Examining the Factors Behind Political Candidates' Success

Brandon Hong, Eddie Liu, Zayan Khanmohammed, Amy Zhong

DATA C102: Data, Inference, and Decisions

May 8, 2023

Data Overview

The dataset we decided to use was FiveThirtyEight's compiled dataset looking at Democratic and Republican primary elections for the U.S. Senate, U.S. House, and governor in 2018. The data was split into two separated comma-separated-value (CSV) files, one for Democratic candidates which had 811 entries, and one for Republican candidates which had 774 entries. There were some overarching columns present in both files which represented general information about the candidates such as the state represented, the district the candidate represented, the office they were running for, and the percentage of the vote the candidate received in their respective primary. The motivation behind this separation, however, was because most attributes were related to specific groups/public figures' endorsements.

Since this dataset includes all candidates appearing on the ballot, not counting races featuring incumbents, this is considered a census. The only group that would be considered to be excluded would be those in races featuring an incumbent candidate. The motivation behind excluding these types of races is that an incumbent has many more variables affecting their voting results. Our data is more focused on new candidates and how their background and endorsements can affect their candidacy.

In our compiled dataset, each row represents a candidate running for a political office in the 2018 election cycle. Since candidates can only run for one office at a time, there is no overlap between candidates and our data. This data was compiled since endorsement data and the percentage of votes won, as well as demographic information for candidates, are all public and easily accessible. A lot of the general information was supplied by Ballotpedia, a nonpartisan online political encyclopedia. While there is uncertainty about whether candidates knew they would be included in this dataset, there is plenty of data surrounding candidates during election season so that candidates and political groups have an accurate representation of how they stand. Moreover, the fact that all this information is publicly available and found on each candidate's website signifies that they understand their data is susceptible to data collection. Our dataset is not modified for differential privacy, as each row displays the name of the candidate. The granularity of our data is at an individual level, which is logical since the analysis being conducted is identifying trends across candidates to develop broader claims about the progress of the two major political parties in the U.S. Since our data collected all candidates from races without an incumbent, there were not any concerns about selection bias or convenience sampling. Furthermore, since the only quantitative measurement conducted is the percentage of the vote received by the candidate in his or her primary, we did not need to account for measurement errors. This dataset is observational, scraping publicly available information to create profiles for these candidates and identify patterns and trends for future elections.

One key attribute that would be useful for our analysis is the amount of funding a candidate received. To promote themselves, candidates across all levels of government spend money on advertising. After all, how are voters supposed to choose a candidate if they are not aware of them? Since we believe that funding is a critical indicator of political success, it would be interesting if this information was available in this dataset and we could quantify this relationship.

There were many missing binary values in the columns of the Democratic and Republican datasets. To combat this issue, we operated under the assumption that a missing value meant that the candidate did not receive that specific endorsement. Besides any necessary exclusions where there were missing values for our primary outcomes, this dataset did not need any true preprocessing as this data was previously cleaned by FiveThirtyEight.

Research Questions

Research Question 1: Multiple Hypothesis Testing & Causal Inference

Our first research question was “Does having the support of certain groups or individuals cause a Democratic candidate to win their primary elections?” In our exploratory data analysis, we were able to observe trends in the effectiveness of different notable endorsements. We thought that multiple hypothesis testing and causal inference techniques could provide us with a way to explore these trends and see if certain supports are statistically significant. We chose these techniques since multiple hypothesis testing allows us to examine several hypotheses on the effect of the numerous endorsements simultaneously. Causal inference is useful since we are trying to make a causal claim of a treatment (support of the Democratic party) on whether a candidate wins their primary. If we can determine the effect of groups, specifically the Democratic party, on a primary election, we can determine the relative importance of these endorsements. We also hypothesize that a similar analysis can be used for the Republican party and then use this information to compare which party is more heavily reliant on endorsements to decide the outcome of primaries. A limitation of multiple hypothesis testing is that the more tests we conduct, the probability of finding a significant result by chance alone also increases. A limitation of causal inference is that it relies on the assumption that all confounders are accounted for which is not possible when measuring success and by violating this assumption, we may introduce some bias into our estimates.

