbent_Status, Gender, N_Givers, Primary_Winner,
         General_Election_Vote_Percent, District_Pres_VS, Parties,
         Vote.Against.Majority, ticker, Total_Donations, Sector, Industry,
         Total_Democrat, Total_Republican, Number_of_Contributions,
         Total_Donation, Democrat_Republican_Ratio)

#[1] 7,231,383     37
dim(df)

df <- df %>% filter(Attribute == Top_Topic)

#[1] 1,400,898     37
dim(df)

#####Data Preparation#####################################################

df %>% head()


#Energy Decision tree (Change the topic and the sector to create each tree and lasso regression)
topic <- c("tw.energy")
top.weight <- 0.25
sector <- c("Energy")

filtered.df <- df %>%
  filter(
    Top_Topic %in% topic,
    Top_Topic_Weight >=top.weight,Sector %in% sector,
    Total_Donation > 0) %>% group_by(
      bill_id,bonica_rid,Vote,Vote.Against.Majority,Industry,Incumbent_Status) %>% summarize(
        TD = log(sum(Total_Donation)),
        bill.year = max(year),
        sponsor.cosponsor = as.integer(any(Sponsor == 1 | Cosponsor == 1)),
        political.party = max(Party),
        seat = max(Seat) == "federal:senate",
        gender = max(Gender),
        n.givers = max(N_Givers),
        general.election.vote.percent = max(General_Election_Vote_Percent),
        mean.democrat.republican.ratio.for.companies = mean(Democrat_Republican_Ratio)) %>%
ungroup()
```

```
filtered.df <- filtered.df %>% mutate(
  Vote = as.factor(Vote),
  Vote.Against.Majority = as.factor(Vote.Against.Majority),
  Industry = as.factor(Industry),
  Incumbent_Status = as.factor(Incumbent_Status),
  bill.year = as.factor(bill.year),
  political.party = as.factor(political.party),
  gender = as.factor(gender),
  sponsor.cosponsor = as.factor(sponsor.cosponsor)
)

# Create the model matrix excluding the response variable and other non-predictors
x <- model.matrix(~ . -1 - Vote.Against.Majority - bill_id - bonica_rid - bill.year,data = filtered.df)  # -1
to omit intercept

# Ensure the response variable is also appropriately factored and numeric
y <- as.numeric(as.factor(filtered.df$Vote.Against.Majority)) - 1  # subtract 1 to make it 0-based for
glmnet

# Checking for NAs and handle them if necessary
complete_cases <- complete.cases(x, y)
x <- x[complete_cases, ]
y <- y[complete_cases]
###Modeling#################################################

# Lasso regression

lasso_model <- glmnet(x, y, family = "binomial", alpha = 1)


dev.new()
plot(lasso_model, xvar = "lambda", label = TRUE)

# Memory Usage Reached! Following code won't work
cv_lasso <- cv.glmnet(x, y, family = "binomial", alpha = 1)
plot(cv_lasso)
best_lambda_lasso <- cv_lasso$lambda.min

# Decision Tree

filtered.df$Vote.Against.Majority<-factor(filtered.df$Vote.Against.Majority)

sample.data<-sample.int(nrow(filtered.df), floor(0.8*nrow(filtered.df)),
                        replace = F)
train<-filtered.df[sample.data, ]
test<-filtered.df[-sample.data, ]

y.test<-test[,"Vote.Against.Majority"]

train<-filtered.df[sample.data, ] %>% select(-bonica_rid,-bill_id)
```

```
test<-filtered.df[-sample.data,] %>% select(-bonica_rid,-bill_id)

library(rpart)
library(rpart.plot)

tree <- rpart(Vote.Against.Majority ~ ., data = train, method = "class")
rpart.plot(tree, main = "Decision Tree for Energy Sector")

topic <- c("tw.healthcare")
top.weight <- 0.25
sector <- c("Healthcare")

