Data 102 Final Project

Alex Lopitz, Mickey Piekarski, Shiyao Lu, and Tyler Nadig

10 May 2021

1 Data Overview

For our final project, we decided to analyze two datasets, one from FiveThirtyEight [Meh18] and the other from the Federal Election Commission (FEC) [FEC18].

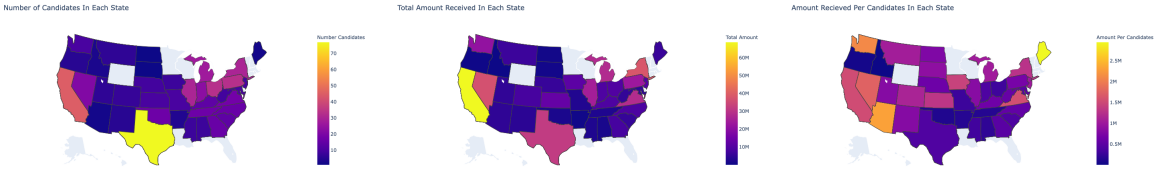
The FiveThirtyEight dataset compiled information on the 811 candidates who appeared on the ballot in the 2018 Democratic primaries for Senate, House of Representatives, and governor, not counting races featuring a Democratic incumbent. This dataset is a census as it includes every candidate in our target population of 2018 Democratic primary candidates. While there are certainly barriers that make it difficult for certain groups to enter politics, the dataset does not exclude any particular group of candidates once they succeeded in getting their name on a primary ballot. Note that the exclusion of candidates from primaries including incumbents actually benefits our analysis as we've removed any incumbency effects from our analysis. Of course, this limits our ability to make conclusions about primary elections that do include incumbents. All candidates included in this dataset should be well-aware of the collection and use of this data as candidates for public office know their information will be made available to the public. Additionally, this dataset is very granular as each row represents one candidate in our target population with features including district/state's partisan lean, number of endorsements, race, veteran status, and election outcome. This high granularity should increase confidence in our ultimate findings. As this dataset is a complete census, there's little reason to concern about selection bias, measurement error, or convenience sampling. It would be nice if the data included demographic information on Republican primary candidates so we could include them in our analysis. Even so, this dataset will allow us to address important questions regarding Democratic election primaries.

The dataset from the FEC compiled campaign financing information on federal candidates in both primary and general elections. So, unlike the FiveThirtyEight dataset, it does not include governors. Yet, similar to the FiveThirtyEight dataset, it's a census as it includes every candidate who's run for federal office. It also should not systematically exclude any group of candidates as long as they succeeded in getting their name on a ballot for federal office. Again, candidates for public office know they are required to make certain funding and spending information public, so the collection and use of this data should not surprise anyone. This data is highly granular as each row represents a candidate with features including total campaign receipts, individual contributions, party contributions, and more. As with the the FiveThirtyEight dataset, the high granularity will increase confidence in our findings. As noted previously, with census data there is little reason to worry about selection bias, measurement error, or convenience sampling as everyone in the target population should be included in the dataset. Unfortunately, there is a significant amount of dark money in politics and not all political spending comes directly from a candidate's campaign. If we had features on outside group spending, this would give us a more complete analysis of the impact of spending on election outcomes.

2 EDA

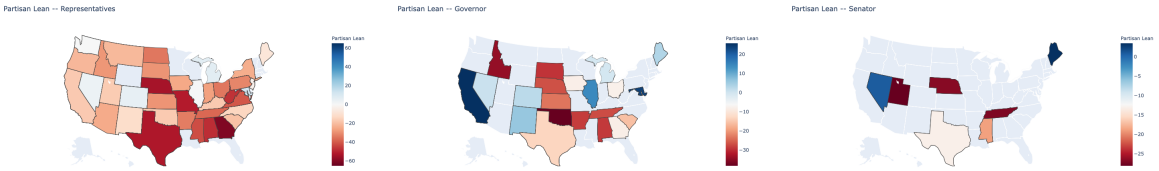
2.1 Visualizations

To visualize how the number of candidates and funding is distributed across states, we started by grouping total number of candidates, total funding, and average candidate funding by state.



