Uncovering the Truth Behind Primary Election Endorsements and Financing

Sammantha Garcia, Salvador Ramirez Jr., Momo Siu, Rane Tzeng
University of California, Berkeley
DATA C102: Data, Inference, and Decisions
Professor Ramesh Sridharan, Professor Jacob Steinhardt
December 12, 2022

Table of Contents

Research Question 1:

Do endorsements for Democratic and Republican candidates cause them to win in the primary election?

Data Overview

Our team retrieved the following two datasets from the *GitHub* repository provided by *FiveThirtyEight*: `rep_candidates.csv` and `dem_candidates.csv`. These datasets included census data for the Republican and Democratic candidates running for U.S. Senate, U.S. House, and governor in the 2018 primary election. *FiveThirtyEight* collected data from several resources, such as: *Ballotpedia*, *New York Times*, *VoteSmart*, and the websites of candidates. According to the *FiveThirtyEight* repository, races that included Democratic and Republican incumbents were neglected from these datasets, which brings concern for selection bias. Since all of the data was publicly available, we believe the candidates are aware of the collection of data but might be unaware of the use of it.

Both datasets, `rep_candidates.csv` and `dem_candidates.csv`, include the following features:

- `Candidate`: name of the candidate in 2018 primary election
- `State`: the U.S. state a candidate is running in
- `District`: the congressional district a candidate is running for
- `Office Type`: the office a candidate is running for
- `Race Type`: whether the election was regular or special
- `Race Primary Election Date`: when the primary election was held
- `Primary Status`: whether the candidate won or lost the primary election
- `Primary Runoff Status`: whether the candidate was involved in a runoff election or not
- `General Status`: if the candidate made it to the November ballot
- `Primary %`: percentage of the vote received by the candidate in their primary
- `Won Primary`: if the candidate won the primary (‘Yes’) or lost the primary (‘No’)

In addition, the `rep_candidates.csv` includes information on whether candidates were endorsed or supported by the following:

- `Rep Party Support?`
- `Trump Endorsed?`
- `Bannon Endorsed?`
- `Great America Endorsed?`
- `NRA Endorsed?`
- `Right to Life Endorsed?`
- `Susan B. Anthony Endorsed?`
- `Club for Growth Endorsed?`
- `Koch Support?`
- `House Freedom Support?`

- `Tea Party Endorsed?`
- `Main Street Endorsed?`
- `Chamber Endorsed?`

The `dem_candidates.csv` includes additional personal information about the candidates:

- `Gender`: gender of the candidate
- `Partisan Lead`: the partisan lean where the candidates' election was held
- `Race`: the race of the candidate
- `Veteran?`: whether the candidate served in the armed forces ('Yes') or did not ('No')
- `LGTQ?`: whether the candidate is LGBTQ ('Yes') or is not ('No')
- `Elected Official?`: whether the candidate has previously held elected office ('Yes') or has not ('No')
- `Self-Funder?`: whether the candidate self-funded their campaign ('Yes') or did not ('No')
- `STEM?`: whether the candidate has a background in STEM ('Yes') or has not ('No')
- `Obama Alum?`: whether the candidate worked in the Obama Administration ('Yes') or did not ('No')

The `dem_candidates.csv` also includes information on whether the candidates received endorsements from the following:

- `Dem Party Support?`
- `Emily Endorsed?`
- `Gun Sense Candidate?`
- `Biden Endorsed?`
- `Warren Endorsed?`
- `Sanders Endorsed?`
- `Our Revolution Endorsed?`
- `Justice Dems Endorsed?`
- `PCCC Endorsed?`
- `Indivisible Endorsed?`
- `WFP Endorsed?`
- `VoteVets Endorsed?`
- `No Labels Support?`

The `dem_candidates.csv` dataset included more personal information about the Democratic candidates, making it more granular compared to the `rep_candidates.csv` dataset. Each row in the datasets represents one candidate in the primary election. Since we have more information about Democratic candidates, this might impact the interpretation of our findings because there are more variables to consider and make assumptions for. *FiveThirtyEight* also used several different external resources to acquire personal information for candidates in this political party. Most personal data was retrieved from the candidates' websites, therefore,

FiveThirtyEight could only access information that the candidates consented to publicizing. Therefore, it might be potentially inaccurate and introduce bias into our study. In addition, for both datasets, there are blank cells for when a candidate's information is unknown or unretrievable. Lack of information could negatively affect our conclusions and produce inaccurate results.

Exploratory Data Analysis

To gain insight into the relationship between election outcome and number of endorsements, we created density plots for each political party, based on whether or not they were endorsed.

It's important to note that a candidate has a primary % of 100% if they are running unopposed or are nominated by convention.

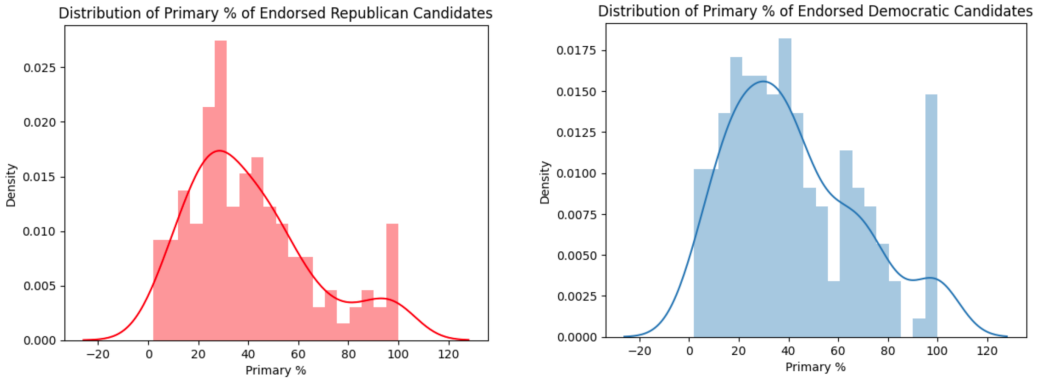


Figure 1a: Distribution of Primary % of Endorsed Candidates

In Figure 1a, we can see that both of the distributions roughly have the same shape and are skewed to the right. These visualizations also show that the majority of endorsed candidates for both political parties have a primary % around 30%.