filtered.df <- df %>%
  filter(
    Top_Topic %in% topic,
    Top_Topic_Weight >=top.weight,Sector %in% sector,
    Total_Donation > 0) %>% group_by(
      bill_id,bonica_rid,Vote,Vote.Against.Majority,Industry,Incumbent_Status) %>% summarize(
        TD = log(sum(Total_Donation)),
        bill.year = max(year),
        sponsor.cosponsor = as.integer(any(Sponsor == 1 | Cosponsor == 1)),
        political.party = max(Party),
        seat = max(Seat) == "federal:senate",
        gender = max(Gender),
        n.givers = max(N_Givers),
        general.election.vote.percent = max(General_Election_Vote_Percent),
        mean.democrat.republican.ratio.for.companies = mean(Democrat_Republican_Ratio)) %>%
ungroup()

filtered.df <- filtered.df %>% mutate(
  Vote = as.factor(Vote),
  Vote.Against.Majority = as.factor(Vote.Against.Majority),
  Industry = as.factor(Industry),
  Incumbent_Status = as.factor(Incumbent_Status),
  bill.year = as.factor(bill.year),
  political.party = as.factor(political.party),
  gender = as.factor(gender),
  sponsor.cosponsor = as.factor(sponsor.cosponsor)
)

# Create the model matrix excluding the response variable and other non-predictors
x <- model.matrix(~ . -1 - Vote.Against.Majority - bill_id - bonica_rid - bill.year,data = filtered.df)  # -1
to omit intercept

# Ensure the response variable is also appropriately factored and numeric
y <- as.numeric(as.factor(filtered.df$Vote.Against.Majority)) - 1  # subtract 1 to make it 0-based for
glmnet

# Checking for NAs and handle them if necessary
complete_cases <- complete.cases(x, y)
```

```
x <- x[complete_cases, ]
y <- y[complete_cases]
###Modeling#################################################

# Lasso regression

lasso_model <- glmnet(x, y, family = "binomial", alpha = 1)


dev.new()
plot(lasso_model, xvar = "lambda", label = TRUE)

# Memory Usage Reached! Following code won't work
cv_lasso <- cv.glmnet(x, y, family = "binomial", alpha = 1)
plot(cv_lasso)
best_lambda_lasso <- cv_lasso$lambda.min

# Decision Tree

filtered.df$Vote.Against.Majority<-factor(filtered.df$Vote.Against.Majority)

sample.data<-sample.int(nrow(filtered.df), floor(0.8*nrow(filtered.df)),
                        replace = F)
train<-filtered.df[sample.data, ]
test<-filtered.df[-sample.data, ]

y.test<-test[,"Vote.Against.Majority"]

train<-filtered.df[sample.data, ] %>% select(-bonica_rid,-bill_id)
test<-filtered.df[-sample.data,] %>% select(-bonica_rid,-bill_id)

tree <- rpart(Vote.Against.Majority ~ ., data = train, method = "class")
rpart.plot(tree, main = "Decision Tree for Healthcare Sector")


###
###
### Decision Tree for Transportation Sector


topic <- c("tw.transportation")
top.weight <- 0.25
sector <- c("Transportation")

filtered.df <- df %>%
  filter(
    Top_Topic %in% topic,
    Top_Topic_Weight >=top.weight,Sector %in% sector,
    Total_Donation > 0) %>% group_by(
      bill_id,bonica_rid,Vote,Vote.Against.Majority,Industry,Incumbent_Status) %>% summarize(
```

```
      TD = log(sum(Total_Donation)),
      bill.year = max(year),
      sponsor.cosponsor = as.integer(any(Sponsor == 1 | Cosponsor == 1)),
      political.party = max(Party),
      seat = max(Seat) == "federal:senate",
      gender = max(Gender),
      n.givers = max(N_Givers),
      general.election.vote.percent = max(General_Election_Vote_Percent),
      mean.democrat.republican.ratio.for.companies = mean(Democrat_Republican_Ratio)) %>%
ungroup()

filtered.df <- filtered.df %>% mutate(
  Vote = as.factor(Vote),
  Vote.Against.Majority = as.factor(Vote.Against.Majority),
  Industry = as.factor(Industry),
  Incumbent_Status = as.factor(Incumbent_Status),
  bill.year = as.factor(bill.year),
  political.party = as.factor(political.party),
  gender = as.factor(gender),
  sponsor.cosponsor = as.factor(sponsor.cosponsor)
)