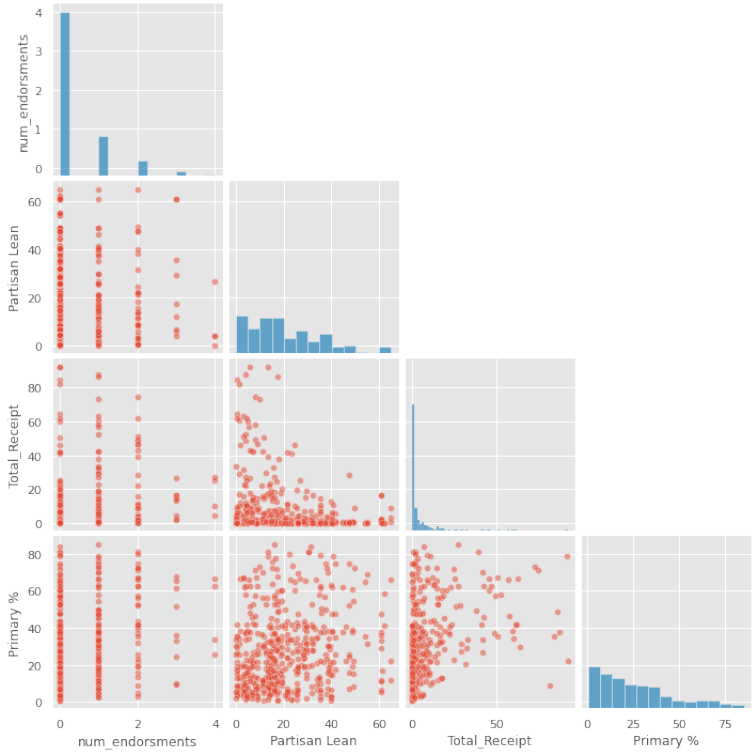
From the above three graphs, we observe that our datasets are missing some values regarding a few states. This likely stems from not every state having elections in 2018 and our dataset not including candidates from primaries with incumbents. In the data we're using, Texas has the most candidates, California has the most funding, and Maine has the most funding per candidate.

There are three office types in the FiveThirtyEight dataset, and each of them have different values of average 'Partisan Lean' per state. Therefore, we grouped by office type to visualize how partisan lean differs across our data.



From the above three figures, we observe that our data does not have much information about Governor and Senate primary races. Therefore, we'll only use Representative data when addressing partisan lean in our our causal inference question.

This following visualization shows the relationships between our treatment, outcome, instrumental variable, and confounding variable for our causal inference question.



SNS Pairplot

2.2 Data Cleaning

The data we chose to use was, for the most part, in a fairly clean state when loaded. Nevertheless, a bit of manipulation was needed in order to get the datasets to fit our needs.

2.2.1 Binary Encode Columns

The columns: 'Veteran?', 'LGBTQ?', 'Elected Official?', 'Self-Funder?', 'STEM?', 'Obama Alum?', 'Party Support?', 'Emily?', 'Guns 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?,' all needed to be encoded as binary variables (0 or 1) from their initial string state ("No" or "Yes") to fit our needs. The 'Race' column was encoded as 1 for 'White', and 0 for 'Nonwhite'.

Lastly, the column 'Primary Status' was changed to a boolean with 'True' for Advanced, and 'False' for Lost, in reference to if a candidate moved on from the primary.

2.2.2 Adding First and Last Name Columns to Both Datasets

Both datasets contained differently formatted columns with candidate names. In order to create consistency across both datasets, regular expression pattern extraction techniques were used to create 'candidate first name' and 'candidate last name' columns, cutting out middle names, suffixes, and other irregularities.

2.2.3 Calculating Total Number of Endorsements

We then decided to use the total number of endorsements a candidate received as a feature in our models and thus added a column summing the endorsement indicator columns encoded earlier.

