Written Report

Introduction

Research Question 1:

How does the number of candidates in a race, as well as money raised and spent on the 2018 primary campaigns affect whether candidates will advance or not? We also want to look at which variables are most predictive.

Research Question 2:

Does having a STEM background cause a Democratic candidate to win the primary?

Data Overview

Data Generation Methods

The data provided was comprehensive and included all primary candidates, therefore the data was a census.

Additional Data Sources

In addition to the Democrats and Republican datasets, we chose to add a financial dataset that includes the summary financial information about each candidate registered with the FEC or appearing on an official state ballot for House, Senate, or President. We wanted to use the variables in the finance dataset as features in our GLM model.

Groups Systematically Excluded From Data

For the second research question, we chose to only focus on the Democrats because some variables that were included in the Democrats dataset were not included in the Republican dataset. For the first research question, we used the entire population from the dataset provided.

Participants' Awareness of Data Collection/Use

In regards to finances, the participants were aware of the data collection. This is because candidates must disclose some campaign finance information as required by law.

Data Granularity

Each row represents one candidate running for office (Representative, Senate, Governor). This makes the interpretability of our findings simple, as we more or less want to understand whether a candidate advances in the primary or not. The granularity of our question matches that of the dataset.

Possible Concerns

– Selection bias

Selection bias is most likely not a concern because our dataset is a census which includes the entire population.

– Measurement error

Our dataset can contain measurement errors. For instance, for the financial dataset, it is possible for there to be human error in inputting accurate and correct numbers into the system. However, this type of measurement error is most likely extremely minimal due to the fact that providing false information, intentional or not, is illegal when it comes to reporting your financial information to the government.

– Convenience sampling

Convenience sampling is not a concern because the entire population is represented in our dataset. None of the individuals in our dataset were selected based on their ease of access or availability.

Differential Privacy Modifications

The dataset was not modified. As explained earlier, candidates are required to disclose such information. This consensual agreement negates the need for differential privacy.

Missing Important Features

For the second research question regarding causal inference, we wanted to use features such as "Partisan Lean" and "Race". These features were not present in the Republican dataset, so we had to focus our research question on the Democratic dataset instead. We believe these features are important to answer our causal inference question because they are likely to be confounders. The Republican dataset also did not have the "STEM?" variable, which is our treatment variable.

Missing Data

There are a few columns with missing data. The "Race" column has a considerable amount of missing rows. Specifically, we found that there were 122 missing values out of 562 data points for the Democratic Party. Based on the dataset's description, "Race" was left blank if they could not identify the candidate's race or ethnicity. To do this, they checked the candidate's website to see if they provided information about their race or ethnicity. If they couldn't find information about their race or ethnicity, they spent no more than two minutes searching for information about it online. For data points with missing values for "Race", we simply excluded them from our analysis because there's no reasonable way we could interpolate the missing values for "Race" without spending a significant amount of time filling them in manually.

The column "STEM?" had a few missing values, specifically 9 missing values out of 562 data points for the Democratic Party. According to the dataset's description, missing values for "STEM?" were due to a lack of a website for the candidate. Since there were only a very few missing values, we excluded them from the dataset as we believe it would not have a significant effect on our causal inference model.

Cleaning/Pre-processing

For the GLM, we removed the outliers based on the “Net_Contribution” column to remove candidates who contributed an incredibly large amount of money. We also added some columns to the data, including one for the number of candidates running in any given race, and changing the “Advanced” column from “Yes” or “No” into a binary 0 or 1 column. We also dropped the rows that had values of NaN in them for the features we were looking at.

For the causal inference model, we turned some of the categorical binary variables ("Self-Funder?", "Race", "STEM?", and "Won Primary") into numerical binary variables of 0's and 1's for the model to properly work.

Research Questions

Research Question 1:

How does the number of candidates in a race, as well as money raised and spent on the 2018 primary campaigns affect whether candidates will advance or not? We also want to look at which variables are most predictive.

The real-world decisions we could make by answering the question is for prospective candidates, what kind of spending would give them the best chance, or for info on their opponents if the spending they’re doing is enough. It would give candidates more info to go by then and is a table that could be updated each election cycle to keep data accurate.

Why Our Method is a Good Fit

We’re using a Generalized Linear Model and nonparametric methods to answer our question. For the GLM, we believe it’s a good fit for our question because GLMs are able to work with multiple predictor variables and we’re able to really control the variables we use to fit the model. The results from a GLM are also really interpretable which is helpful in a case like this where a candidate might want to find out what specifically they can do to increase their odds of winning.

Method Limitations

A limitation of using a GLM is that there are so many features available in our dataset, and so many factors in general that can contribute to voting results that can’t be measured, so it’s just not possible to take everything into account with a GLM and there will always be an underlying error with the model from the factors that aren’t able to be taken into account.

Research Question 2:

Does having a STEM background cause a Democratic candidate to win the primary?

As students majoring in a STEM-related field, we wanted to determine if candidates would have an advantage at winning the primary election if they had a STEM background, given that they were Democratic. If the question at hand were true, it would be a major political advantage to have a STEM background. For instance, political parties might consider prioritizing candidates with a STEM background, and political strategists could advise candidates to emphasize their STEM credentials more in their campaigns. By answering this question, we hope it could provide valuable real-world insights that guide candidates with a STEM background toward winning the primary election.

