Congressional Survival

Neil Horning
Women in Data
Nov. 20 2019

Contents

Problem:

Based on voting history, what type of congressional member is most likely to survive? One that meets partisan preferences of a district or the expectations of their own partisan base?

Client / Stakeholder:

AUDREY DENNEY

D, CA-1

Audrey Denney is a Democratic congresswoman from California District 1, she was recently elected after a very close contest with Doug LaMalfa in 2020, riding the historic blue wave that swept Trump out of office. CA-1 is a rural, traditionally Republican district with a partisan lean of R+22.53. She wants to know what type of voting strategy will help her survive next election.

Exploratory Data Analysis

	bill	date	democratic	description	document_number
(116, 'Senate', 1, 1)	{'bill_id': 's1-116', 'number': 'S.1', 'sponsor_id': 'R000595', 'api_uri': 'https://api.propublica.org/congress/v1/116/bills/s1.json', 'title': 'A bill to make improvements to certain defense and security assistance provisions and to authorize the appropriation of funds to Israel, to reauthorize the United States-Jordan Defense Cooperation Act of 2015, and to halt the wholesale slaughter of the Syrian people, and for other purposes.', 'latest_action': 'Held at the desk.'}	2019-01-08	{'yes': 4, 'no': 41, 'present': 0, 'not_voting': 0, 'majority_position': 'No'}	A bill to make improvements to certain defense and security assistance provisions and to authorize the appropriation of funds to Israel, to reauthorize the United States-Jordan Defense Cooperation Act of 2015, and to halt the wholesale slaughter of the Syrian people, and for other purposes.	1

Meta Data

			party	D			ID	D		
			state	MA	CA	NJ	VT	WI	MA	HI
			dw_nominate	-0.774	-0.710	-0.611	-0.526	-0.512	-0.506	-0.498
			member_id	W000817	H001075	B001288	S000033	B001230	M000133	H001042
			name	Elizabeth Warren	Kamala Harris	Cory Booker	Bernard Sanders	Tammy Baldwin	Edward J. Markey	Mazie K. Hirono
congress	chamber	session	roll_call							
116	Senate	1	1	No	No	No	No	No	No	No
			2	No	No	No	No	No	No	No
			3	No	No	No	No	No	No	No
			4	No	No	No	No	No	No	No
			5	Yes	Yes	Yes	Yes	Yes	Yes	Yes

5 rows × 100 columns

Member Votes

Data Sets:

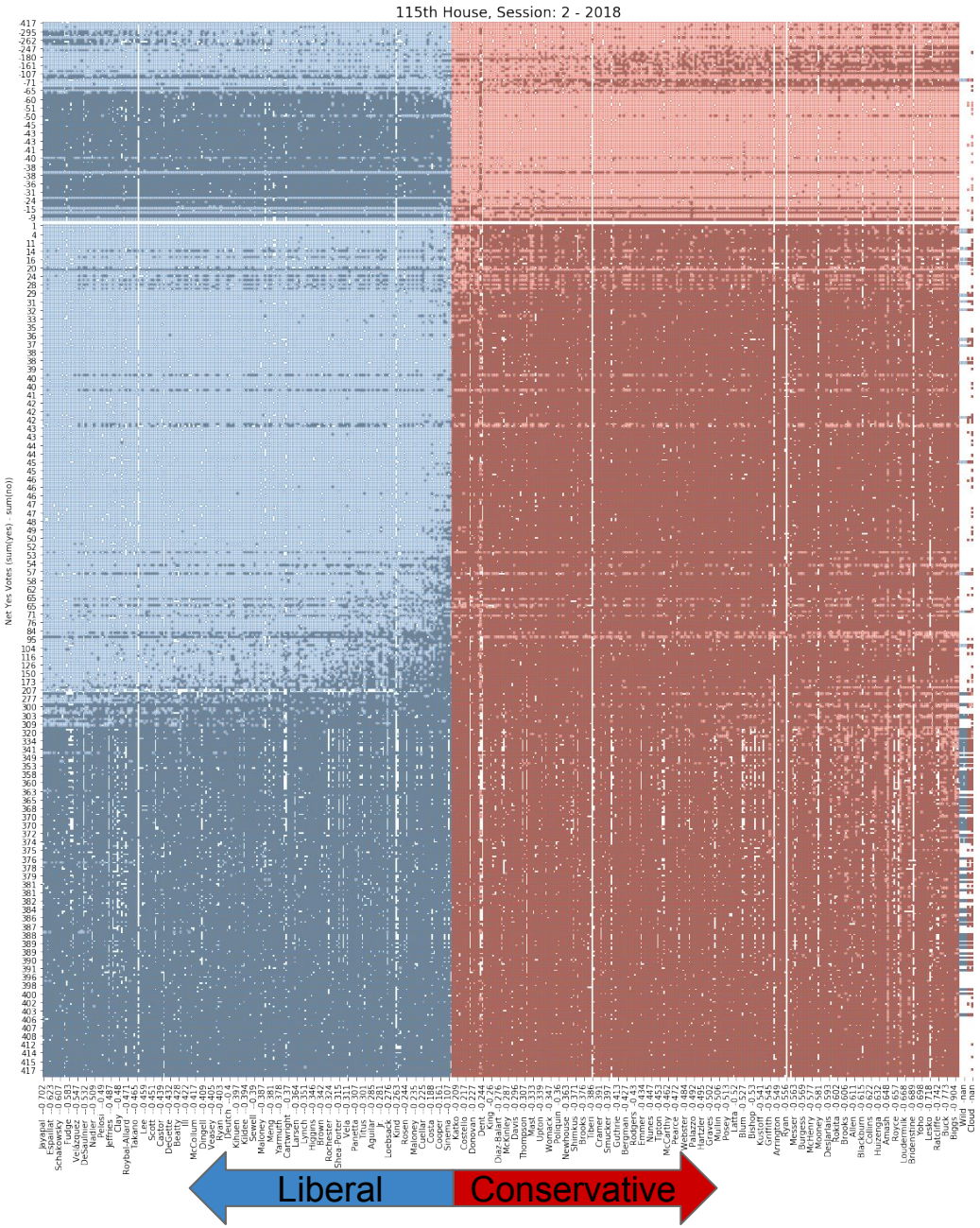
- congressional voting records via the [ProPublica Congress API](https://projects.propublica.org/api-docs/congress-api/).
- Partisan lean of districts and states available from FiveThirtyEight:

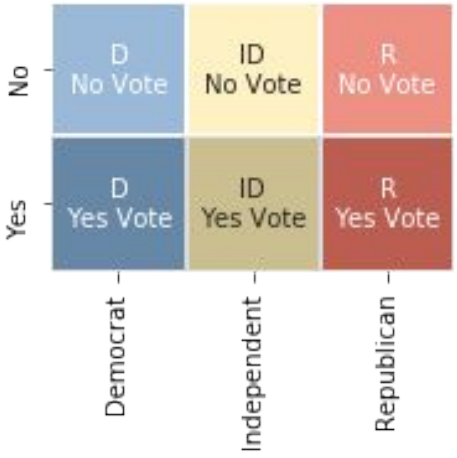
Cleaning/Wrangling:

Created Python functions to:

1. Request the ProPublica API for voting metadata for all months for a given congress chamber and year, and return a tidy Pandas DataFrame.
2. Take the index of the metadata for a given year and use it to request each member's position for each vote in that year.
3. Handle missing vote position data by filling rows with 'NaN' values
4. Take a given year (or congress session) and chamber, return corresponding DataFrame of all the metadata and vote positions, cache this data into csv format for automatic quick loading of additional requests.
5. Detect and update out of date csv cache, by fetching only the missing data.
6. Convert district lean from FiveThirtyEight from positive (R+x|D+x) values into continuous -1 to +1 scale compatible with ideological score.

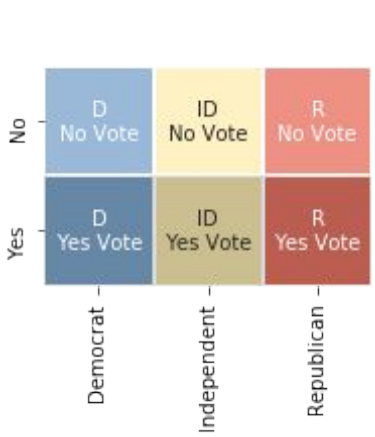
What patterns are apparent in the data?