# Create the model matrix excluding the response variable and other non-predictors
x <- model.matrix(~ . -1 - Vote.Against.Majority - bill_id - bonica_rid - bill.year,data = filtered.df)  # -1
to omit intercept

# Ensure the response variable is also appropriately factored and numeric
y <- as.numeric(as.factor(filtered.df$Vote.Against.Majority)) - 1  # subtract 1 to make it 0-based for
glmnet

# Checking for NAs and handle them if necessary
complete_cases <- complete.cases(x, y)
x <- x[complete_cases, ]
y <- y[complete_cases]
###Modeling#################################################

# Lasso regression

lasso_model <- glmnet(x, y, family = "binomial", alpha = 1)


dev.new()
plot(lasso_model, xvar = "lambda", label = TRUE)

# Memory Usage Reached! Following code won't work
cv_lasso <- cv.glmnet(x, y, family = "binomial", alpha = 1)
plot(cv_lasso)
best_lambda_lasso <- cv_lasso$lambda.min

# Decision Tree
```

```
filtered.df$Vote.Against.Majority<-factor(filtered.df$Vote.Against.Majority)

sample.data<-sample.int(nrow(filtered.df), floor(0.8*nrow(filtered.df)),
                        replace = F)
train<-filtered.df[sample.data, ]
test<-filtered.df[-sample.data, ]

y.test<-test[,"Vote.Against.Majority"]

train<-filtered.df[sample.data, ] %>% select(-bonica_rid,-bill_id)
test<-filtered.df[-sample.data,] %>% select(-bonica_rid,-bill_id)

tree <- rpart(Vote.Against.Majority ~ ., data = train, method = "class")
rpart.plot(tree, main = "Decision Tree for Transportation Sector")


### Decision Tree for Financial Sector


topic <- c("tw.banking.and.finance")
top.weight <- 0.25
sector <- c("Financial")

filtered.df <- df %>%
  filter(
    Top_Topic %in% topic,
    Top_Topic_Weight >=top.weight,Sector %in% sector,
    Total_Donation > 0) %>% group_by(
      bill_id,bonica_rid,Vote,Vote.Against.Majority,Industry,Incumbent_Status) %>% summarize(
        TD = log(sum(Total_Donation)),
        bill.year = max(year),
        sponsor.cosponsor = as.integer(any(Sponsor == 1 | Cosponsor == 1)),
        political.party = max(Party),
        seat = max(Seat) == "federal:senate",
        gender = max(Gender),
        n.givers = max(N_Givers),
        general.election.vote.percent = max(General_Election_Vote_Percent),
        mean.democrat.republican.ratio.for.companies = mean(Democrat_Republican_Ratio)) %>%
ungroup()

filtered.df <- filtered.df %>% mutate(
  Vote = as.factor(Vote),
  Vote.Against.Majority = as.factor(Vote.Against.Majority),
  Industry = as.factor(Industry),
  Incumbent_Status = as.factor(Incumbent_Status),
  bill.year = as.factor(bill.year),
  political.party = as.factor(political.party),
  gender = as.factor(gender),
  sponsor.cosponsor = as.factor(sponsor.cosponsor)
```

```
)

# Create the model matrix excluding the response variable and other non-predictors
x <- model.matrix(~ . -1 - Vote.Against.Majority - bill_id - bonica_rid - bill.year,data = filtered.df)  # -1 to omit intercept

# Ensure the response variable is also appropriately factored and numeric
y <- as.numeric(as.factor(filtered.df$Vote.Against.Majority)) - 1  # subtract 1 to make it 0-based for glmnet

# Checking for NAs and handle them if necessary
complete_cases <- complete.cases(x, y)
x <- x[complete_cases, ]
y <- y[complete_cases]
###Modeling#################################################

# Lasso regression

lasso_model <- glmnet(x, y, family = "binomial", alpha = 1)


dev.new()
plot(lasso_model, xvar = "lambda", label = TRUE)