Why Our Method is a Good Fit

Causal inference is a good fit for this question because we want to determine the cause and effect that the treatment (having a STEM background or not having a STEM background) has on the outcome (winning the primary election or not winning the primary election). By finding causality and accounting for all the confounders, we could determine whether having a STEM background truly causes a candidate to win the primary election, and whether the result of winning or losing the primary election is not just partly due to the correlation between the confounder variables and the outcome variable.

Method Limitations

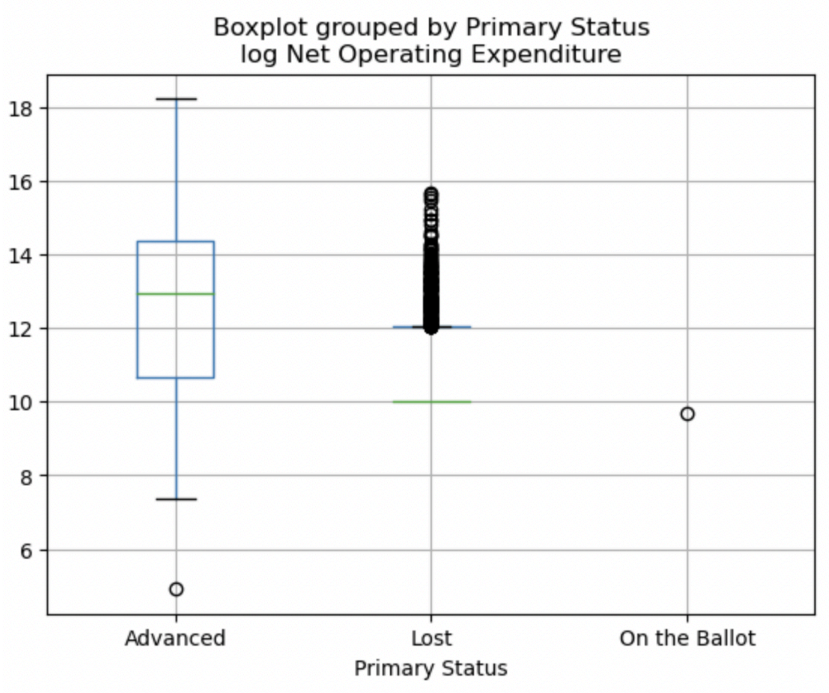
For our method, causal inference heavily relies on the assumption that all confounding variables are accounted for, otherwise our model would not properly work under a false assumption. Here, we picked all the confounding variables that made the most sense to us. It would have been better to mathematically or visually find the correlation between each variable with the treatment variable and the outcome variable to determine if the variable is a confounding variable or not. However, since our combined dataset had 99 columns, it was not feasible to do it mathematically or visually.

EDA

Research Question 1:

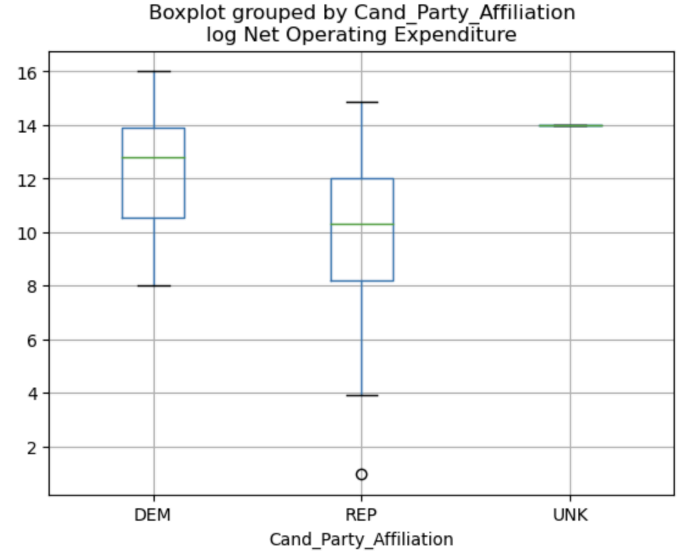
How does money raised and spent on the 2018 primary campaigns affect whether candidates will advance or not? We also want to look at which variables are most predictive.

(Qualitative)