2.2.4 Calculating Number of Candidates in Each Respective Primary

Another calculated feature we created was the number of candidates in each respective primary race. To add this feature, a group by function was used on the candidates' respective district (The 'District' column) and then the sum of each group was added as a feature for each group member.

2.2.5 Inner Join on FiveThirtyEight and FEC Data

In order to create a clean dataframe to run our models on with all the features we wanted to included, we decided to join the two datasets on candidate first name, last name, and state, in order to ensure the proper rows were joined since there was no valid relational key between the two sets. We then manually reviewed the data to ensure a proper join since the number of rows was small (¡1000). In the process of the join we lost the candidates running for governor as the financial FEC data did not include those candidates.

2.2.6 Filtering Outliers

The final process we performed on our joined dataframe was filtering out candidates that received 100 percent of their primary vote. This is because these candidates were involved in uncontested races and we felt they were irrelevant outliers to the models we were attempting to build.

3 Research Questions

3.1 Research Question 1

For our first question, we're interested in the causal effect of campaign funding on a Democratic candidate's 2018 primary vote share. By answering this question, future candidates in Democratic primaries can better allocate their time and resources as the political environment isn't likely to change drastically from 2018 to 2022. For example, if funding doesn't increase a candidate's chance of winning, then candidates would be better off campaigning instead of going to fundraisers. Our research question is a perfect fit for causal inference because we're interested in the ability of one specific feature, money raised, to increase vote share.

3.2 Research Question 2

For our second question, we want to predict a candidate's share of the primary vote based on their personal demographics, endorsements, and funding. A 'candidate profile' if you will. We will approach this by fitting a Gaussian GLM with the identity link function and a random forest. By answering this question, we gain a better understanding of what characteristics benefit a candidate running in a Democratic primary. Our results can also demonstrate how certain features systematically benefit specific candidates over others. Our research question is a good fit for prediction with GLMs and nonparametric methods, as we have data on multiple features including personal demographics, endorsements, and funding, and we're interested in how those features contribute to election results. While we can't prove causality with our GLM and nonparametric models, we can analyze what features are statistically significant and contribute most to election outcomes.

4 Research Question 1: Causal Inference

4.1 Methods

For our first question, we're interested in causal impact of funding on the primary percentage of a campaign. Thus our treatment variable is 'Total Receipt,' i.e. money received by a campaign, and our outcome variable is 'Primary %,' i.e. percent of the primary vote a candidate receives. Unfortunately, there are many potential confounders that make this analysis difficult. One potential confounder is the variable 'Partisan Lean,' i.e. the average difference between how a state or district voted in the past two presidential elections and how the country voted overall. More partisan districts might have more competitive primaries (because the winner of the primary likely wins the general) and thus there might be more candidates running which could decrease vote share. It's also likely that there are confounders not included in our dataset. For example, socioeconomic factors like a state or district's income per capita can impact a candidate's ability to both raise money and increase their name recognition, making it difficult to raise their primary vote share. Therefore the unconfoundedness assumption likely doesn't hold as these socioeconomic confounders aren't included in our dataset, and thus we can't account for them in our model. To adjust for this dilemma, we decided to use the variable 'num_endorsements,' i.e. a candidate's total number of endorsements, as an instrumental variable. So we're assuming that 'num_endorsements' is independent of the confounder 'Partisan Lean,' which appears reasonable based on the SNS Pairplot in the EDA section, and affects 'Primary %' only through 'Total Receipt.'