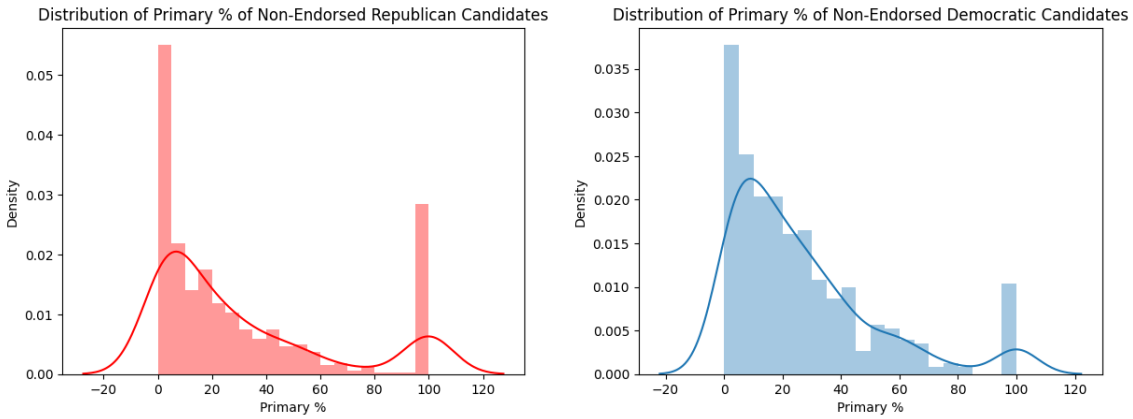


Figure 1b: Distribution of Primary % of Non-Endorsed Candidates

In Figure 1b, there are fewer candidates that are non-endorsed. Similar to Figure 1a, both visualizations roughly have the same shape and are skewed towards the right. However, the majority of these non-endorsed candidates have a primary % around 10-15%.

Observing the distributions of primary %, we can see that endorsed candidates generally have a higher primary % whereas non-endorsed candidates have a lower primary %, which shows support for our research question. For Figures 1a and ab, there were no data cleaning steps necessary.

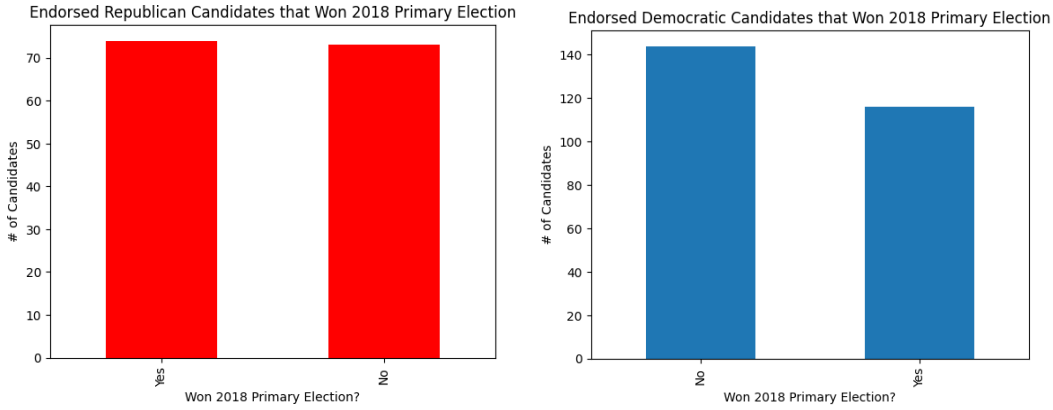


Figure 2a: Number of Endorsed Candidates that Won the Primary Election

Figure 2a shows that there are more endorsed Democratic candidates that lost the primary election than won the primary. Surprisingly, we see that about half of the endorsed Republican candidates won the primary election and the other half lost. However, upon closer examination, there's slightly more endorsed Republican candidates that won than have lost which supports our research question. Since this figure shows that there's a positive relationship between endorsed Republican candidates and election outcome, we're interested in following up on this and seeing if these results will be consistent with our findings.