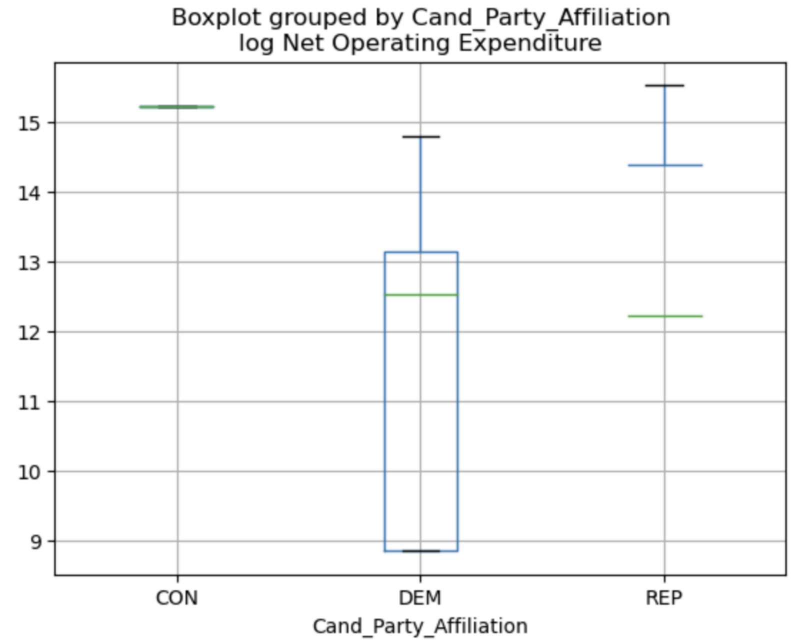


Box plot of Log Net Operating Expenditure broken down by Primary Status (Advanced and Lost). We see that those who advanced had a wider range of expenditure, and was generally higher than those who lost. It seems that those who lost stayed in a similar range (but with many exceptions). We will investigate these effects further by looking into different states and breaking them down by party.

California

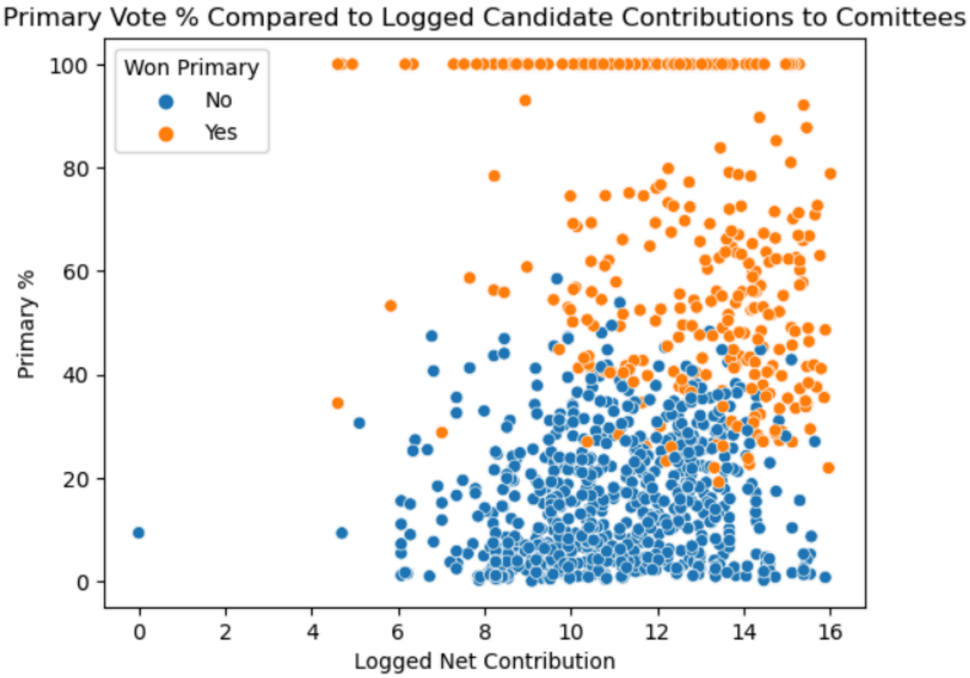


West Virginia

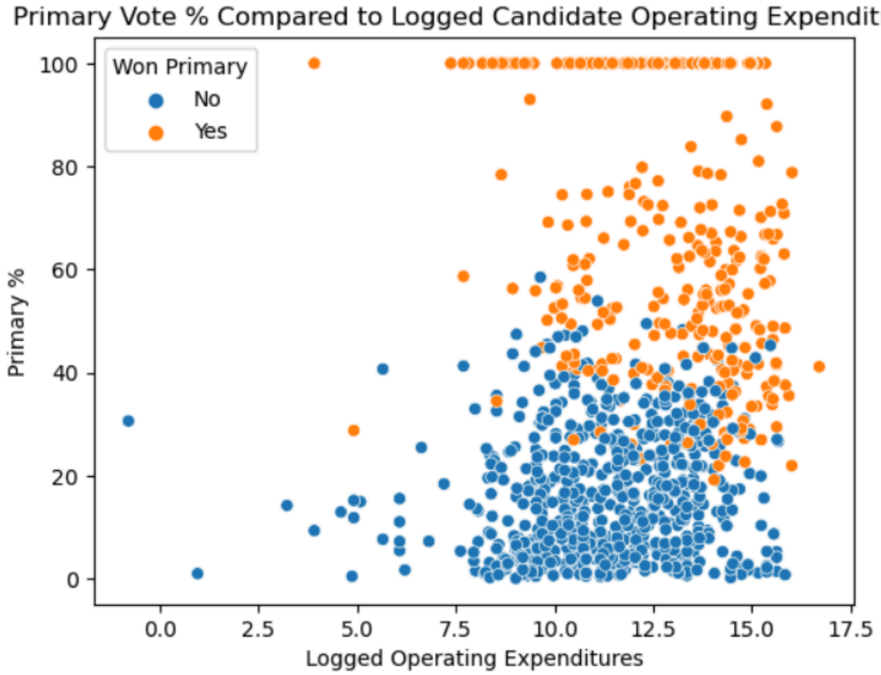


Looking at California, we see that Democrats spent more than Republicans. In West Virginia on the other hand, Democrats spent less than Republicans. This indicates that in our model, we should potentially add Party Affiliation as well as State into our model.