4.2 Results

First, we conducted OLS regression with 'Total_Receipt' as our response variable and 'Partisan Lean' and 'num_endorsments' as our regressors. This was meant to analyze the impact of our confounder and IV on our treatment variable. We found that as 'Partisan Lean' increases by 1, 'Total Receipt' increases by $5,000, and for every additional endorsement a candidate receives, their 'Total Receipt' increases by about $767,000. In reality, the effect of most confounders cannot be measured, but using 'Partisan Lean' as our confounder allows us to analyze the validity of 'num_endorsements' as an instrumental variable. Clearly both our confounder and IV impact our treatment variable which is promising, but it's worth noting the impact of our confounder is significant only if we use a p value of 0.10, not 0.05. Still, with evidence that both our IV and confounder impact our treatment variable, we ran OLS regression on 'Primary %' with 'Total_Receipt' and 'Partisan Lean' as our regressors. This was meant to establish the "true" impact of money raised on primary vote share when we account for 'Partisan Lean' as a confounder (True in parentheses given there are likely more unaccounted for confounders). We found that both an additional $100k raised and a 'Partisan Lean' increase of 1 led to a 0.84% increase in primary vote share. In other words, if we control for the confounder 'Partisan Lean,' each $100k raised increases a candidate's primary vote share by 0.84%. When we dropped 'Partisan Lean' and conducted a naive OLS regression on vote share with only our treatment variable 'Total Receipt' and an intercept term, we found that for every $100k raised, a candidate's primary vote share increases by 0.51%. In other words, when we fail to account for the potential confounding variable 'Partisan Lean,' we underestimate the impact of 'Total Receipt' by about 0.33%.

With an understanding of our confounders affect on our treatment and outcome variables, we then ran OLS regression on our treatment variable 'Total Receipt' with an intercept term and our IV 'num_endorsements' as the regressor. This resulted in predicted values for 'Total Receipt,' which we then regressed on primary vote percentage. For every increase of $100k in our predicted 'Total Receipt' variable, we found a 1.59% increase in vote share. So using our IV, we would actually overestimate the impact of an additional $100k on primary vote share by 0.75%. Due to this discrepancy, we're hesitant to conclude that each $100k raised causes a 1.59% increase in primary vote share as the IV 'num_endorsements' doesn't seem to accurately control for the confounder 'Partisan Lean.' Thus it's difficult to conclude that using 'num_endorsements' as an IV would also control for potential confounders like socioeconomic factors that we couldn't include in our model.

4.3 Discussion

One major limitation in the design of our causal inference question is that our IV 'num_endorsements' doesn't have very high variance, and thus it's difficult to use 'num_endorsements' to acutely predict our treatment variable, 'Total Receipt.' Additionally, every OLS regression we conducted assumes a linear relationship between our regressors and the response variable, which likely doesn't hold in reality. It wouldn't be surprising if the impact of 'Total Receipt' on 'Primary %' decreases after a certain amount of money is raised by a candidate. It's also possible that our IV 'num_endorsements' impacts primary vote share directly rather than through our treatment variable. For example, Clyburn's endorsement of Biden in the 2020 Democratic Primary seems to have helped a campaign that wasn't bringing in anywhere close to as much money as its competitors. Thus if we tried to answer this question again with additional data, a more random IV that only impacts our outcome variable through our treatment variable would be extremely useful. Perhaps 'Partisan Lean' could be that IV. I would also be interested in further splitting our data set into House and Senate races as Senate races are likely more expensive than most House races. For all of the reasons listed above, we're skeptical in concluding there's an exact causal relationship of funding on primary vote percentage.

5 Research Question 2: Prediction with GLMs and non-parametric methods

5.1 Methods

For our second question, we're interested in predicting a Democratic candidate's primary vote share using the following binary categorical variables: Race (white or non-white), Veteran status, LGBTQ, Elected Official, Self Funder, and whether or not the candidate has a background in STEM. We also used the following quantitative variables: number of endorsements, money raised, and number of candidates running in a primary. We selected these variables assuming that more endorsements, more money raised, and fewer opponents likely correlates with higher shares of the primary vote (It's worth noting that even though we failed to determine the exact causal effect of funding on primary vote share, there can still be an association between the two variables). Additionally, the categorical features we selected, a candidate's "profile," allow us to include qualitative features which have the potential to be just as impactful as our quantitative features when predicting primary vote share.