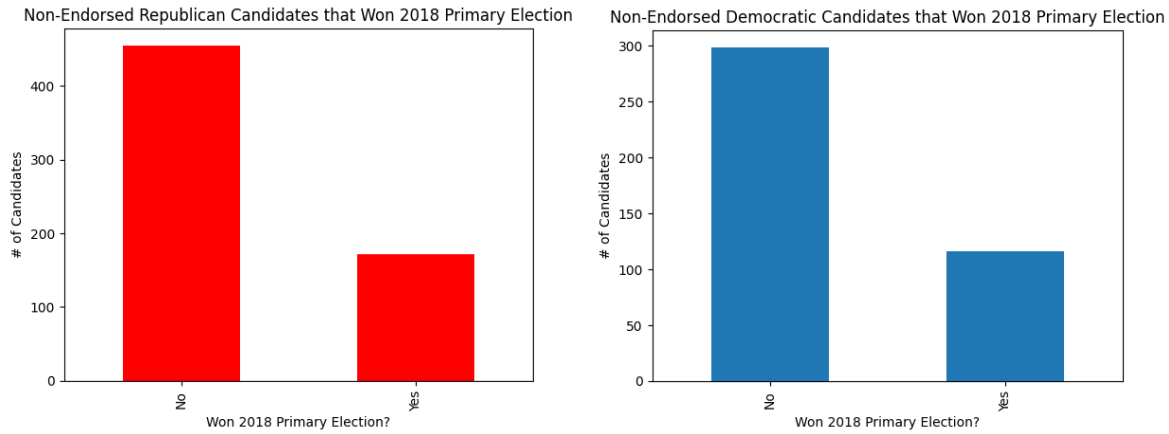


Figure 2b: Number of Non-Endorsed Candidates that Won the Primary Election

Figure 2b shows that the number of non-endorsed candidates that lost the primary election exceeds the number of non-endorsed candidates that won; this observation is applicable to both political parties and it shows support for our research question. For Figures 2a and 2b, there were no data cleaning steps necessary.

Models

Our research question is: Do endorsements for Democratic and Republican candidates cause them to win in the primary election? By answering this research question, voters in the U.S. will be able to form genuine and accurate opinions on candidates because they will know who and what organizations drive their campaigns. Voters will also know the true policy views of the candidates which would affect their voting outcome.

To understand the direct effect of endorsements on election outcomes, our team decided to use causal inference because we are proving that endorsements cause specific outcomes for the primary election. For this study, the treatment (Z) and potential outcome (Y) variables are:

***Z*:** whether or not a candidate receives endorsements

$Z_i = 0$: if the candidate does not receive an endorsement for the 2018 primary election

$Z_i = 1$: if the candidate does receive an endorsement for the 2018 primary election

***Y*:** the outcome of the 2018 primary election for a candidate

$Y_i(0)$: if the candidate lost the 2018 primary election

$Y_i(1)$: if the candidate won the 2018 primary election

The units i for this observational study represent a single candidate in the 2018 primary election. For example, since there are 811 Democratic candidates on the ballot for this specific election year, there are 811 units in that dataset.

Since we chose to have the treatment represented as a binary variable, we recorded the number of endorsers that a candidate has. If a candidate has zero endorsers, they are represented with a '0' and if a candidate has one or more endorsers, they are represented with a '1'.

For each Democratic candidate in the `dem_candidates.csv` dataset, we counted the total number of endorsers for the following:

- `Dem Party Support?`
- `Emily Endorsed?`
- `Gun Sense Candidate?`
- `Biden Endorsed?`
- `Warren Endorsed?`
- `Sanders Endorsed?`
- `Our Revolution Endorsed?`
- `Justice Dems Endorsed?`
- `PCCC Endorsed?`
- `Indivisible Endorsed?`
- `WFP Endorsed?`
- `VoteVets Endorsed?`
- `No Labels Support?`

Similarly, we counted the total number of endorsers for Republican candidates:

- `Rep Party Support?`
- `Trump Endorsed?`
- `Bannon Endorsed?`
- `Great America Endorsed?`
- `NRA Endorsed?`
- `Right to Life Endorsed?`
- `Susan B. Anthony Endorsed?`
- `Club for Growth Endorsed?`
- `Koch Support?`
- `House Freedom Support?`
- `Tea Party Endorsed?`
- `Main Street Endorsed?`
- `Chamber Endorsed?`

Confounding variables (X_D) for the Democratic candidates include the following:

- `Partisan Lean`
- `Veteran?`
- `LGBTQ?`
- `Elected Official?`
- `Self-Funder?`

- `STEM?`
- `Obama Alum?`
- `Party Support?`
- `Office Type_Governor`
- `Office Type_Representative`
- `Office Type_Senator`
- `Race Type_Regular`
- `Race Type_Special`
- `Race_Nonwhite`
- `Race_White`

Confounding variables (X_R) for Republican candidates include the following:

- `Office Type_Governor`
- `Office Type_Representative`
- `Office Type_Senator`
- `Race Type_`
- `Race Type_Regular`
- `Race Type_Special`
- `Rep Party Support?_No`
- `Rep Party Support?_Yes`

Since this is an observational study and not a randomized experiment, we are assuming that the treatment and outcomes are conditionally independent given the set of confounding variables for each respective political party. We are using all of the observed confounding variables in the datasets, therefore, the unconfoundedness assumption holds for this observational study. To adjust for confounding variables, we will be using unconfoundedness and conditional independence, more specifically, we will be calculating outcome regression. Also, there aren't any colliders and instrumental variables in our study.

Results

After setting up the treatment, potential outcomes, and the confounders, we fitted the data using logistic regression from `sklearn.linear_model` and calculated the estimated treatment effect for each political party. For the candidates in the Republican Party, the estimated treatment effect was 0.189 and the candidates in the Democratic Party had an estimated treatment effect of 0.197. Based on these results, it seems that there's barely any causal effect between our treatment and potential outcomes. However, we can assert that the treatment has a stronger causal effect on election outcomes for candidates in the Democratic Party compared to candidates in the Republican Party.

In our EDA, Figures 2a and 2b show how many candidates won the primary election based on whether or not they received endorsements. For Republican candidates, endorsed and non-endorsed, the visualizations are consistent and support our research question. For the Democratic Party, the visualizations show that the majority of the candidates still lost the primary election whether they were endorsed or not. This does not support the results from calculating estimated treatment effects as the Republican Party had a lower estimated treatment effect than the Democratic Party.