(Quantitative)



When we switch to looking at candidate contributions, we can see that the contributions in the highest range, from 14–16, have candidates almost exclusively winning their primary if they contribute that much to committees, party affiliated or not. With so many points on the lower end despite spending, it’s possible that there were just so many candidates in the primaries that the vote got massively split amongst those with a really low % of votes received.

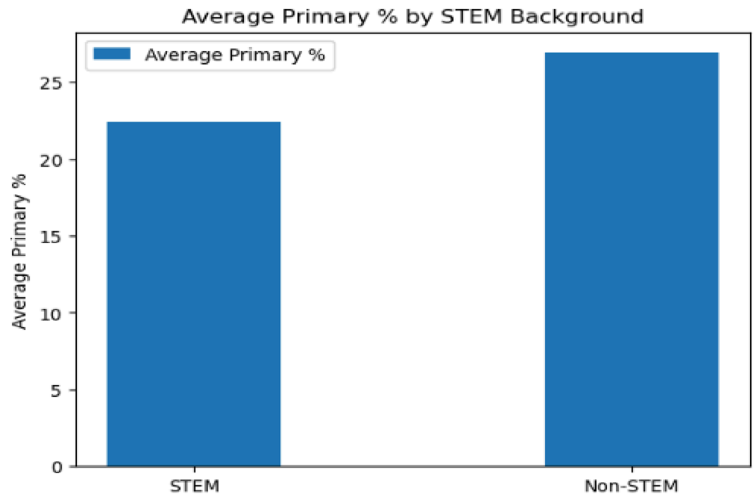


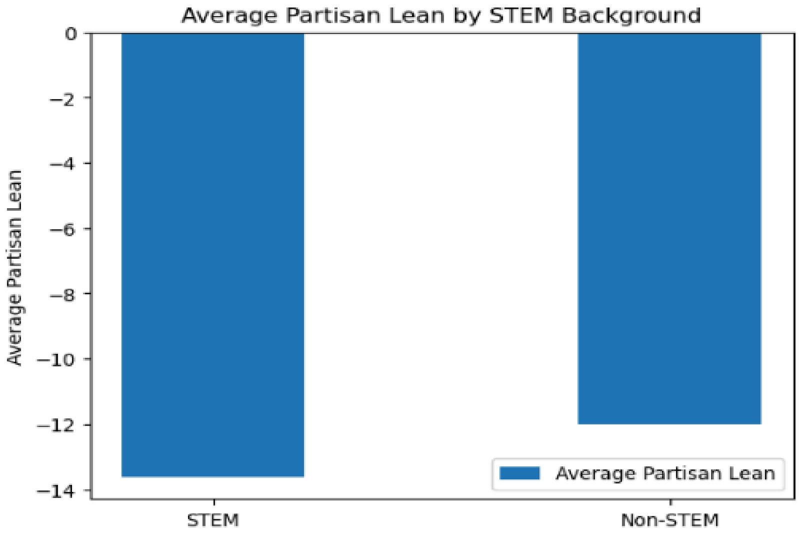
Similarly, when looking at operating expenses for each candidate, we see even more of a bunch up in the 7.5–15 range, and a resoundingly high chance of winning above that. The fact that there are so many candidates bunched up with such high expenditures while still receiving such a small percentage of the votes might indicate that there’s somewhat of a floor to what a candidate needs to spend to even be able to stand a chance of winning, a luxury that many ambitious grassroots candidates may not be able to afford.

Research Question 2:

Does having a STEM background cause a candidate to win the primary?

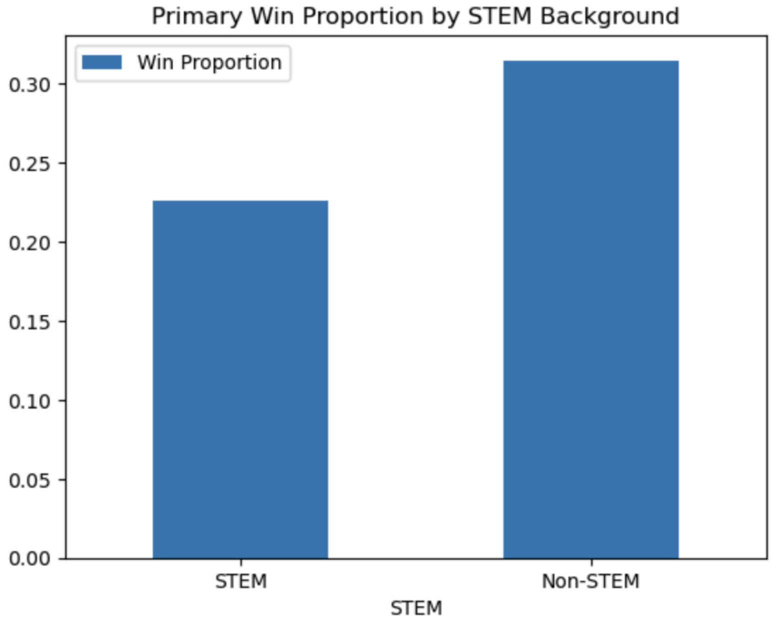
(Quantitative)