For our GLM, we'll be using a Gaussian distribution with an identity link function to predict a candidates primary vote share because we're not predicting counts so a Poisson distribution and a Negative Binomial distribution would not be sound choices. However, this does mean we're assuming a linear relationship with independent, Gaussian errors. We will fit the regression to the subset of candidates from our dataset that did not run uncontested as to account for obvious outliers. The model will be fit on a training set with 70% of the candidate data and tested on the remaining 30% to observe its accuracy when applied to new data.

We will then use a random forest as our nonparametric method as random forests tend to have high accuracy without over fitting like decision trees. Additionally, they require few assumptions unlike our GLM model which assumes linearity and Normal errors. Finally, we can interpret results from our random forest through the LIME explanatory method. We will also train/test the random forest model on a standard 70%/30% split of the candidates.

5.2 GLM Results

5.2.1 Regression Output

The results from our GLM can be seen below:

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	281
Model:	GLM	Df Residuals:	271
Model Family:	Gaussian	Df Model:	9
Link Function:	identity	Scale:	196.81
Method:	IRLS	Log-Likelihood:	-1135.8
Date:	Mon, 10 May 2021	Deviance:	53335.
Time:	16:50:04	Pearson chi2:	5.33e+04
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
x1	0.5156	1.902	0.271	0.786	-3.212	4.243
x2	-0.3711	2.434	-0.152	0.879	-5.141	4.399
x3	1.0912	4.027	0.271	0.786	-6.801	8.984
x4	10.5945	2.730	3.880	0.000	5.243	15.946
x5	-8.5562	4.491	-1.905	0.057	-17.358	0.245
x6	-1.7016	2.078	-0.819	0.413	-5.774	2.370
x7	5.8954	1.294	4.557	0.000	3.360	8.431
x8	-5.4872	0.444	-12.346	0.000	-6.358	-4.616
x9	4.186e-06	5.83e-07	7.177	0.000	3.04e-06	5.33e-06
const	45.6660	2.846	16.045	0.000	40.088	51.244

The coefficients prior to the constant correspond to the following list of features in order: Race, Veteran?, LGBTQ?, Elected Official?, Self-Funder?, STEM?, Number of Endorsments, District Candidate Count, Total Receipt of Funding.

5.2.2 Gaussian RMSE

Here we have the RMSE of the Gaussian GLM on the training and test sets.

```
Training set error for guassian GLM: 13.777004033411748
Test set error for guassian GLMS:    14.269859488265022
```

5.3 Bootstrap Standard Errors

Included below are bootstrap estimates of the standard errors run on 1000 iterations, for each feature coefficient. We see that the bootstrap standard errors mostly fall in line with the standard errors we received directly from our model.

```
Bootstrap std error for Race(x1): 1.855
Bootstrap std error for Veteran (x2): 2.320
Bootstrap std error for LGBTQ (x3): 4.372
Bootstrap std error for Elefted Official (x4): 2.614
Bootstrap std error for Self-Funder (x5): 4.607
Bootstrap std error for STEM (x6): 1.944
Bootstrap std error for num_endorsements(x7): 1.518
Bootstrap std error for district_candidate_count (x8): 0.437
Bootstrap std error for Total_Receipt (x9): 0.000
Bootstrap std error for const|: 3.027
```

5.3.1 Interpretation

The GLM found a positive association between share of primary vote and Race, LGBTQ, Elected Official, number of endorsements, and Total_Receipt. On the other hand, Veteran, Self Funder, STEM backgound, and the Number of Candidates in primary all had negative associations. Interestingly, only Elected Official, Number of Endorsements, Number of Candidates in primary, and Total_Receipt were statistically significant at the 5% level. Below is an interpretation of these statistically significant variables:

- If a candidate was an elected official, then their primary vote share increased by 10.59%, on average.
- If a candidate was self funded, then their primary vote share decreased by 8.56%, on average (significant at the 10% level).
- For each additional endorsement a candidate received, their primary vote share increased by 5.9%, on average.
- For each additional opponent in the primary, a candidate's primary vote share decreased by 5.49%, on average
- For each additional $100k received by a candidate, their primary vote share increased by 0.41%, on average. (This seems small, but is enlarged given candidates are often raising millions of dollars)

While these interpretations make sense and the coefficients of the remaining insignificant features are interesting, the log-likelihood indicates that this model is not an excellent fit for the outcome data and thus applying this model to new data is questionable.