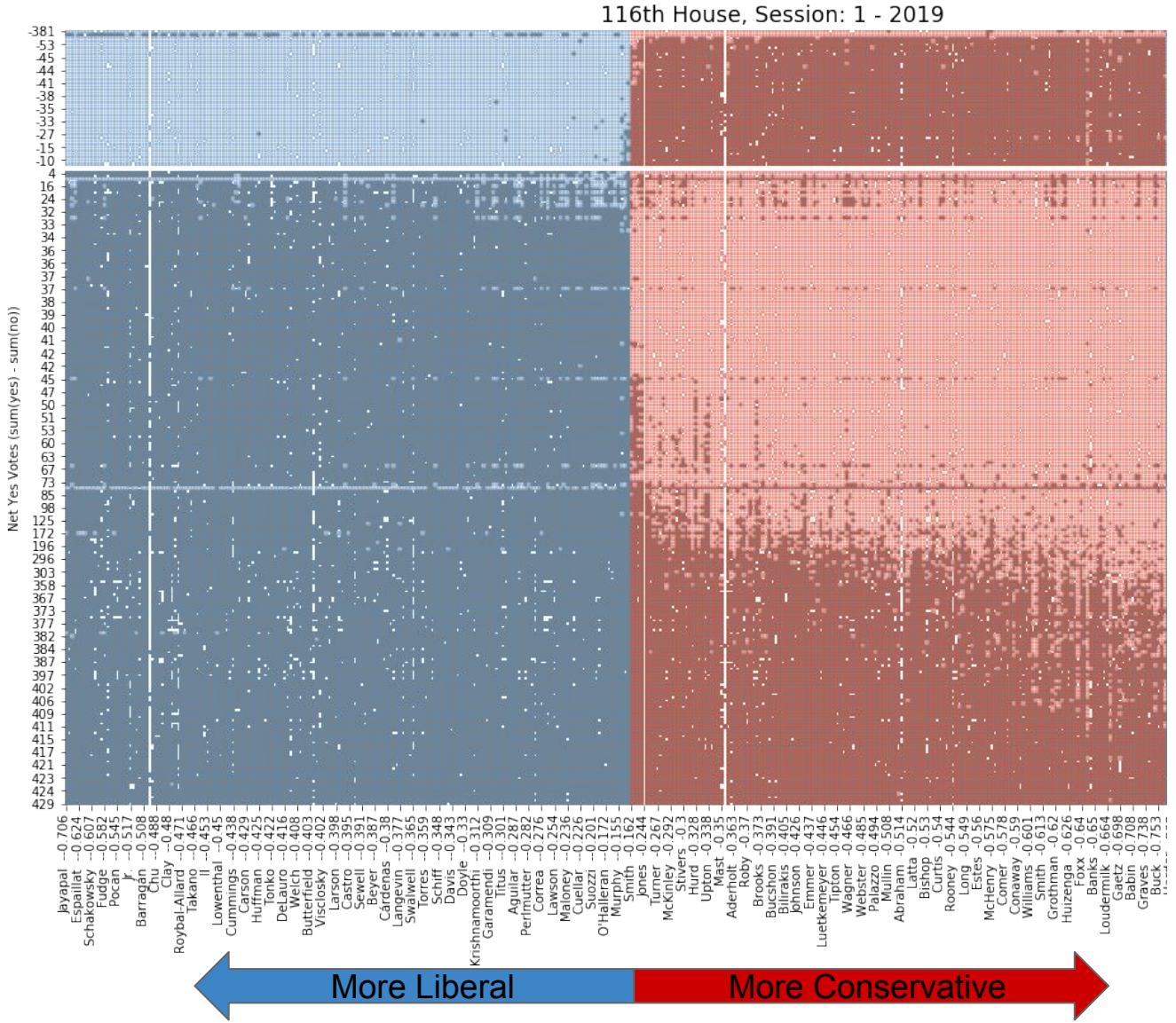




The 115th congress, session 2 of 2018, has been displayed here in a heat map with each column representing a member of congress, and each row representing a given roll call vote. Because the Republicans controlled this chamber, they fully supported nearly all of the votes that passed, and appear to be a solid red rectangular block of yes votes. However, Democrats' voting habits can be differentiated by the narrow curve of increasing support from right to left, as the bills become more popular.

What patterns are apparent in the data?



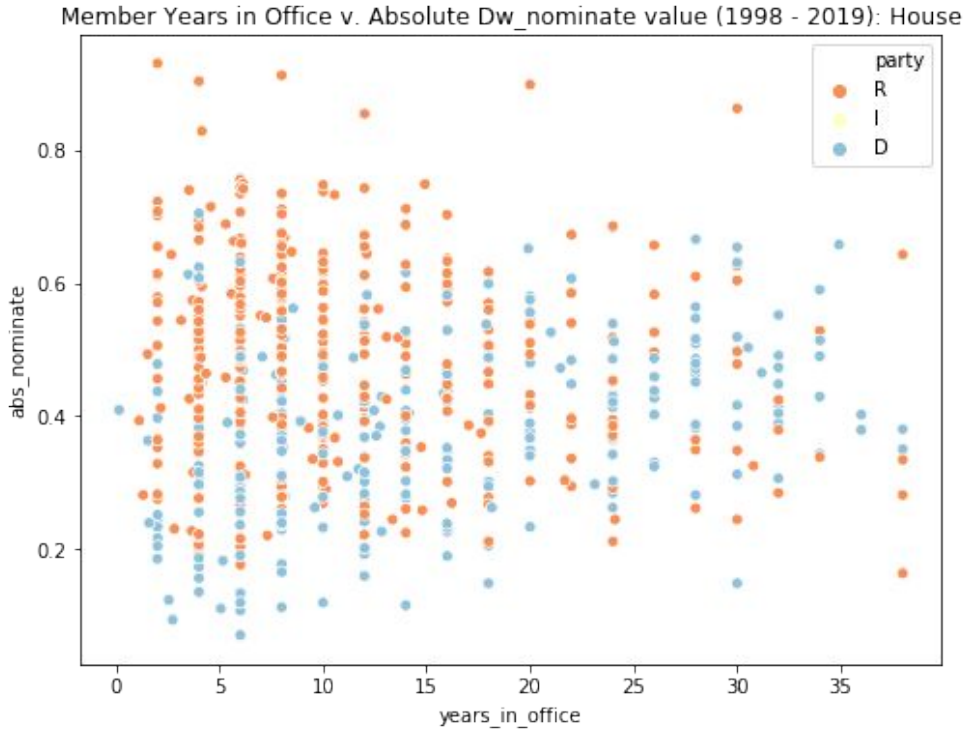


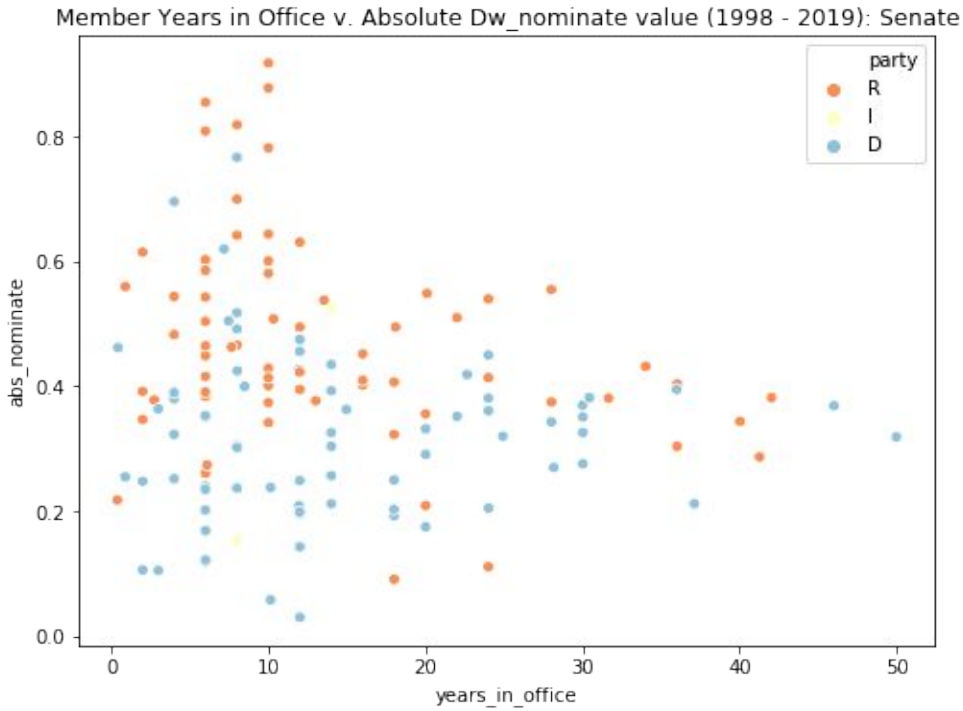
Primary Republican Differences

The same pattern can be seen in the current house (2019), except that the democrats control the chamber. Liberal / conservative scores are determined by pre-calculated “dw_nominate” values included in the ProPublica API, some of which haven’t been calculated yet.