The biggest limitation of our method came from the fact that this was an observational study. Observational studies are generally more difficult when it comes to establishing a causal relationship between variables because researchers have to make a lot of assumptions about the data. Making incorrect assumptions could potentially lead to inaccurate results and conclusions so our team had to make careful assumptions. To help answer our causal research question, it would've been useful to obtain data on how much each candidate received from their endorsers. With this additional information, we can get a better understanding of the effects of endorsed candidates; analyzing how much a candidate receives rather than how many endorsements they have might be more accurate in predicting whether or not they will win the primary election. Regardless of the results, our team is fairly confident in the causal relationship between our chosen treatment and potential outcomes because we believe that political candidates are always seeking money and are willing to compromise their values in order to be endorsed and win the election.

Conclusion

Following our results, we found that there wasn't a causal relationship between endorsed candidates and election outcomes. Our results are not generalizable because the confounding variables used in this study will vary from candidate to candidate. In addition, our findings are fairly broad because we don't know the significance and the impact of the endorsements on candidates, for instance.

Knowing that there's a causal relationship between endorsed candidates and election outcomes, it would be essential for the public to know who or what organizations endorse the candidates. While this information is currently public, it is still not advertised enough for voters to know. That said, there should be a policy that forces candidates to explicitly reveal who they are being endorsed by and how much they are receiving. This would change a lot of voters' perspective of the candidates and, therefore, their voting decision.

Since the data was collected through the candidates' website, we only have information that they consented to releasing publicly, which was a big limitation in the data. We could not account for this in our analysis because we don't know if candidates were truthful about who or what organizations endorsed them. Some future studies that could build on our work might include data analysts that want to examine how greater financing from endorsements causes candidates to win or how the number of endorsements might cause an increase in primary %.

Research Question 2

Given financing contributions for a given democratic candidate, can we predict the proportion of votes they will receive in a primary election?

Data Overview

Similar to our first research question, we used the `dem_candidates.csv` dataset. This dataset includes information on the Democratic candidates running for U.S. Senate, U.S. House, and governor in the 2018 primary election. Combined, there was a total of 801 candidates in the dataset.

Financing data was collected by the *Federal Election Commission (FEC)* and accessible from their website. Our dataset of interest was `candidate_summary_2018.csv`, which contains information about each candidate that appears on an official state ballot for House or Senate in 2018 or is registered with the FEC. In total, there were 3,793 candidates in this dataset. Below, we list our attributes of interest.

- `Cand_Name`: Name of Candidate
- `Cand_Party_Affiliation`: Candidate Party Affiliation
- `Individual_Itemized_Contribution`: Contributions from an individual to the campaign that exceeds $200
- `Individual_Unitemized_Contribution`: Contributions from an individual to the campaign that does not exceed $200
- `Other_Committee_Contribution`: Contributions from PACs (Political Action Committees) and other candidates
- `Party_Committee_Contribution`: Contributions from Party Committees
- `Cand_Contribution`: Contributions from the candidate
- `Transfer_From_Other_Auth_Committee:` Contributions from other committees working for a candidate's election

The FEC dataset also consisted of aggregates (sum) of the previous attributes. We chose to ignore those columns and instead use the individual attributes to more accurately capture the difference between contribution types.