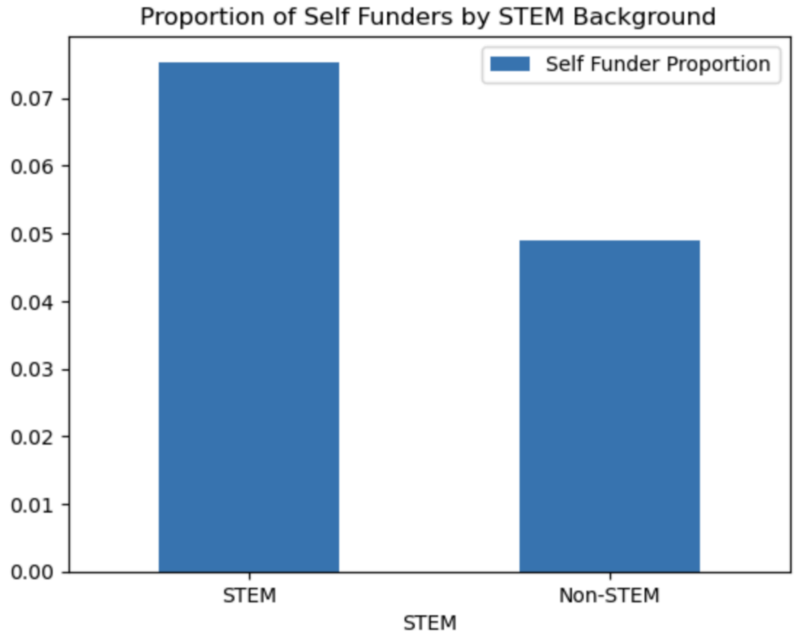




Based on the two graphs above, we can see that the average primary % is almost 5% lower and the average partisan lean is almost 2 lower for having a STEM background. This suggests that on average, having a STEM background may put a candidate at a slight disadvantage when it comes to winning the election. However, this is not a strong indication due to the disproportionate balance of data points between STEM and non-STEM. The % of STEM Democratic candidates is much lower, with only 18% of the data points being STEM and 82% being non-STEM (there are only 138 data points for STEM candidates compared to 619). Some possible reasons for the disproportionate balance of data points between STEM and non-STEM: people with a STEM background are less likely to be eligible to be a candidate for the primary election since there are way more non-STEM candidates than STEM candidates, people with a non-STEM background are more likely to sign up to be eligible to be a candidate than people with a STEM background (for whatever reason(s)). So having a STEM background might not be a strong factor in a candidate's likelihood of winning the primary election.

(Qualitative)





Based on the first graph looking at the primary win proportion against STEM backgrounds, we can see that there is a clear difference in the proportion of wins for those with a STEM background and for those without. Specifically, those without a STEM background have a proportion of wins that is 0.09 more than those with. While this may seem like a significant difference, we should further our research and attempt to account for the difference in data between the two groups, as well as look into potential confounders.

In particular, the second graph looks at one of the potential confounders we identified, the self-funding of the candidate. In this graphic, we can see that those with a STEM background do have a proportion about 0.025 higher than those without a STEM background. This difference doesn't seem to be particularly different, so we may want to consider some other possible confounders as well.

Prediction with GLMs and Nonparametric Methods

Methods and Prediction Goals (Question 1)

We are trying to predict how much of a percent of the primary vote would a candidate receive based on our chosen features. Our features are:

- **Logged Net Contribution** - A candidate's contributions to their campaign can be a good sign of their campaign doing well or receiving a lot of support, it should absolutely be included as a feature
- **Logged Operating Expenditures** - Similarly, a candidate's expenditures can show that they may be more willing to keep their candidacy going and may mean they're able to last longer in the race and garner more votes
- **Logged Total Loan** - If a candidate is taking out loans, they may not be receiving a lot of support from voters in their district, and so they may have to resort to loans in order to keep up their presence in the race
- **Number of Candidates in Race** - It's possible that the more candidates there are in a race, the less vote any other given candidate will receive, so this feature is also worth looking into

GLM Description

We primarily utilized logistic regression for our models because we are predicting a binary variable (Advanced or Not advanced in the primaries). Logistic regression returns values between 0 and 1 corresponding to the predicted probability of belonging to the "Advanced" category.

The prior one we chose to use for our Bayesian model was the default one provided by bambi. This is due to our lack of domain knowledge on election trends and distributions.

The primary assumption with our GLM is that the variables we are investigating are indeed linearly related.