5.4 Random Forest Results

5.4.1 RMSE

For the second part of this research question, a random forest regression model was fit to the outcome variable 'Primary %,' with the same features used in our GLM model.

The random forest model was trained and tested on a 70/30 split of the data and then measured for accuracy with a standard root mean squared error metric (RMSE). The results on the training and test sets were as follows:

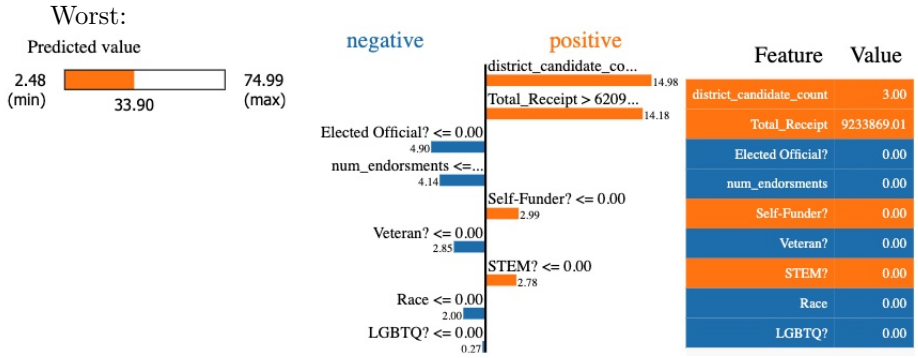
```
Training set error for random forest: 5.261226192942388
Test set error for random forest:     13.173634060256042
```

As an absolute measure of fit, the RMSE for the training set is somewhat promising. A standard deviation of 5% for the unexplained variance of the model is not terrible. The RMSE of the test set is notably worse at 13%, but still better than the RMSE of our GLM on the test set. Still, in an election, 13% of the vote could be a noticeable swing especially when a large number of candidates are running.

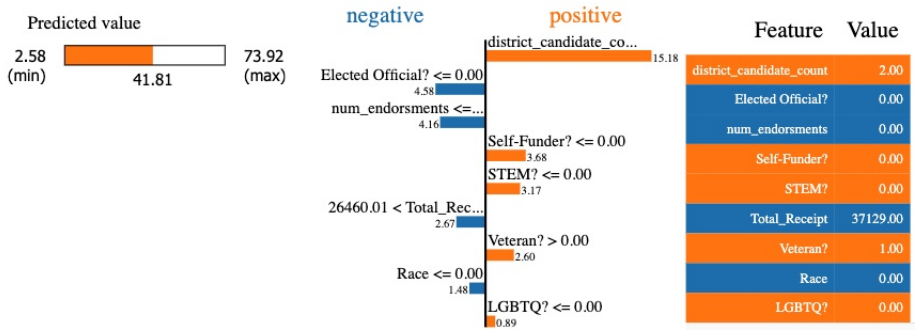
5.4.2 LIME

These are the LIME outputs for the two candidates our model did the worst job predicting primary vote share:

Worst:



Second Worst:



And here are their entries in our dataset:

Candidate	District	Primary %	Race	Veteran?	LGBTQ?	Elected Official?	Self-Funder?	STEM?	num_endorsments	district_candidate_count	Total_Receipt	forest_pred
Andrew Janz	U.S. House California District 22	78.830002	0.0	0.0	0.0	0.0	0	0.0	0.0	3	9233869.01	40.564100
Robert Kennedy	U.S. House Alabama District 1	80.769997	0.0	1.0	0.0	0.0	0	0.0	0.0	2	37129.00	42.896601

As we can see, the difference in what our model predicted (forest_pred) and the actual vote share the candidates received (Primary %), is quite large. It does makes sense though that these two candidates had the worst predictions as both of them received astronomically high vote shares, each around 80%.