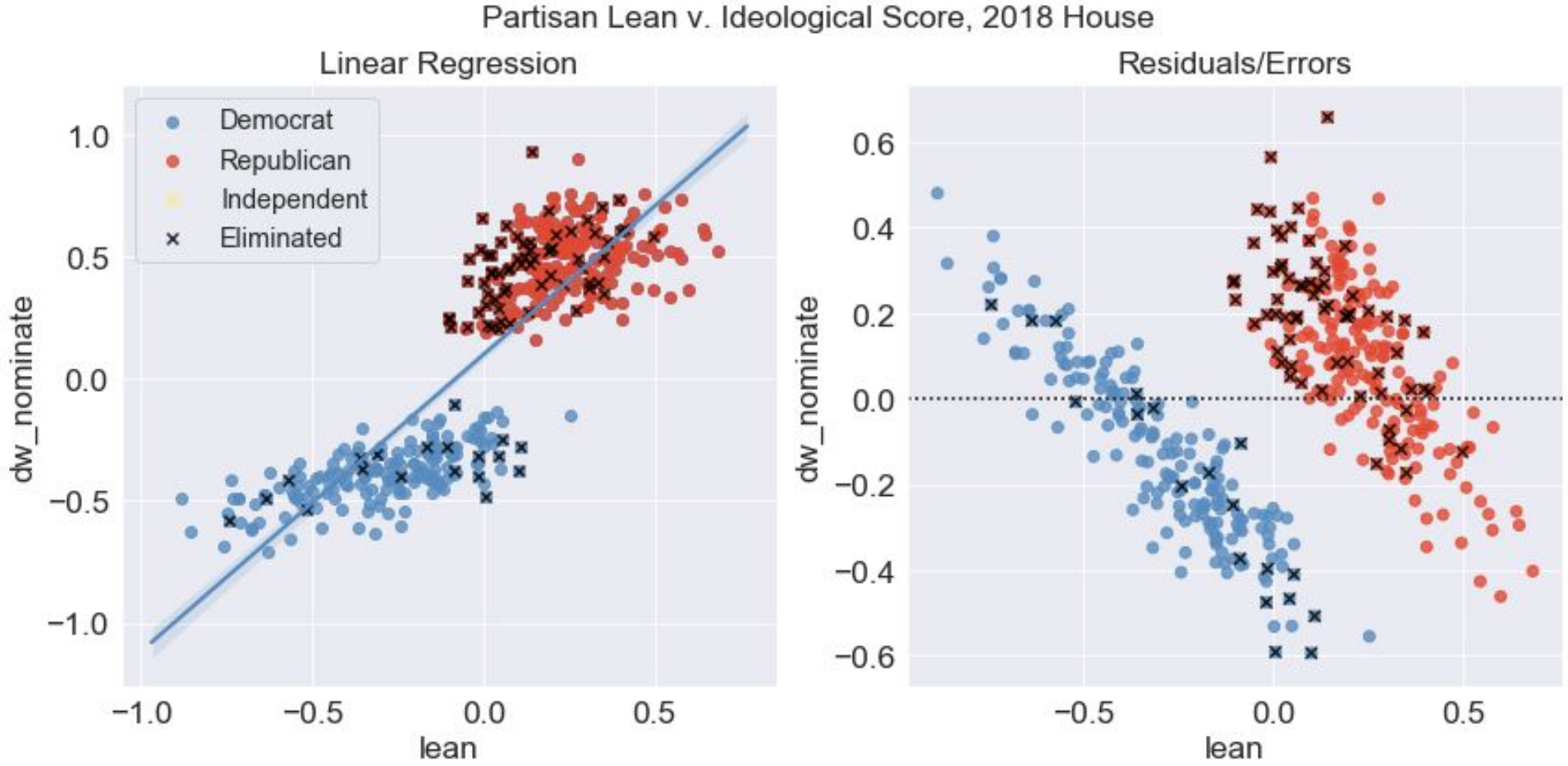
Which members tend to survive?

Longer lasting members tend to be more toward the center of their respective side of the political spectrum. Senate dw_nominate scores appear to approach .37 distance from 0 as years in office increase. House, scores appear to converge toward .5 distance from 0. Outlier dw_nominate scores are all held by house Republicans.



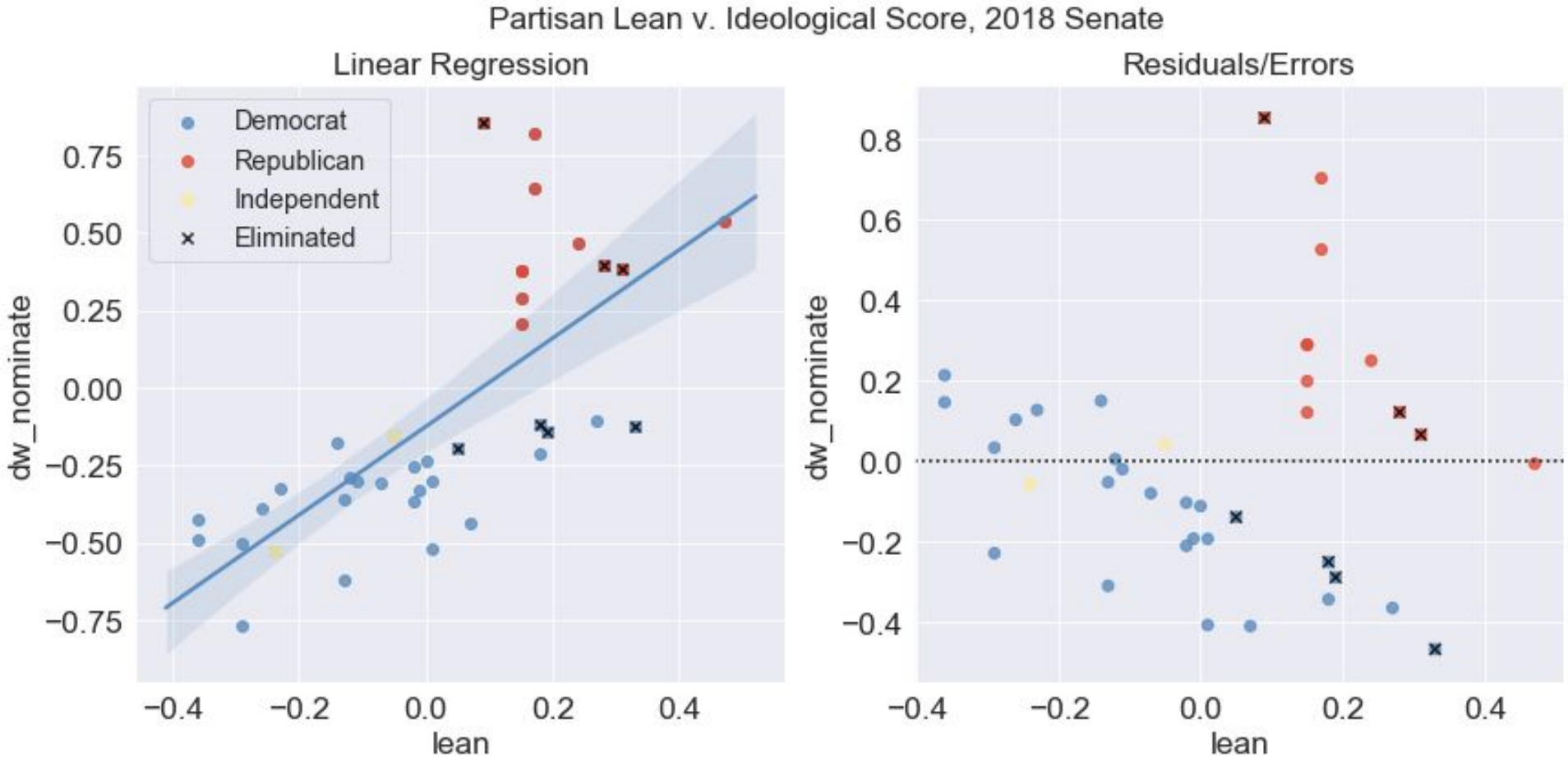


Hypothesis: Members who vote more in line with the party preferences of their constituents survive longer



Both parties are less moderate than district lean would predict by linear regression in swing districts/states, and less extreme than district lean would predict in highly partisan areas. In the residual plots on the right, members who are the most out of line with what is predicted by partisan lean are toward the upper and lower ends of the plot. It appears that those members most out of line with their district lean have been eliminated more frequently. This is particularly stark among House Democrats in districts near a partisan lean of 0

Hypothesis: Members who vote more in line with the party preferences of their constituents survive longer



(Only senators facing election in 2018 shown)

Those more extreme than predicted, seem to be eliminated with more frequency than those more moderate than predicted. This apparent trend can be investigated further with data from more years.

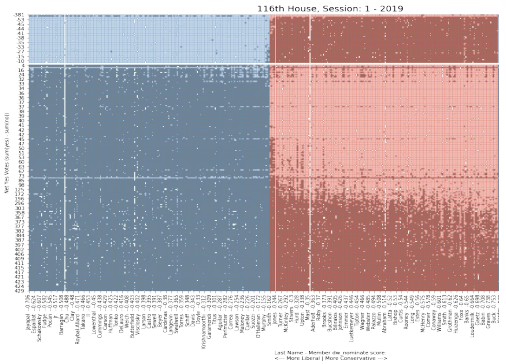
Machine Learning

Dimension Reduction (PCA)

"Dw_nominate" is a multi-dimensional scaling method, developed in the 80's where congressional voting records are interpreted along two dimensions. Let's see if we can "roll our own" dw_nominate scores with Principal Component Analysis.