Nonparametric Methods Description

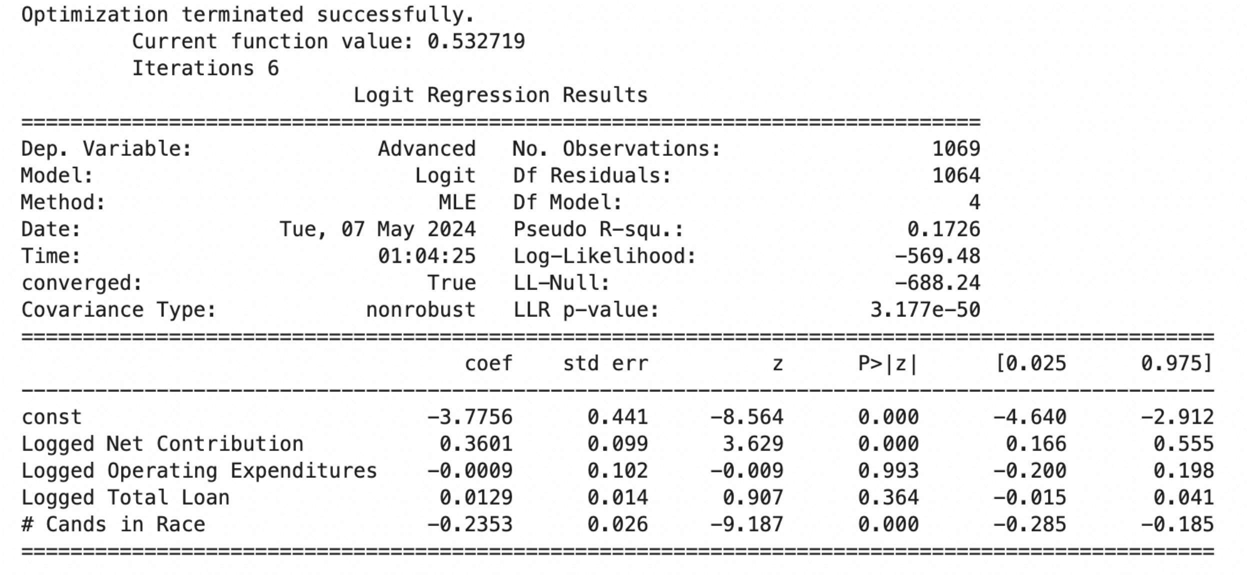
We chose to use a random forest because we wanted to look at variables that might not have been linearly related. This makes a non-parametric method ideal.

Model Performance Evaluation

We generated AIC scores for multiple frequentist GLM models and compared them to pick the best one. For the random forest, we primarily looked at test accuracy.

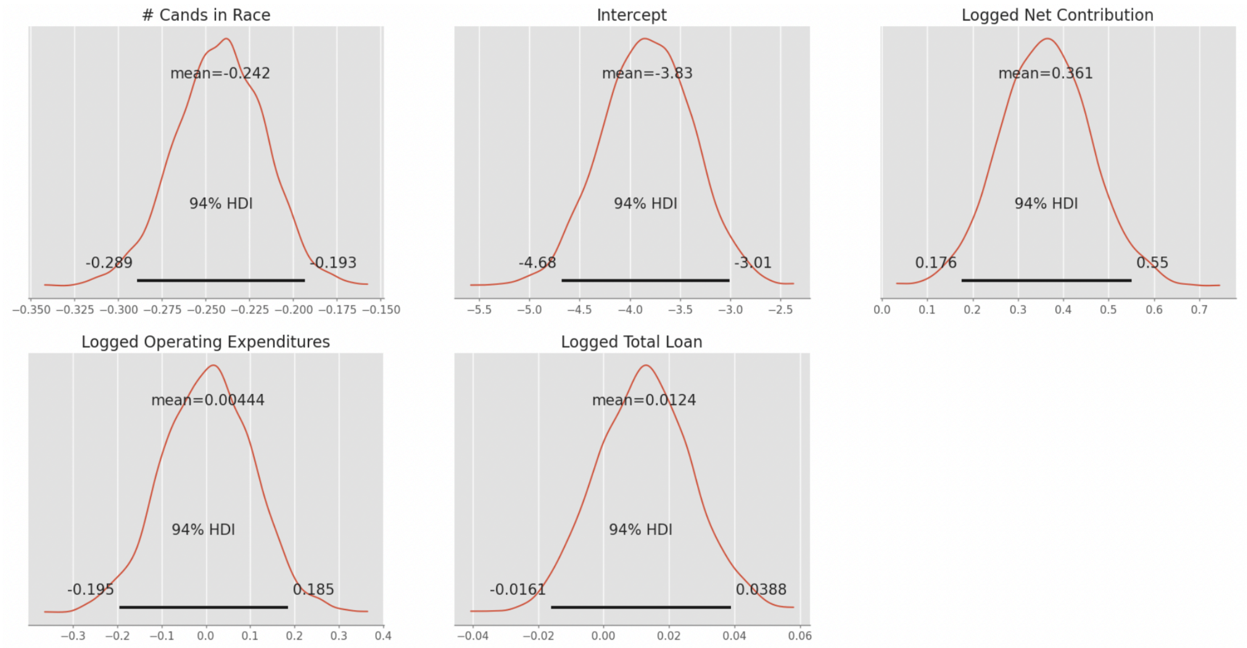
Results (Question 1)

Frequentist Model Results



```
Optimization terminated successfully.
         Current function value: 0.532719
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:               Advanced   No. Observations:                 1069
Model:                          Logit   Df Residuals:                     1064
Method:                           MLE   Df Model:                            4
Date:                Tue, 07 May 2024   Pseudo R-squ.:                  0.1726
Time:                        01:04:25   Log-Likelihood:                -569.48
converged:                       True   LL-Null:                       -688.24
Covariance Type:            nonrobust   LLR p-value:                 3.177e-50
=================================================================================================
                                    coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                            -3.7756      0.441     -8.564      0.000      -4.640      -2.912
Logged Net Contribution           0.3601      0.099      3.629      0.000       0.166       0.555
Logged Operating Expenditures    -0.0009      0.102     -0.009      0.993      -0.200       0.198
Logged Total Loan                 0.0129      0.014      0.907      0.364      -0.015       0.041
# Cands in Race                  -0.2353      0.026     -9.187      0.000      -0.285      -0.185
=================================================================================================
```

Bayesian Model Results



The frequentist and the Bayesian model results mirror each other. In both cases, Operating Expenditure and Total Loan are not predictive features. We can tell this by noting the HDIs, and the high z-scores.