Exploratory Data Analysiss

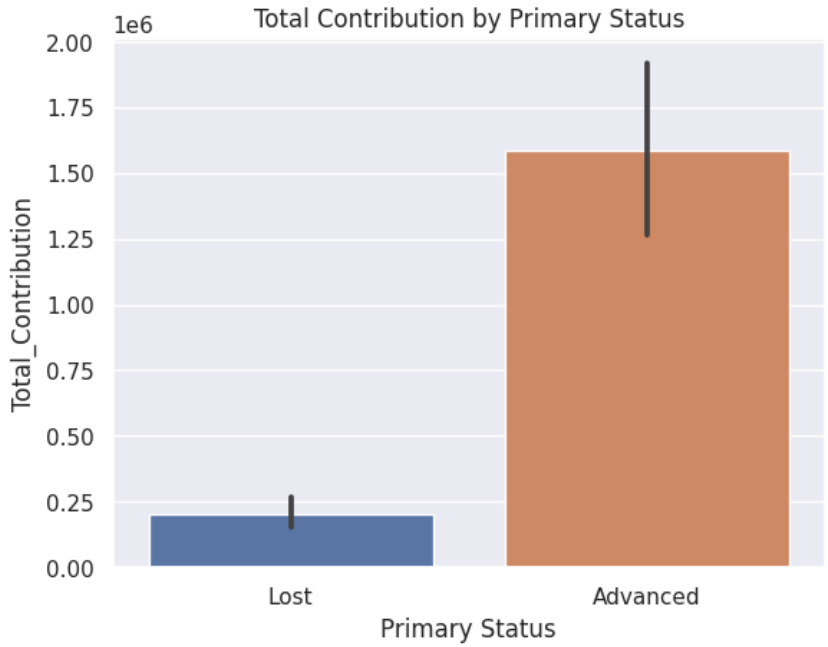


Figure 3: Contribution Amount by Contribution Category

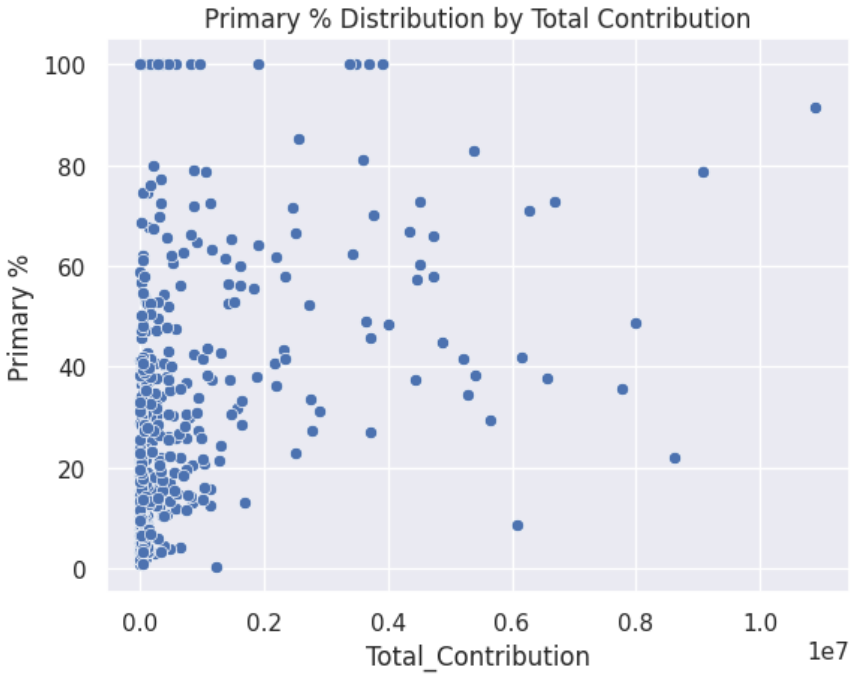


Figure 4: Distribution of Different Types of Contributions by Amount

2. Describe any trends you observe, and any relationships you may want to follow up on.

Because we are trying to predict the primary % from contribution categories, we wanted to see if there was a difference in total contribution amounts between candidates who lost and those who advanced. Looking at the visual, there appears to be a significant difference in the contribution amount between candidates who advanced from the primary election and those who lost, with those who advanced having significantly higher contribution amounts than those who lost. This tells us that there is probably some correlation between contribution amount and winning the primary election. We will explore this relationship further in our models.

Additionally, we compared the total contribution of the candidates to the primary % to see if we could observe any overall trends between total contributions and the proportion of the primary vote attained. Looking at the visual above, there appears to be a high percentage of total contributions that are centered around 0. There primary % attained among this cluster of candidates ranges from 0% to 80%, so there does not appear to be a correlation between these data points. However, when we cross the $200,000 threshold, there appears to be a vague positive correlation between the total amount of contributions and the primary % of the vote attained. This information reveals to us that there might be some correlation to explore between contributions and primary % of the vote.

Data Cleaning

To start off, certain candidates had NaN values in their *Primary %* attribute, resulting in their removal from the dataset. Generally, the financing data was clean, and our attributes of interest had no NaN values.

Our main challenge when looking to create our models was the difference in candidate names between the two datasets.

	Candidate	State	District
0	Anthony White (Alabama)	AL	Governor of Alabama
1	Christopher Countryman	AL	Governor of Alabama
2	Doug "New Blue" Smith	AL	Governor of Alabama

In the first three lines of `dem_candidates`, shown on the left, there are two candidates with labels inserted into their names. Similar issues plagued the financing data, with prefixes just such as “MR.” present in a significant amount of candidate names. In another instance, *Roger Ruggles* was identified as *Ruggles, Roger William Dr.* in our financing data.

To solve the problem of mismatching names, two main techniques were used to ensure that names were accurately mapped and that we would keep as many candidates as possible. String Replacement using Regex in Python allowed us to remove nicknames, prefixes, and suffixes. Then, names were split into First and Last subparts, which were then compared with

each other for a match. Of 801 democratic candidates, we matched 614. That means that about 75 percent of the candidate data we had access to was used. Upon matching all names, a new column was created to merge the datasets in preparation for training the model.

Models

What we are trying to predict is the percentage of votes a Democratic candidate receives in the primary using their financing information. The features we included in our prediction are as follows:

- `Primary %`: outcome variable
- `Individual_Itemized_Contribution`
- `Individual_Itemized_Contribution`
- `Individual_Unitemized_Contribution`
- `Other_Committee_Contribution`
- `Party_Committee_Contribution`
- `Cand_Contribution`
- `Transfer_From_Other_Auth_Committee`

The reason we chose these specific features is because they gave us a holistic view of what types of contributions go into candidate financing, and they are mutually exclusive. For example, we chose to exclude ‘Individual_Contributions’ from our model because it overlapped with ‘Individual_Itemized_Contributions’ and ‘Individual_Unitemized_Contributions’. Using the GLM, we planned to predict the weighted impact each variable had on the percentage of the vote a candidate received in the primary election.

GLMs

To predict the proportion of votes a candidate receives in the primary election, we used a Generalized Linear Model (GLM) to model the effect of different types of contributions on the primary % that a candidate receives. Because we used several variables to model this relationship, we decided to use a Gaussian (Normal) GLM to predict this relationship because we were working with a lot of continuous, unbounded variables. The link function in this model is the identity function, and the likelihood is Gaussian. The assumptions that we are making in using this model are linearity, homoscedasticity (constant variance), normality, and independence.