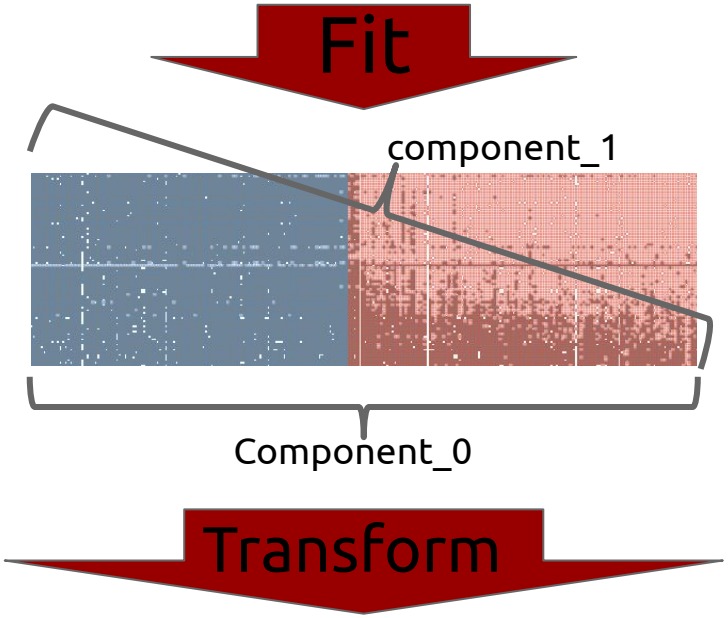
Most of the variation in the data is explained by component '0' with component '1' providing much less, but probably enough to be useful, and component '2' providing even less than that. This means it is likely possible to interpret results in 2 dimensions quite easily

>100d:

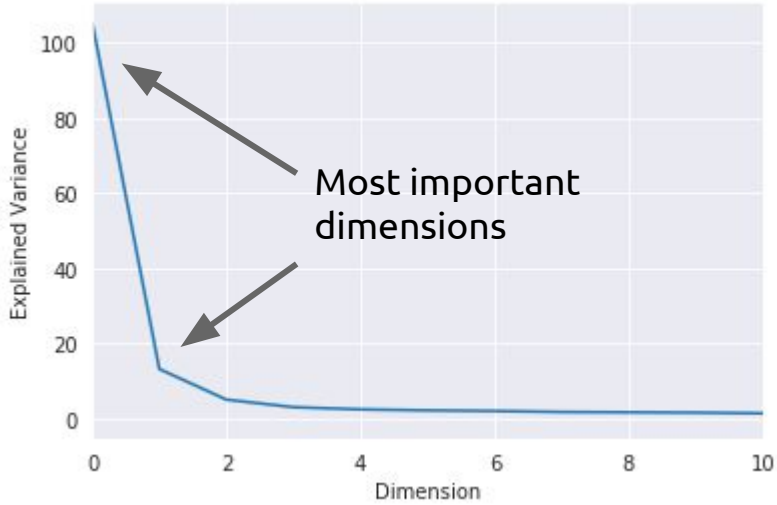


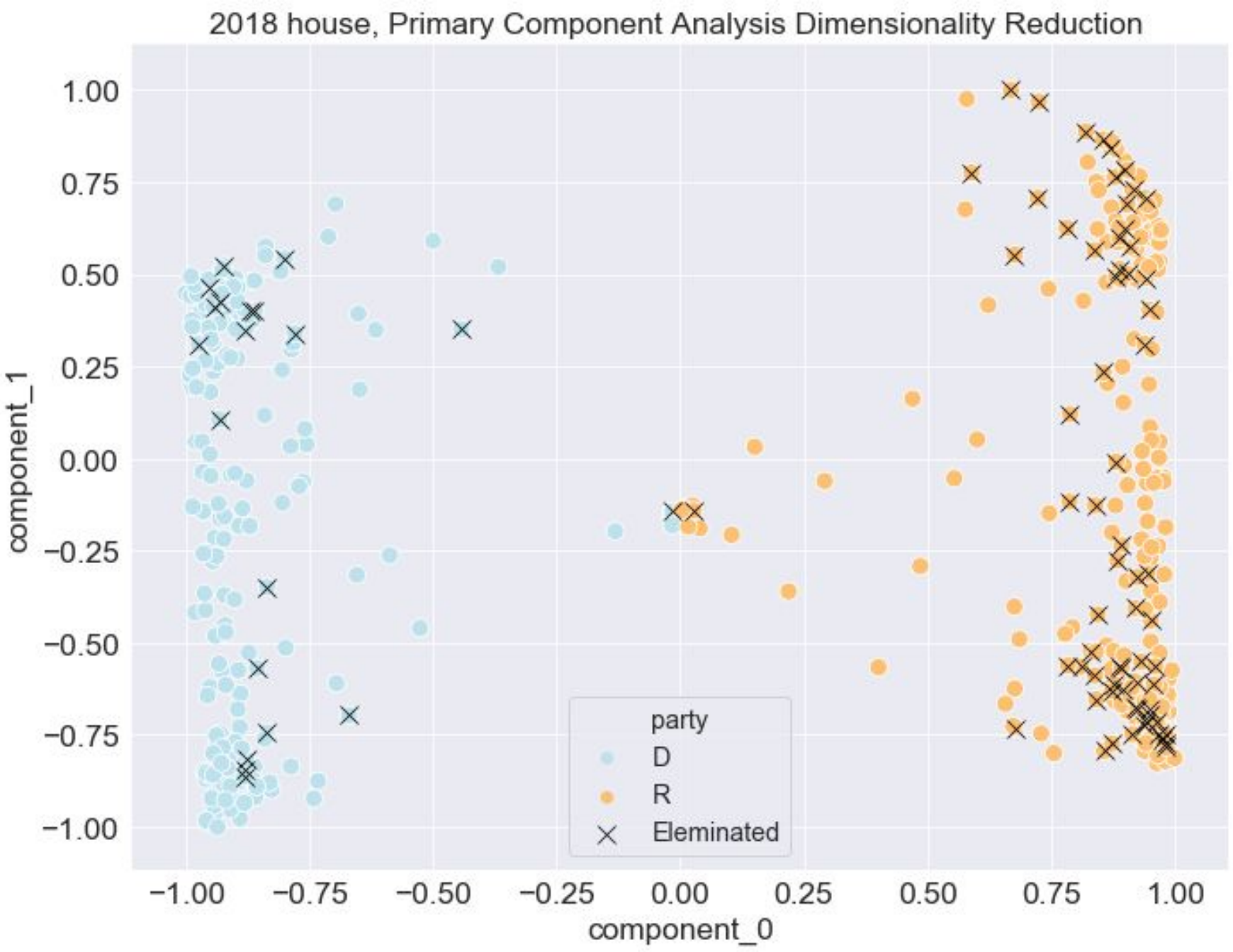
PCA reduces the hundreds of dimensions represented by individual votes by repeatedly finding an axis that represents the largest variance in the data, and projecting it into its own dimension.

Crude artistic interpretation.

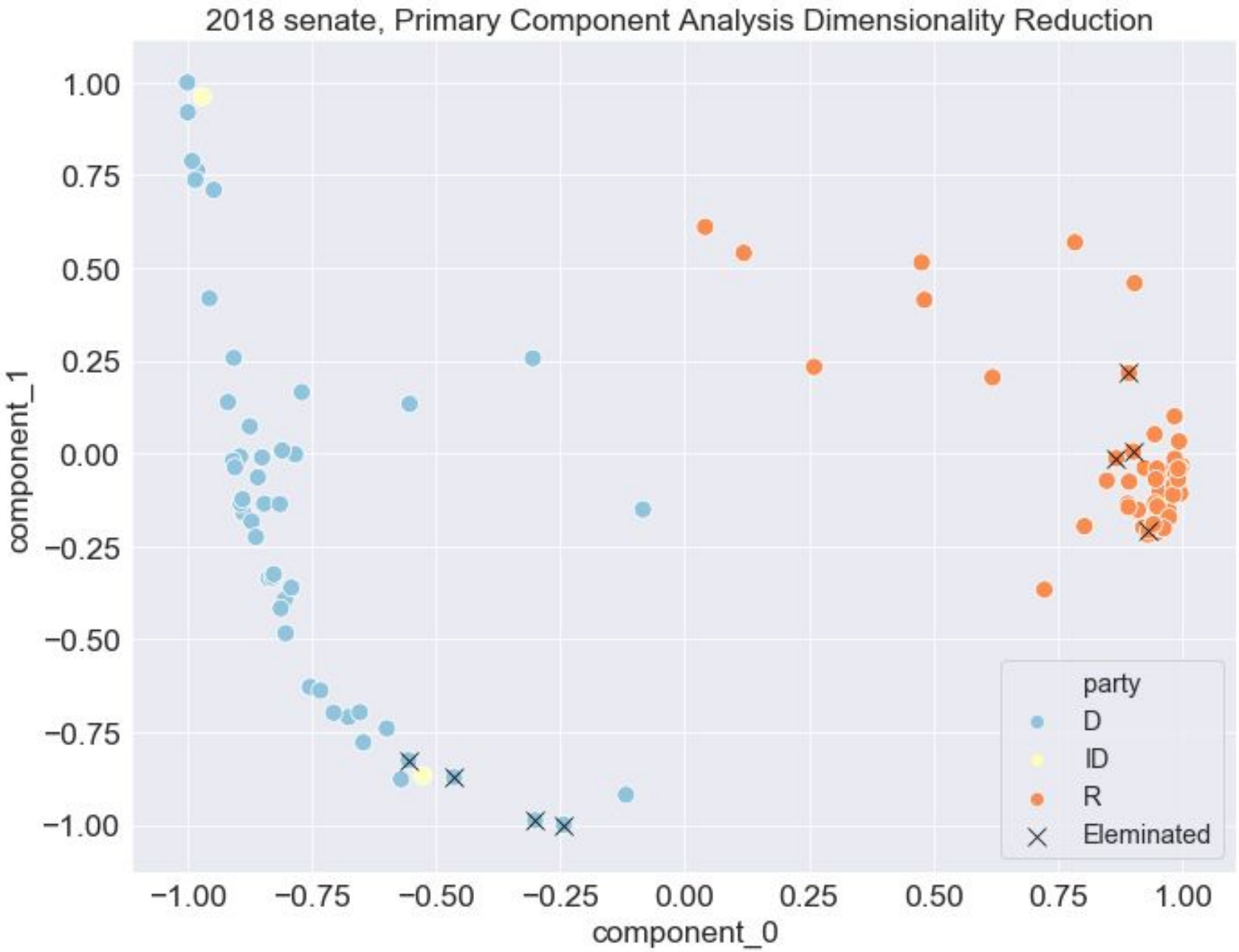


2d:

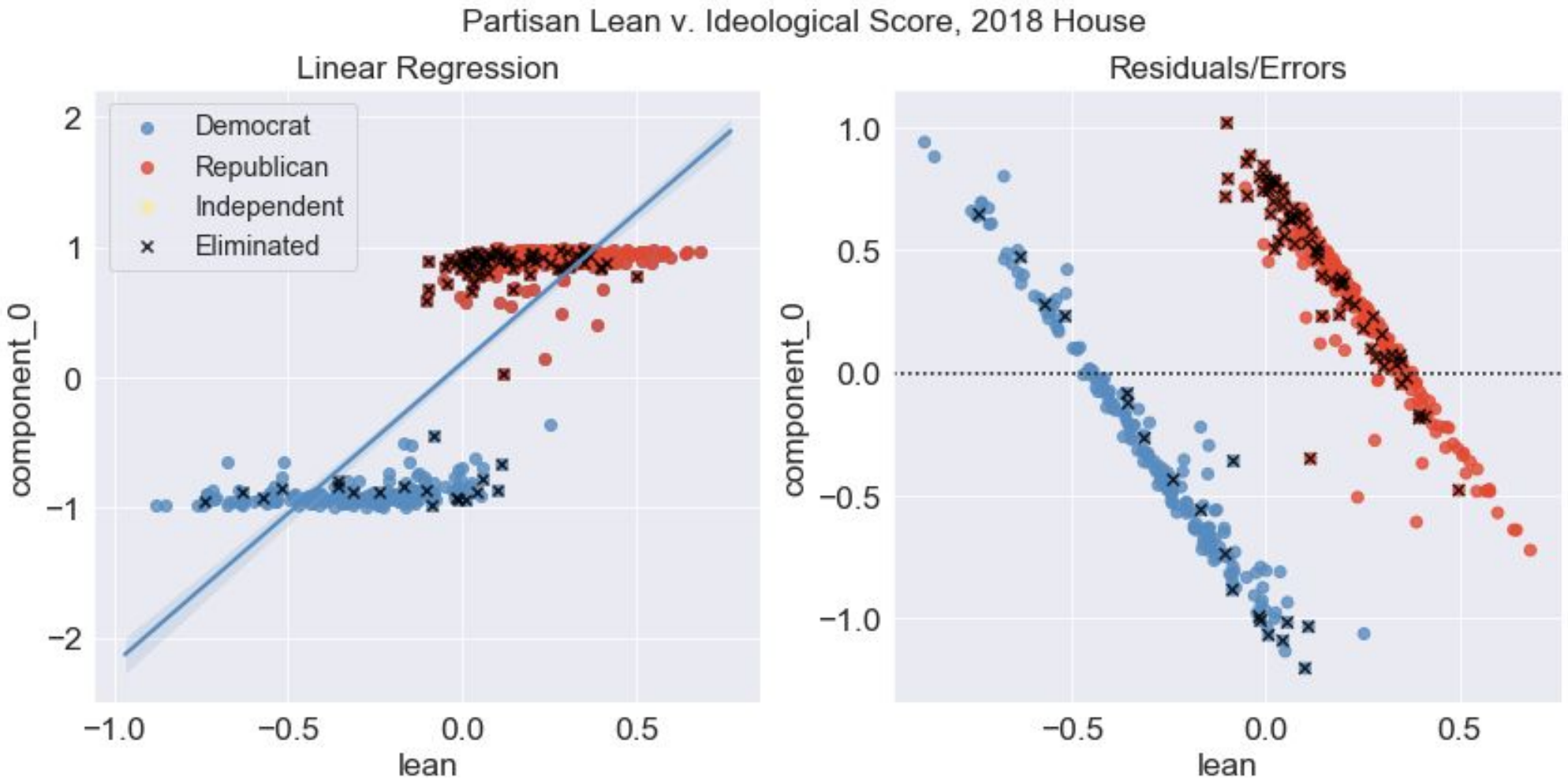




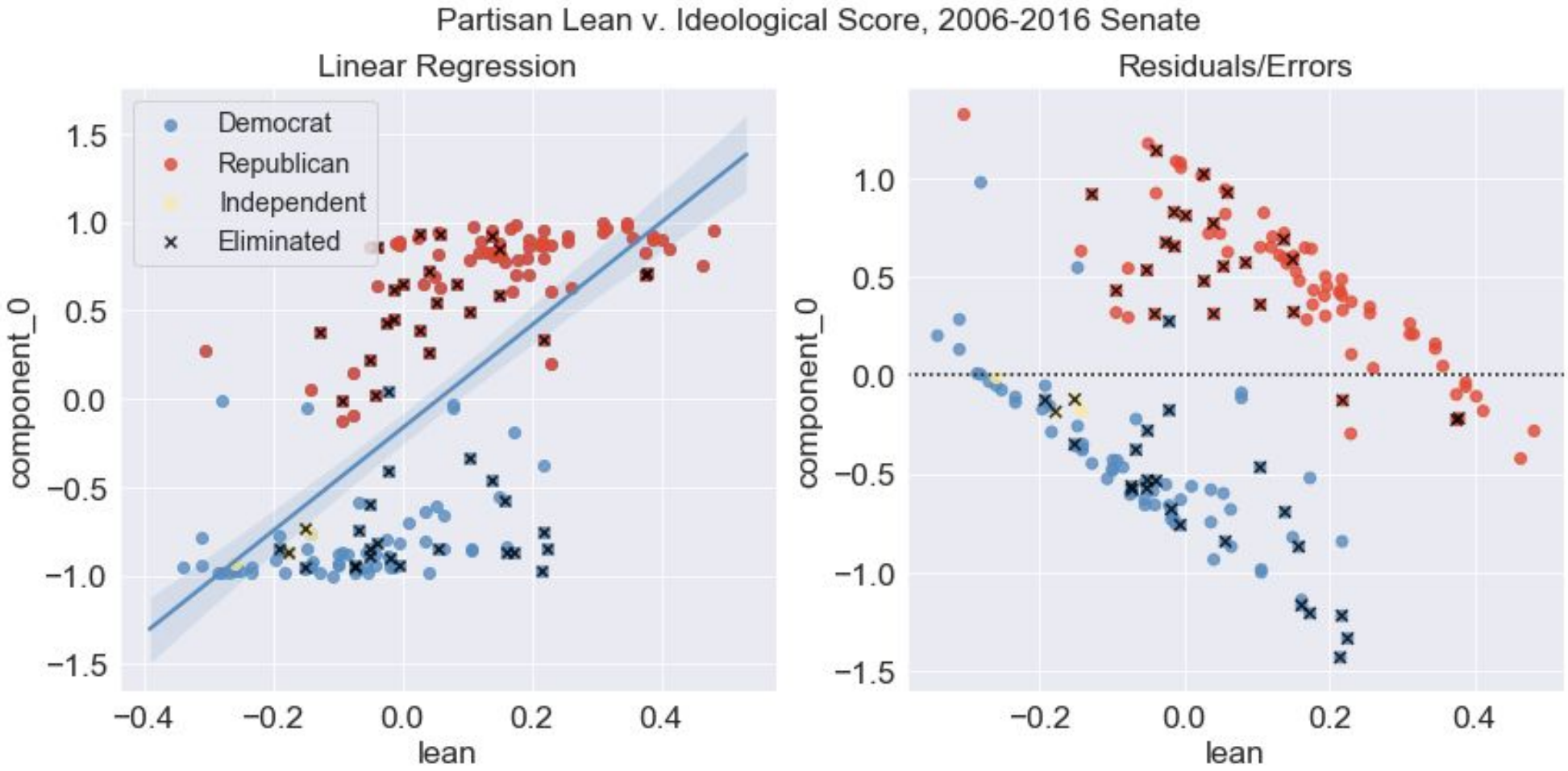
As shown, PCA can neatly separate the parties from one another based entirely on voting records, for both the house …



… and the Senate. It can probably be assumed that "component_0" roughly corresponds with liberal vs. conservative voting records. "Component_1" might align with social values, but this won't be clear without further analysis



While primary "component_0" seems to leave the parties a bit flat compared to dw_nominate scores, perhaps it can predictor of a given member's chance of being eliminated, when combined with partisan lean. However, historical partisan lean for house races is not yet open data, making incorporation of previous house races quite difficult.