Using LIME to explain the predictions from our random forest, we see that for the two candidates with the worst predicted primary vote share, a larger value for 'district_candidate_count' as well as a value of 1 for 'Self Funded' contributed to decreases in expected primary vote share while larger values for 'Total_Receipt' as well as a value of 1 for 'Elected Offical' contributed to increases in expected primary vote share. These results are consistent with the trends observed in our GLM model.

5.5 Discussion

Ultimately, the random forest had more accurate predictions than our GLM as it performed (slightly) better on our test data. This likely has to do with a failure to satisfy the assumptions necessary to use a Gaussian GLM. With fewer assumptions, our non-parametric model seems to have better picked up on the trends in our data. Although, one limitation of the random forest is that it can be difficult to interpret its results, unlike the GLM. This is exactly why we used the lime package to analyze two of the model's worst predictions as it provides insight into how predictions were made. We're not very confident in our ability to apply this GLM or nonparametric model to future elections as there's no guarantee these associations will hold true in the future. In fact, it's possible both models will decrease in accuracy as time goes on. For example, it doesn't seem like experience as an elected official is as valued as it used to be. Additionally, there will likely be more LGBTQ and non-white candidates as public opinion and demographics shift. It is especially interesting though that our random forest can be as accurate as it is without taking into consideration the partisan lean of both candidates and their district. It makes us wonder what factors really contribute to how Americans vote if we can predict vote share without explicitly considering a candidate and their electorate's ideological position. We could possibly improve our model by including more candidates and thus expanding our training data. Perhaps this could be achieved by including candidates from local elections. Of course, trends in local politics likely differ from that of national politics.

6 Conclusion

In regards to our causal inference question, we were unable to determine the exact impact funding has on vote share. But our GLM did find that money raised and experience as an elected official has a positive association with vote share while the number of candidates running in a primary along with whether or not the candidate is self funded both had statistically significant

negative associations with vote share. Our nonparametric random forest also relied heavily on the features our GLM found significant when predicting a candidate's primary vote share. Based on these results, it's clear that while money might not cause a candidate to win, there is still a positive correlation between the two. Perhaps money is a signal of a candidates popularity and thus people give more to the candidates they think will win. This could also explain why self funding has a negative associating with vote share. Perhaps self funded candidates don't have the support needed to fundraise and earn votes. And it's intuitive as to why more candidates in a primary field hurts one's chances of winning.

Ultimately, we aren't suggesting that candidates should not fundraise and try to earn money. Clearly campaigns need money to pay for staff, advertising, get out the vote efforts, and more. But without proof of a causal relationship between money raised and primary vote share, it does call into question that vast sums of money that go into political campaigns today. Is all of this spending necessary? Could political spending be more impactful if it was allocated toward more pressing matters? More research into the effects of political funding and spending will certainly be needed to better understand how much campaign funding is enough.

It's worth noting that one major limitation of our findings is that they're all based on Democratic primary data from 2018. Thus there is no guarantee that our findings will hold in the upcoming 2022 elections and beyond. If we were to conduct future studies to build off this work, we would like to see a similar analysis on Republican primaries as well as general elections as it's possible different features are significant in Republican primaries, and perhaps spending has a more apparent causal relationship in general elections. We would also like to see more tests of causality with different instrumental variables so we can better understand the causal effect of money in politics.

References

[FEC18] FEC. *candidate_summary-2018*. 2018. URL: `https://www.fec.gov/files/bulk-downloads/2018/candidate_summary_2018.csv`.

[Meh18] Dhrumil Mehta. *primary-candidates-2018*. 2018. URL: `https://github.com/fivethirtyeight/data/blob/master/primary-candidates-2018/dem_candidates.csv`.