Thus, we created a GLM that got rid of those two features. The resulting model was accurate 72% of the time in comparing the predictions to the truth.

Random Forest Results

We created two random forests: one with the same features as the GLM, and the other with more categorical features (Political Party and State). Both random forests achieved a test accuracy of around 74%.

Uncertainty Estimations

We measured uncertainty using accuracy. Given the accuracy of around 75%, we expect our model to predict wrong about 25% of the time. This holds for all models.

Discussion (Question 1)

Our random forest performed better, but only slightly. We think this is because we are modeling relationships between variables that are mostly linear. Going forward, we suggest using the GLM for future datasets, as it is more interpretable. This includes applications on future datasets of this nature.

Given the nature of the data, this model could be applied to future datasets, as they will contain similar information about future candidates.

Model Fit

We examined two frequentist logistic regression models: one containing information about operating expenditures and loans, and the other without that information. We found that the second model had a lower AIC, meaning that it was able to achieve similar results with less information.

For both the GLM and the random forest, we achieved an accuracy of around 75%

Results Interpretation

Expanding upon the previous question, we found that two of the most predictive features for whether a candidate advances in the primary were the number of candidates in their race and the net contributions to the campaign. Operating expenditures and loans provided proved to not be useful information.

Model Limitations

Some limitations both models have is that there are so many features that we just can’t feasibly include them all, and many of them may not even be worth including at all either. Not to mention, there’s a lot that goes into voting for a candidate that isn’t in our features, like likability, history, etc., and it’s not possible to quantify everything and make a model that will be highly accurate without having those features.

For the random forest, a limitation of the model is that it’s not very interpretable. We have the results from the validation set and their accuracy, but because of the nature of random forests, it’s difficult to interpret how each feature contributed compared to the interpretability of GLMs.

Potentially Useful Additional Data

Additional data that would be useful for improving our model would be poll data. Poll data would be the closest we can get to a “likeability” / “population” feature and it would help gauge how a candidate is doing in the current climate, making it an especially useful feature for predicting.

Results Uncertainty

Given the accuracy of around 75%, there seems to be a good amount of uncertainty. This is due to the limited information we worked with (two variables proved to be predictive). It’s possible that with more variables (unrelated to finances), the uncertainty would decrease.

Causal Inference

Methods (Question 2)

Treatment/Outcome Variables

- **Treatment:** STEM? (Yes / No)
  - If the candidate stated they came from a STEM background.
- **Outcome:** Won Primary (Yes / No)
  - If the candidate won the primary in their district.

Confounding Variables

Our confounders are:

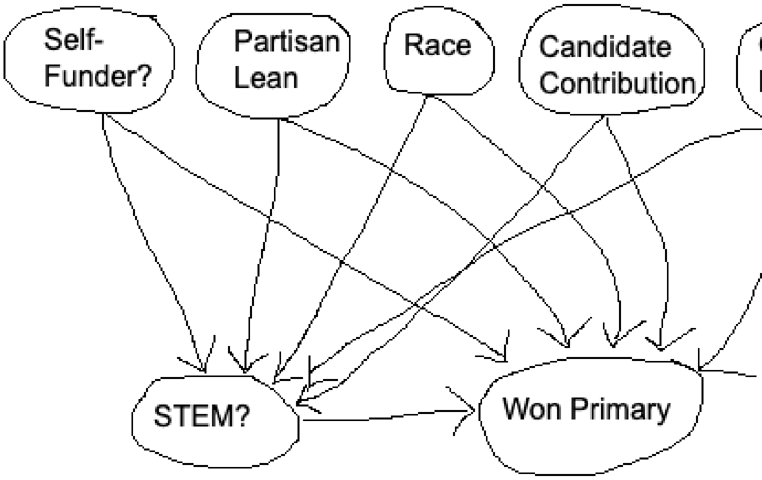
- **Self-Funder?** - Self-funder affects whether you'll have a STEM background or not because you're more likely to have a STEM background if you're self-funded. Self-funder also affects whether a candidate wins the primary in their district because a highly funded election is positively correlated with a higher chance of winning.
- **Race (White/Nonwhite)** - There seems to be a positive correlation between being white and having a STEM background. There also seems to be a correlation between being white and being elected into government. This could be a reflection of racial biases in STEM fields as well as in politics.
- **Partisan Lean** - Depending on the partisan background of the district they come from, this may affect their educational backgrounds and if they went into STEM. Also, partisan lean will most likely affect a candidate's chances of winning in their specific district.
- **Candidate Contribution** - This is the amount of personal wealth a candidate puts towards their campaign. This can affect their chance of winning because higher contributions can lead to better campaigning. This can also affect their STEM background, as people with higher wealth may be more likely to have access to STEM programs or education.
- **Candidate Loan** - This is the amount of loans that the candidate gave to the campaign. This can affect primary wins, as more loans can lead to a better campaign. This can also

reflect STEM backgrounds, as people with more money to loan can have more money towards STEM education.