# Memory Usage Reached! Following code won't work
cv_lasso <- cv.glmnet(x, y, family = "binomial", alpha = 1)
plot(cv_lasso)
best_lambda_lasso <- cv_lasso$lambda.min

# Decision Tree

filtered.df$Vote.Against.Majority<-factor(filtered.df$Vote.Against.Majority)

sample.data<-sample.int(nrow(filtered.df), floor(0.8*nrow(filtered.df)),
                        replace = F)
train<-filtered.df[sample.data, ]
test<-filtered.df[-sample.data, ]

y.test<-test[,"Vote.Against.Majority"]

train<-filtered.df[sample.data, ] %>% select(-bonica_rid,-bill_id)
test<-filtered.df[-sample.data,] %>% select(-bonica_rid,-bill_id)

tree <- rpart(Vote.Against.Majority ~ ., data = train, method = "class")
rpart.plot(tree, main = "Decision Tree for Financial Sector")


### Decision Tree for Environment Sector
```

```
topic <- c("tw.environment")
top.weight <- 0.25
sector <- c("Energy","Utilities")

filtered.df <- df %>%
  filter(
    Top_Topic %in% topic,
    Top_Topic_Weight >=top.weight,Sector %in% sector,
    Total_Donation > 0) %>% group_by(
      bill_id,bonica_rid,Vote,Vote.Against.Majority,Industry,Incumbent_Status) %>% summarize(
        TD = log(sum(Total_Donation)),
        bill.year = max(year),
        sponsor.cosponsor = as.integer(any(Sponsor == 1 | Cosponsor == 1)),
        political.party = max(Party),
        seat = max(Seat) == "federal:senate",
        gender = max(Gender),
        n.givers = max(N_Givers),
        general.election.vote.percent = max(General_Election_Vote_Percent),
        mean.democrat.republican.ratio.for.companies = mean(Democrat_Republican_Ratio)) %>%
ungroup()

filtered.df <- filtered.df %>% mutate(
  Vote = as.factor(Vote),
  Vote.Against.Majority = as.factor(Vote.Against.Majority),
  Industry = as.factor(Industry),
  Incumbent_Status = as.factor(Incumbent_Status),
  bill.year = as.factor(bill.year),
  political.party = as.factor(political.party),
  gender = as.factor(gender),
  sponsor.cosponsor = as.factor(sponsor.cosponsor)
)

# Create the model matrix excluding the response variable and other non-predictors
x <- model.matrix(~ . -1 - Vote.Against.Majority - bill_id - bonica_rid - bill.year,data = filtered.df)  # -1
to omit intercept

# Ensure the response variable is also appropriately factored and numeric
y <- as.numeric(as.factor(filtered.df$Vote.Against.Majority)) - 1  # subtract 1 to make it 0-based for
glmnet

# Checking for NAs and handle them if necessary
complete_cases <- complete.cases(x, y)
x <- x[complete_cases, ]
y <- y[complete_cases]
###Modeling#############################################

# Lasso regression

lasso_model <- glmnet(x, y, family = "binomial", alpha = 1)
```

```
dev.new()
plot(lasso_model, xvar = "lambda", label = TRUE)

# Memory Usage Reached! Following code won't work
cv_lasso <- cv.glmnet(x, y, family = "binomial", alpha = 1)
plot(cv_lasso)
best_lambda_lasso <- cv_lasso$lambda.min

# Decision Tree

filtered.df$Vote.Against.Majority<-factor(filtered.df$Vote.Against.Majority)

sample.data<-sample.int(nrow(filtered.df), floor(0.8*nrow(filtered.df)),
                        replace = F)
train<-filtered.df[sample.data, ]
test<-filtered.df[-sample.data, ]

y.test<-test[,"Vote.Against.Majority"]

train<-filtered.df[sample.data, ] %>% select(-bonica_rid,-bill_id)
test<-filtered.df[-sample.data,] %>% select(-bonica_rid,-bill_id)

tree <- rpart(Vote.Against.Majority ~ ., data = train, method = "class")
rpart.plot(tree, main = "Decision Tree for Environment Sector")
```