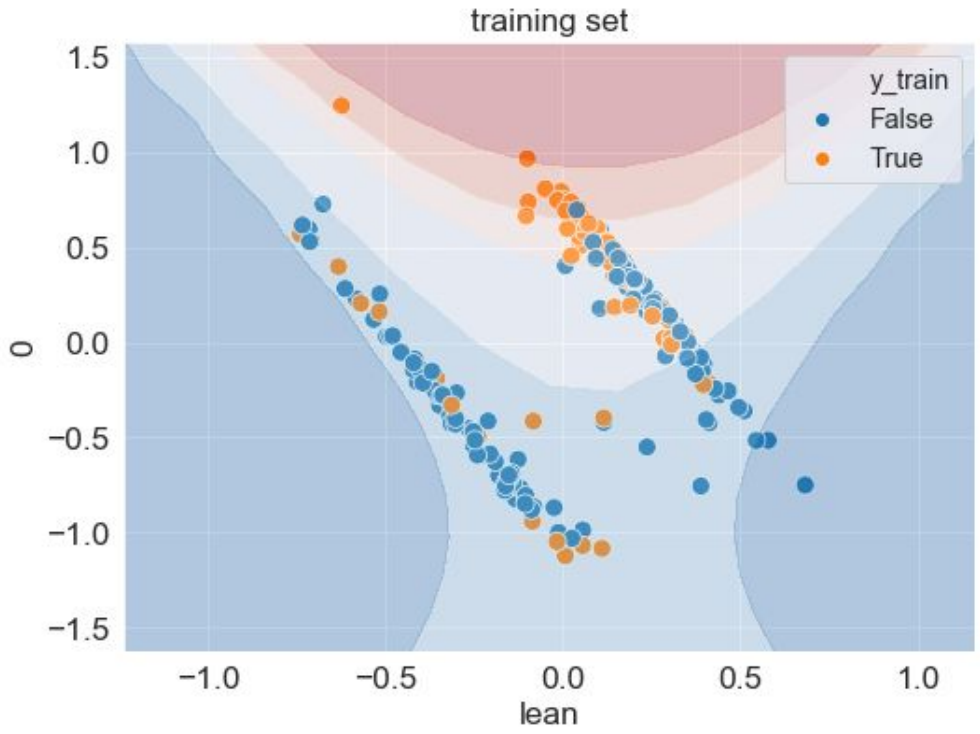
* Partisan lean according to FiveThirtyEight is the average difference between how a state or district votes and how the country votes overall w/ the current presidential election results weighted at 50%, the previous weighted at 25%, and the results from elections for the state legislature weighted at 25%

Historical partisan lean for the Senate can be relatively easily constructed using election returns, allowing the aggregation of data from senate races from 2006 to 2016. Note the apparent zones of safety along the regression line. These Senators are voting “in line” with their constituents.

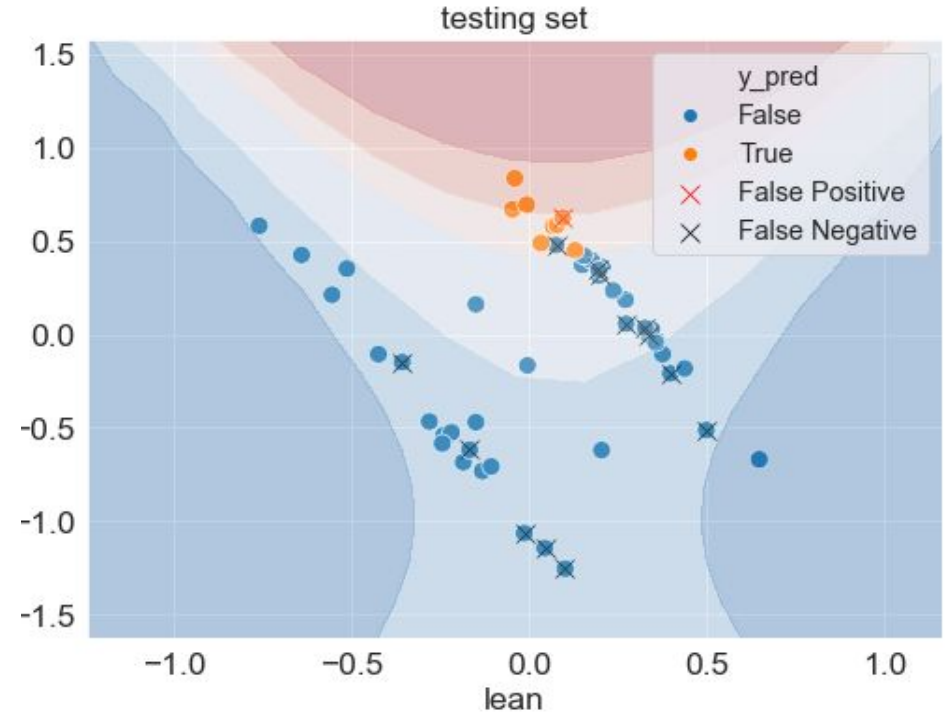
Supervised learning (Classification)

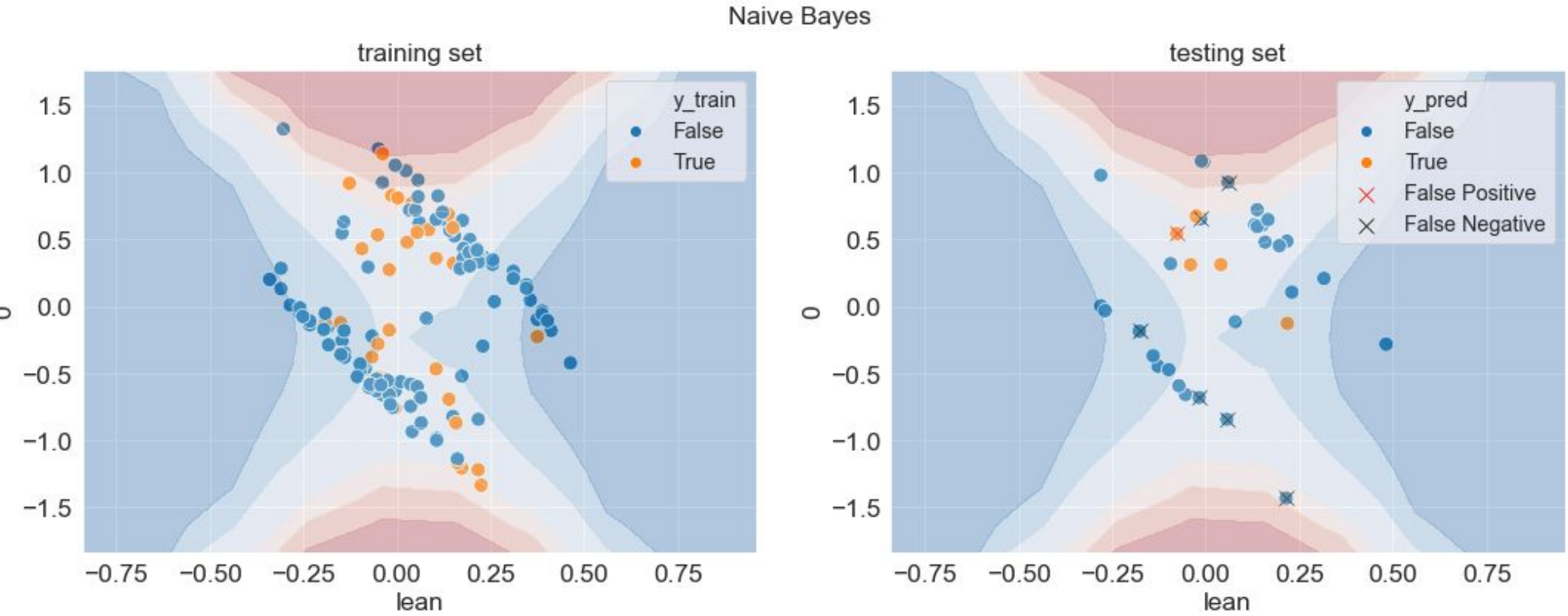
Here the Naive Bayes classifier is being fed dimension reduced voting data from the 115th house, and is being trained to predict which members will be eliminated. In the training set plot (top) members labeled “True”, were eliminated, and “False” were not.

In the testing plot (bottom), the classifier is giving it's best guess with the remaining 4th of the data. The contour lines of the decision function show 2018 was a bad year for republicans and that one year isn’t nearly enough data to generalize.