Adjusting for Confounders
We will use a propensity score model to adjust for confounding variables.

Colliders
There are no colliders in the dataset.

Causal DAG



Results (Question 2)

According to the calculations we've made, having a STEM background has an average treatment effect of -0.31. Assuming that we were able to account for all potential confounders of having a STEM background and winning the primary, this should be a statistically significant estimate. Moreover, utilizing a propensity model, we can assume further accuracy of the different effects of each confounder. In terms of magnitude, -0.31 is a fairly large effect. This can be interpreted as having a 31% smaller chance of winning for candidates having a STEM background. For candidates, this can easily be the difference between winning and losing.

Estimate Uncertainty
There are other possible confounding variables that are not in the data set. Another possible source of uncertainty in our estimate can come from us only looking at Democratic candidates.

The Republican data did not have variables like “STEM?” and the other categorical variables we were looking at. This means our estimates can only be applied to Democratic candidates and not be representative of all candidates.

Discussion (Question 2)

Method Limitations

One of the potential limitations of our methods is that there is the possibility of other confounders not in our dataset. With the data we have, we were able to pick the best confounders we could to account for any confoundedness in our outcomes and treatment. However, it is possible that there are other confounders that we are unaware of or don’t have access to. This means that our estimates can be slightly off if there are other variables that we are not controlling. Another limitation of our methods is that there are some missing values in our data. In other words, some candidates did not have data on the specific variables we used in our models, which meant they had to be excluded. This can be a potential problem if there is a reason why specific values or candidates have missing values, which can ultimately affect the accuracy of the population we are looking at. A final limitation of our model is that there is a difference in the size of the treatment and control. In particular, it seems that only 20% of the candidates in the census have a stem background. This meant our treatment was of size 94 and our control was 341. This can be a problem because our treatment is not equally accounted for in the model, which can take away from its accuracy.

Potentially Useful Additional Data

It would be useful to have more data on Republican candidates. As stated before, the Republican data did not have categorical variables like “STEM?” or the others we used as confounders. If we had access to that data for Republicans we could expand our causal inference to not only Democrats, but all candidates.

Causal Relationship Confidence

We are fairly confident that there is a causal relationship between STEM backgrounds and primary wins because we utilized a propensity model to account for confounders. In addition, while our control and treatment groups are different sizes, they are both large enough to draw significant conclusions from them. Therefore, with the large ETA of -0.31 we calculated, we can be confident that there is a causal relationship even with any of the slight limitations in our methods.

Conclusions

Key Findings

- (q1) We found that the amount of money a candidate spends on their campaign increases how many votes they receive, and that the more candidates in a race, the fewer votes the candidates receive

- (q2) We found that having a STEM background has an average treatment effect on winning primaries of -0.31. This means that candidates with a STEM background have a 31% less chance of winning a primary than those who don't.

Results Generalizability

- (q1) The results are relatively generalizable, as the structure of elections and financial data are similar across years.
- (q2) The results for the second question are not generalizable for all candidates, because STEM backgrounds for Republicans were not available. This means that the causal effect of STEM backgrounds on winning primaries can only be generalized to Democratic candidates.

Call to Action

- (q1) Contributions were a major predictive factor in whether a candidate won or not. While there is a cap on the size of contributions, there is not a cap on campaign spending. This means that those who are backed by ultra-rich donors are more likely to win. Our call to action is to level the playing field by putting a cap on campaign spending. This could help in protecting public interests from the ultra-rich.
- (q2) For people who are looking for a career in politics or running for government, they should not focus on having a STEM background. It is possible that those without a STEM background might have more focus on legal studies or policy which is more appealing for voters.

Merging of Data Sources

- (q1 & q2) Yes, we combined datasets for candidate backgrounds with candidate finances. One of the drawbacks was that some candidates were not on one data set or the other. This meant that some rows were excluded, decreasing the size and accuracy of our data. A benefit was that we were able to consider other important variables in our models.

Data Limitations

- (q1) We limited ourselves to looking at financial data and seeing which ones were most predictive from that. If we truly wanted to predict races, we could add more variables that are not related to finance.
- (q2) The dataset realistically has confounding variables that researchers did not account for since this dataset was not used to find causality between variables, let alone causality between specifically "STEM?" and "Won Primary". As a result, this could lead to biased effects of casual estimates.

Future Studies

- (q1) A future study that could build off our work is we could also look at future trends, seeing how much campaign money contributions change over election cycles and how they affect voting.
- (q2) Looking into backgrounds different from STEM like legal backgrounds, political science, and economics, to name a few.

What We Learned During This Project

One of the main skills gained through this project was practice reading and interpreting APIs and documentation further than what we learned in university classes. Those classes had a lot of the EDA and cleaning already done for us before we even got to see the datasets, so this was a good experience in exploring and cleaning data from scratch and working with it to get the features that best fit our goals. While university classes provided a good introduction to using Python packages such as bambi and statsmodels, we ultimately needed to explore more of the packages on our own in order to answer the questions, and through this, we were able to reinforce our understanding of the topics we learned in university.