Frequentist Model

Our first step was to run the frequentist model of our GLM. Below is the output we received from our GLM using the statsmodel.api library, including the coefficients produced by each of our independent variables:

Generalized Linear Model Regression Results

Dep. Variable:	Primary %	No. Observations:	480
Model:	GLM	Df Residuals:	474
Model Family:	Gaussian	Df Model:	5
Link Function:	identity	Scale:	1105.6
Method:	IRLS	Log-Likelihood:	-2360.0
Date:	Tue, 13 Dec 2022	Deviance:	5.2405e+05
Time:	03:50:47	Pearson chi2:	5.24e+05
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Individual_Itemized_Contribution	1.023e-05	4.16e-06	2.459	0.014	2.08e-06	1.84e-05
Individual_Unitemized_Contribution	5.594e-06	3.49e-06	1.603	0.109	-1.25e-06	1.24e-05
Individual_Unitemized_Contribution	5.594e-06	3.49e-06	1.603	0.109	-1.25e-06	1.24e-05
Other_Committee_Contribution	5.728e-05	1.86e-05	3.072	0.002	2.07e-05	9.38e-05
Party_Committee_Contribution	0.0038	0.001	6.032	0.000	0.003	0.005
Cand_Contribution	5.023e-06	5.31e-06	0.947	0.344	-5.38e-06	1.54e-05
Transfer_From_Other_Auth_Committee	-6.109e-05	1.84e-05	-3.314	0.001	-9.72e-05	-2.5e-05

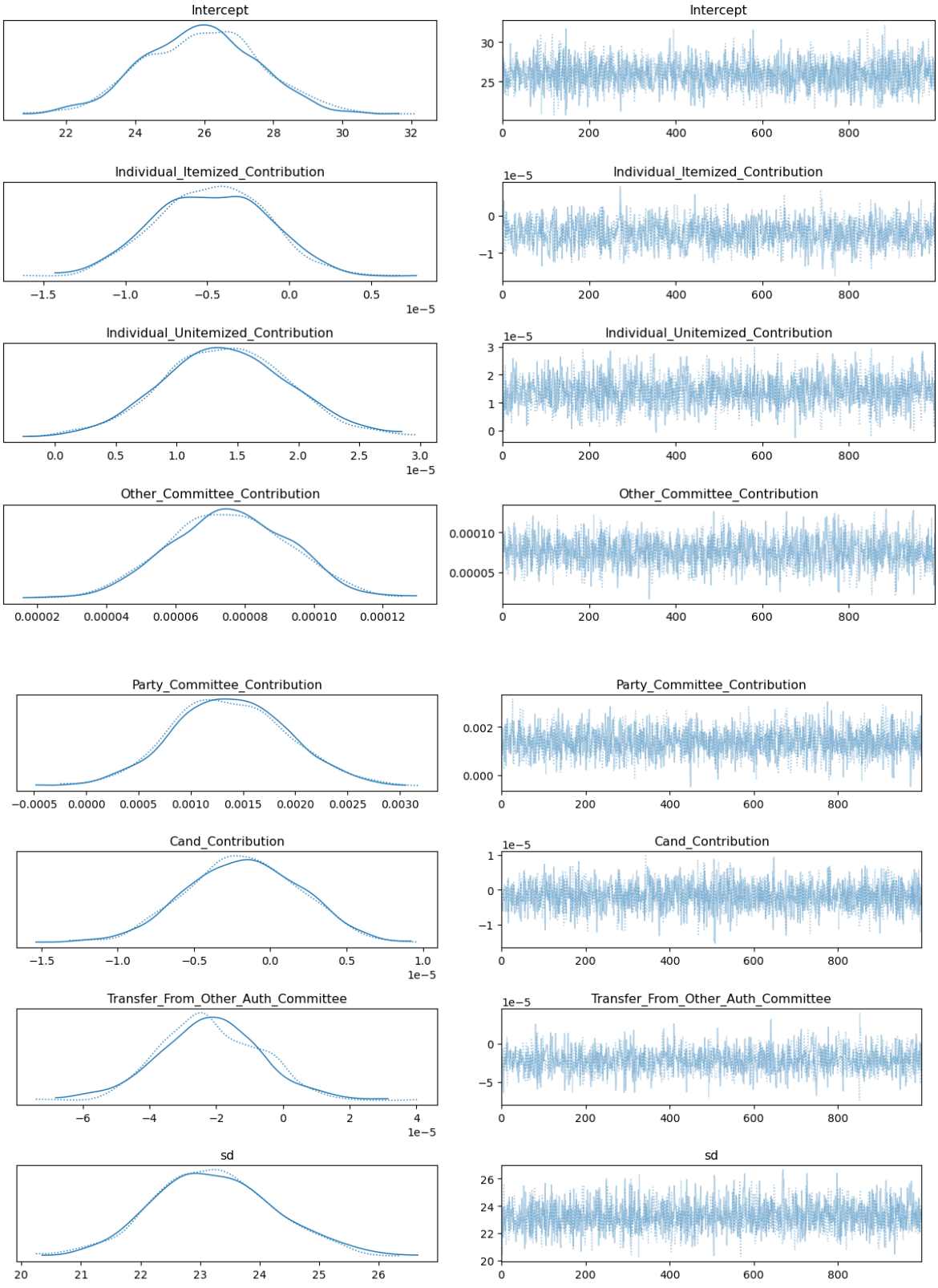
Looking at the coefficients our model produced, we can see that 'Party_Committee_Contribution' produced the highest weighting at 0.0038, and this result is statistically significant; 'Transfer_From_Other_Auth_Committee' produced the second highest weighting at -6.109e-05 (meaning that it reduced Primary % of a candidate), and this result was statistically significant, and 'Other_Committee_Contribution' produced the third highest result at 5.728e-05, and this result was statistically significant.

We then ran the coefficients that we received from this model to predict the Primary % based on the model we produced using the above contribution variables. Our MSE for the difference between predicted and observed outcome values was 1091.775.