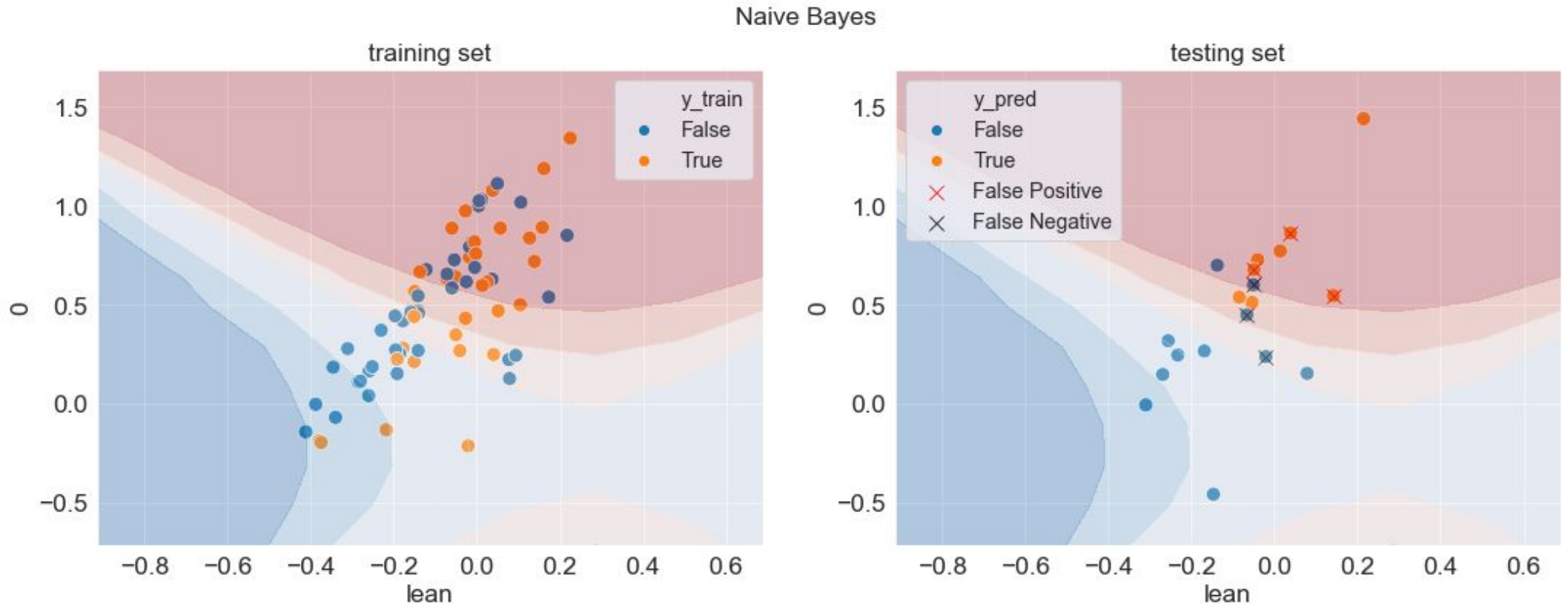
Predict





Here, the same Naive Bayes classifier is being trained on aggregate 2006 - 2016 Senate voting data to predict elimination of Senators. Note the same patterns emerge more symmetrically when combined over 10 elections.

Senators in states with the highest partisan lean in their favor are clearly the safest. However, those in states with partisan leans up to 25 points against them are still in a safer zone if their voting record matches their constituents.



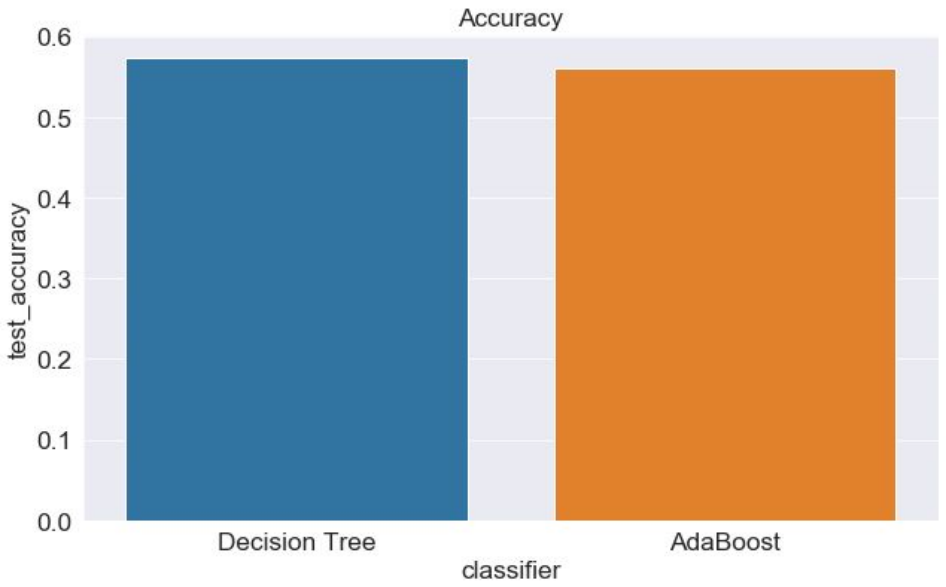
If the parties are overlaid on each other so that higher values of ‘lean’ represent preference for the other party, and the y axis represents distance from the regression line (being “in-line” with constituents). The problem becomes more simple to process, and more algorithms can beat baseline in their predictions.

Improvement against baseline

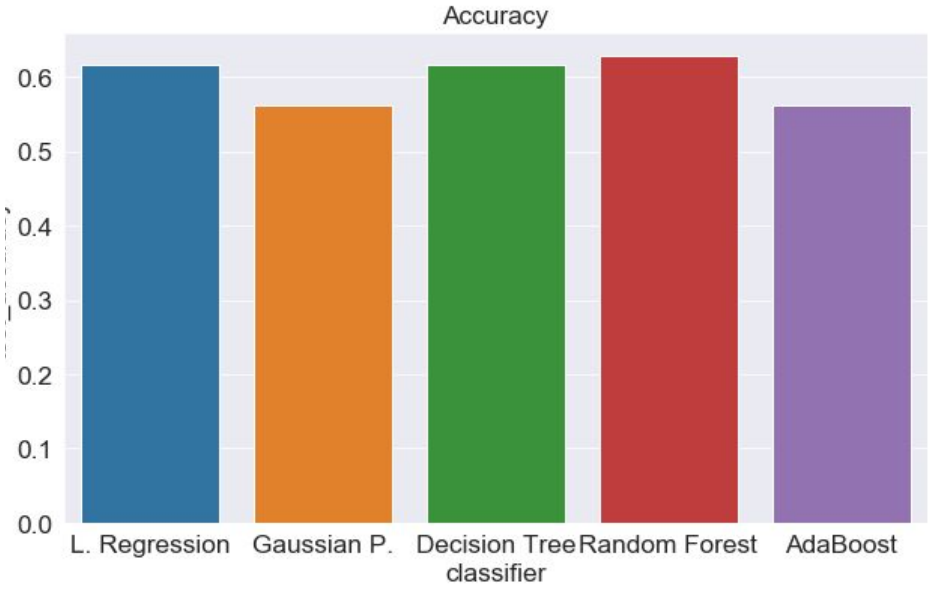
The training and testing sequence is done with each 4rd of the data, using the other two thirds for training, to create a 4-fold cross validation score average, which has been done here across several different classifiers.

Undersampling survivors produces a baseline of 0.55, which an algorithm could achieve by guessing all survived.

By overlaying / transposing parties 5 classifiers achieve a mean score over this baseline.