Bayesian Model

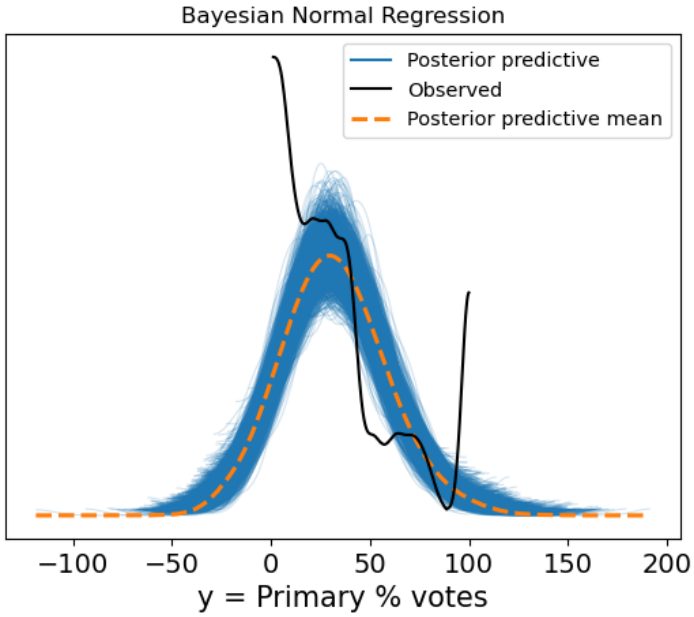
The next model that we decided to use was the Bayesian GLM. For this model, we decided to use a Normal prior, with the parameters, mean=0, and standard deviation=1. As in our Frequentist model, the assumption we made here in assigning a Normal prior was that we are working with several continuous, unbounded variables. So, we decided to assign a Normal prior centered around the mean, with a standard deviation of 1 to test our model.

Below are our results from running our Bayesian GLM. On the left side, we have the histogram of samples for each hidden variable in our model. On the right, we see a distribution of how these variables changed from sample to sample.



Posterior Predictive Checks

Finally, we performed posterior predictive checks on our Bayesian model to validate our model. This would allow us to compare the distribution of our samples produced by our model with the original historical data. The visualization below shows the distribution of PPC samples in blue, outlines the overarching distribution in orange, and compares it to the observed data distribution in black.



As we can see, our Bayesian model does not do a great job at fitting the data. Our observed data does not resemble a Normal distribution whatsoever. This reveals a flaw in our assumptions: we assumed that a Gaussian model would fit our data well because we were modeling several continuous, unbounded variables. Our observed distribution, however, does not follow that of a Gaussian model. Further modeling of this data should be performed to get a representation that mimics the observed values of our dataset.

Nonparametric Methods

To improve upon our previous GLMs, we shifted our focus to two kinds of nonparametric methods: Neural Networks and Decision Trees using the Random Forest Algorithm. Because our prediction of primary % is a prediction task on continuous values from 0-100, we decided that these two methods may be able to provide a high level of accuracy.

Neural Network using scikit-learn

Our first approach to using a nonparametric method involved using a Neural Network through the scikit-learn library. In particular, an MLPClassifier was used. After splitting our dataset into

66 percent training data, our final model was able to test its accuracy using 203 democratic candidate data.

```
▾                                MLPClassifier
MLPClassifier(hidden_layer_sizes=200, max_iter=300, random_state=3)
```

```
1 mean_squared_error(dem_y_pred, dem_y_test)
```

```
562.5741553074827
```

Our mean squared error using this Neural Network is 562.57. This is nearly half of the mean squared error provided using GLMs previously discussed. However, it is worth noting that this is still quite a high error and does not provide confidence in our model's ability to accurately predict the primary % of votes a candidate would receive. Given that primary election results are often very close in terms of the percentage of the total a candidate receives, our model would also not be able to accurately predict whether a candidate would win the primary election or not.

Neural Network using keras

To further tune our neural network, we decided to use keras to add more hidden layers to our neural network as well as fine-tune different parameters with the goal of being able to lower our mean squared error from our previous neural network using scikit-learn.

```
Epoch 68/200
7/7 [==============================] - 0s 2ms/step - loss: 1882.1244
Epoch 69/200
7/7 [==============================] - 0s 2ms/step - loss: 1795.3986
Epoch 70/200
7/7 [==============================] - 0s 2ms/step - loss: 1734.8292
Epoch 71/200
7/7 [==============================] - 0s 2ms/step - loss: 1686.9238
Epoch 72/200
7/7 [==============================] - 0s 3ms/step - loss: 1655.1046
Epoch 73/200
7/7 [==============================] - 0s 3ms/step - loss: 1640.1876
Epoch 74/200
7/7 [==============================] - 0s 3ms/step - loss: 1634.7659
Epoch 75/200
7/7 [==============================] - 0s 3ms/step - loss: 1633.9158
Epoch 76/200
7/7 [==============================] - 0s 2ms/step - loss: 1633.8646
Epoch 77/200
7/7 [==============================] - 0s 3ms/step - loss: 1633.8643
Epoch 78/200
7/7 [==============================] - 0s 4ms/step - loss: 1633.8646
```

However, our mean squared error was not able to improve despite fine-tuning our model.

Random Forest

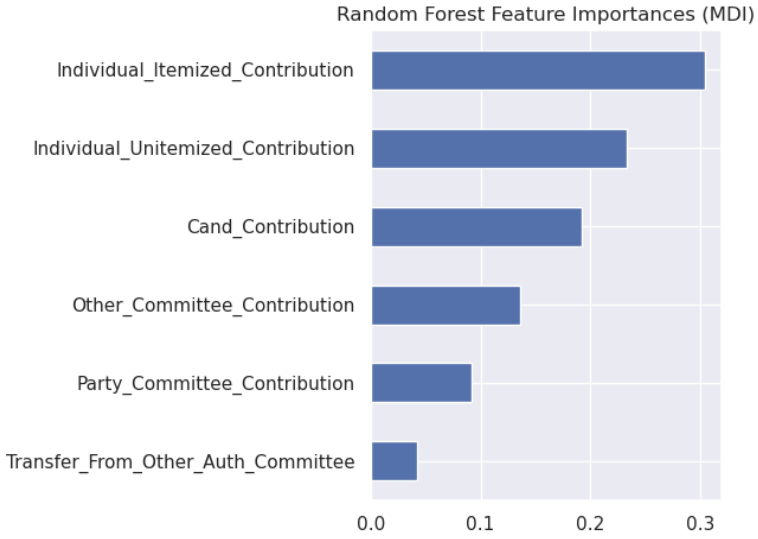
Our final approach involved using a Random Forest consisting of 100 Decision Trees. While we may expect the percentage of votes to be a continuous value that may result in low prediction accuracy for something like a decision tree, it is important to note that a near-perfect prediction

only has a possible set of 100 values. That is, a model that predicts 38% of votes in a primary, in a practical sense, is just as good as one that predicts 38.12% correctly.

```
1 mean_squared_error(dem_rf_y_pred, dem_y_test)
```

```
648.4682375340836
```

However, our results are slightly worse than those of the Neural Network. Part of the reason for a lower prediction accuracy may be related to what was previously discussed.



Another detriment to our predictions may reveal itself by looking at feature importance. Our model only uses 6 attributes, which may be too few for a random forest to correctly predict the percentage of votes and may benefit the Neural Network. As we can see in the above figure, two attributes are about as important as the other 4 attributes.

As with our Neural Network, the financing data is not enough for our models to provide an accurate prediction of the primary percentage of votes.

Results

When looking at the predictions made by all of our previous models, it is clear that none were able to accurately predict the percentage of votes a candidate would receive in a 2018 primary election. Overall, our nonparametric methods were able to provide nearly twice as accurate results when compared to our GLM models. Our neural network produced an MSE of 562.57 and our Random Forest produced an MSE of 648.47, compared to our Frequentist GLM which produced an MSE of 1091.775. We did not find any uncertainty in our GLM predictions.

Despite our nonparametric models performing better than our GLMs, we would not feel comfortable applying these models to future datasets because our error metric was very high,

telling us that our model does not fit our data well. Between our Bayesian and Frequentist GLMs, we noticed that perhaps the Normal distribution was a better fit in the Frequentist model than in the Bayesian model. Looking at the posterior predictive check that we ran, it appears that our prediction did not fit the data whatsoever. While we got a rather large error metric for our Frequentist model, it appears that the Bayesian model performed worse.

Some limitations of our models are that there are other confounding variables that would provide a significant amount of bias when predicting the primary % of the vote obtained. Additionally, the significance of each variable we selected could have been more robust. When we look at the weighting of the variables we selected in each of our models, they did not provide a significant contribution to the primary %, which is why our predictions were probably not accurate.

Our final dataset consisted of around 650 candidates, which hints at enough data for our models to train with. However, as explored in our EDA section and in the previous section, there are too few attributes, with some of them being overemphasized in our models due to their naturally higher values. More financing data, beyond contributions to committees, would prove useful in creating a better model.

Conclusion

Overall, we found that our models were not accurate in predicting the percentage of the primary vote a candidate receives using their financing information. For our GLM models, we found that our Frequentist GLM had a large MSE of 1091.775, and our Bayesian model did not fit the observed data whatsoever. For our nonparametric models, we also saw large errors including an MSE of 562.57 for our Neural Network, and the Random Forest model produced an error of 648.47. Although our errors were large, we can conclude that our nonparametric models outperformed our GLMs.

Because our models performed poorly, our results are not generalizable. If we were to assume that greater financing increases the percentage of the primary vote a candidate receives, then we would propose that the federal government set financing caps on candidates so that candidates with greater financing capabilities do not automatically have an advantage over candidates with less funding. Oftentimes, candidates who receive greater funding have greater personal wealth than candidates with less campaign funding, so this policy would be intended to equalize the playing field.

In terms of the dataset we used, we merged the `dem_candidates.csv` from the *FiveThirtyEight* Github repository with the `candidates.csv` dataset from the *Federal Elections Commision*. In doing so, we were able to aggregate candidate information with their corresponding financing information, which made it possible to model the relationship between our financing variables

and the primary % a candidate received. However, when we merged these datasets, we eliminated many candidates because we were either unable to find an exact match across both datasets, or we were unable to standardize the names across the dataframes.

Some limitations in the data were that there were many null values that made us unable to use accurate information about the candidates to build our models. We filled these blank values with 0, but this could have likely been a misrepresentation of the data. Additionally, the financing information that we used was not entirely clear. It was difficult to intuitively understand whether or not these contributions would actually be impactful on the primary % a candidate receives, so we had to make assumptions about which variables to include in our models.

Future data that may be useful to collect may include the total number of individual contributions, in addition to the total amount of the contributions. This may prove to be a useful metric in predicting the primary % of votes, as it can give insight into how many supporters there are willing to contribute money per individual candidate. While in this project, we only looked at contributions to a candidate, it may be useful to look at campaign spending in addition, which may also have a strong effect. Quantifying that effect, if possible, may also lead to more accurate results in the future, and would be an interesting topic of further study.

Citations

[1] 2018, Five-Thirty-Eight Primary Elections Data,
https://github.com/fivethirtyeight/data/tree/master/primary-candidates-2018

[2] 2018, Federal Elections Commission - United States of America
https://www.fec.gov/data/browse-data/?tab=candidates