Mean cross validation scores - using residuals



Mean cross validation scores - residuals with parties transposed

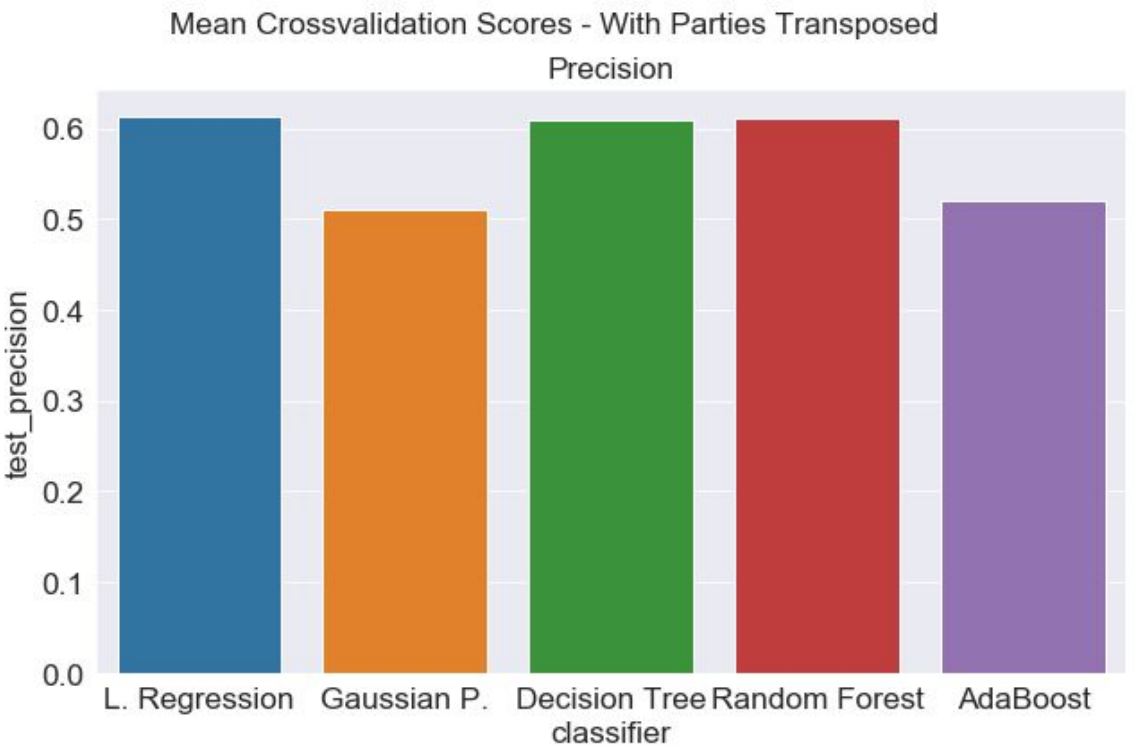
Recall v. Precision

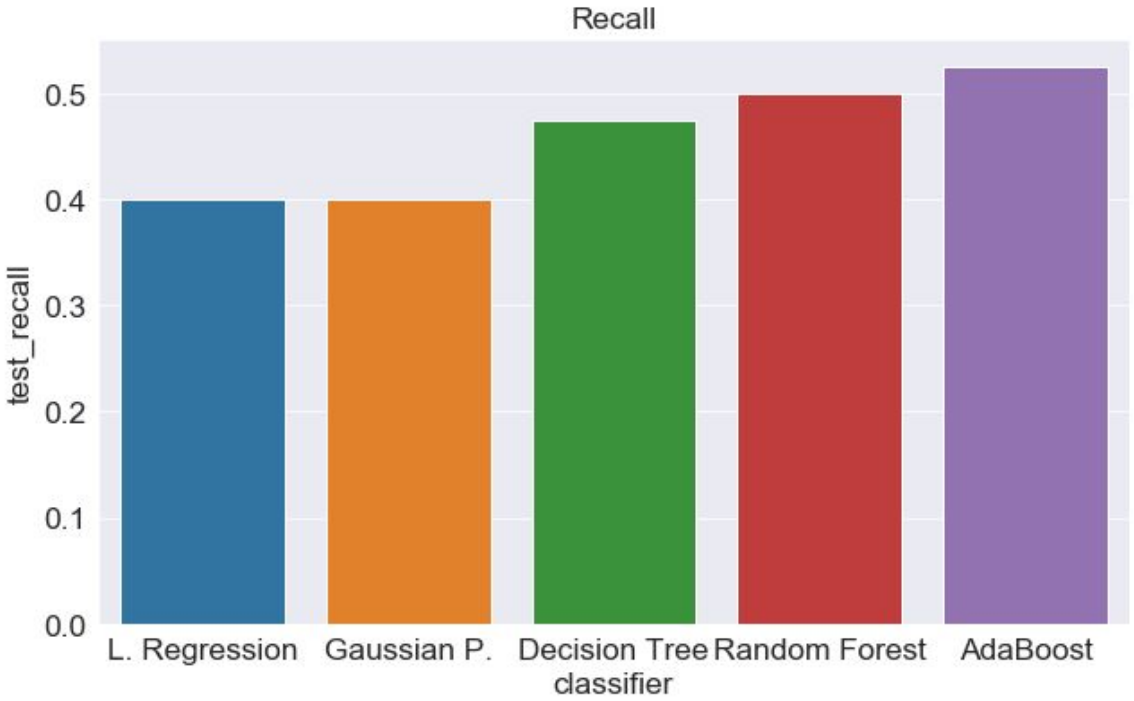
"Precision" is the proportion of positive classifications that are correct.

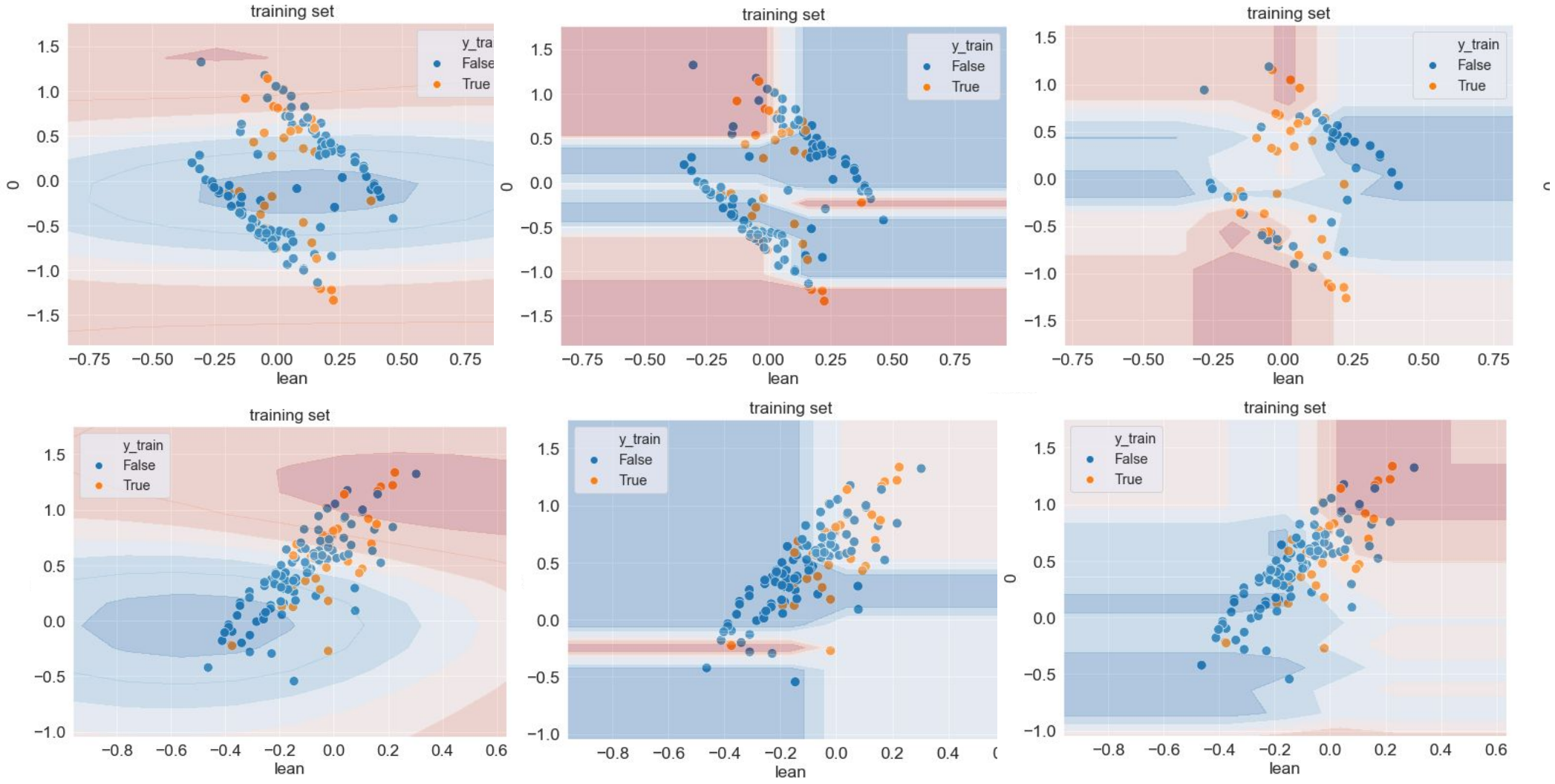
"Recall" is the proportion of total positive values that have been correctly classified.

Random forrest's precision of .6 and recall of .5 means when it predicts a given house member is eliminated it is correct 60% of the time, and is able to identify 50% of those which were eliminated.

Not great; Not terrible. But, we have already answered our hypothesis







Gaussian Process

Decision Tree

Random Forest

The reduction in complexity for the a given machine learning algorithm can be seen in the decision boundary contours of three of the classifiers performing above baseline.

Findings:

1. Voting behavior “in line” with the partisan lean of constituents is an important predictive indicator of congressional survival. Members who do so survive longer
2. Members with voting records more “extreme” than the partisan lean of their constituents are less likely to survive than those with a more moderate one than their constituents
3. There is a correlation between longer surviving members and voting records closer to the center of the ideological spectrum of their party.

Recommendations:

1. To survive in an R+22 district, pursue one of the most moderate voting records in the house. Otherwise survival is highly unlikely.
2. Members take on less risk by voting more “moderate” than the partisan lean in their district than more “radical.” So it is better for the survival of the delegation for members in safe districts to vote less radical than it is for members in swing districts to be less moderate.
3. Find as many opportunities as possible to find common ground and vote in line with constituents without sacrificing core values

Possible Next Steps

1. Bite the bullet. Buy historical partisan lean data.

   As historical partisan lean from FiveThirtyEight and the Cook Political Report is not open, just buy it. Then it will be possible to build a house model. Alternately, assist “open elections” project in producing it.

2. Select and tune the best performing classifier.

   Calculate area under ROC curve, export applicable coeficents, and Identify precise percentage chances of elimination for a given voting score and partisan lean

3. Map which individual votes effected chances of elimination the most.

   This can probably done with Naive Bayes classifier without using dimension reduction.

Appendix

1. Datasets
    a. ProPublica Congress API
    b. Partisan Lean - FiveThirtyEight
        i. District
        ii. State
    c. Election returns
        i. County
        ii. State
2. Project on git-hub
    a. https://github.com/Nhorning/Congressional_Survival