Research Question 2: Prediction with GLMs and Nonparametric Methods

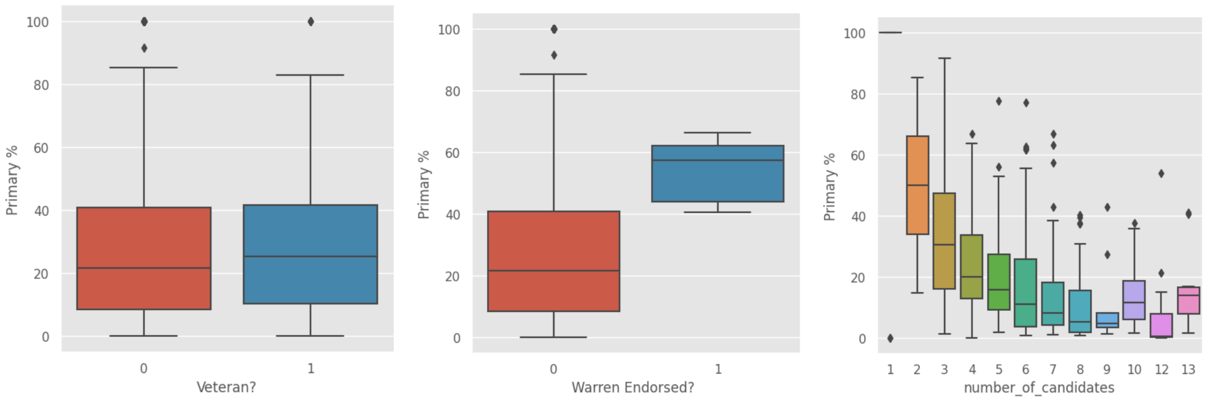
Our second research question was “What percentage of votes will a candidate receive in their primary elections with the support(s) of specified groups/individuals?” We used the prediction option since our research question is about creating a prediction based on certain endorsements/support from groups. We did not delineate a specific group/individual since when we initially analyzed the data, we were unsure which endorsements held the most weight. Additionally, this allowed us to broaden the scope of our analysis, fine-tuning our models with different combinations of endorsements to create the best predictions. Creating Frequentist and Bayesian models will allow us to learn these weights and what distributions serve as the best models for candidate data. Since the attributes of the Republican and Democratic candidates were different, we also had to create separate models for both sets.

This question is important because if one can accurately predict a candidate’s primary vote percentage based on demographic information and the support they received, this indicates how crucial financial and political support can be during an election. One limitation of our prediction model is that our data does not account for important factors that can impact an election, such as the candidate’s ideology, personality, and the demographic makeup of where the candidate is running.

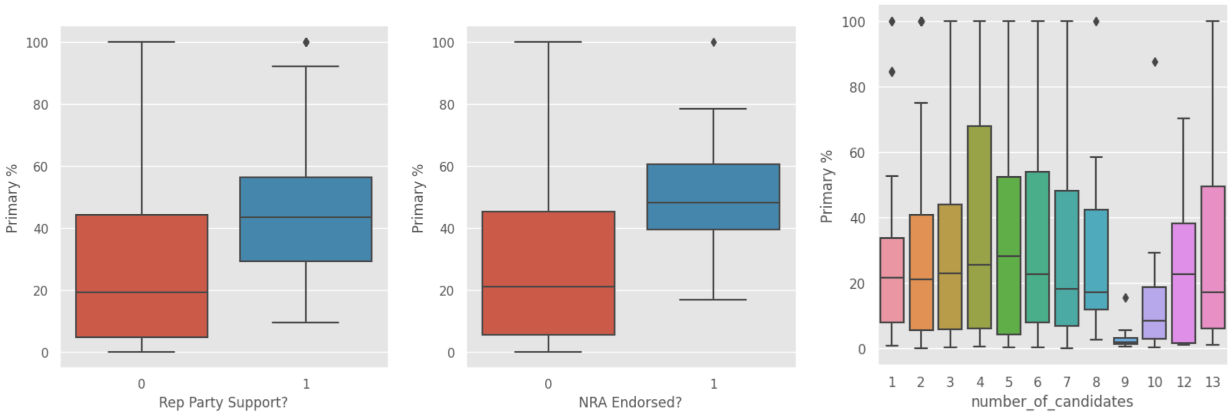
EDA

(See notebook for all visualizations)

Democratic Data



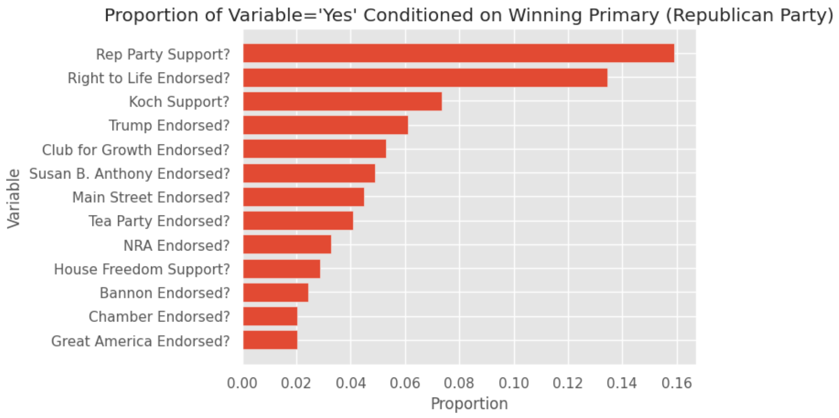
Republican Data

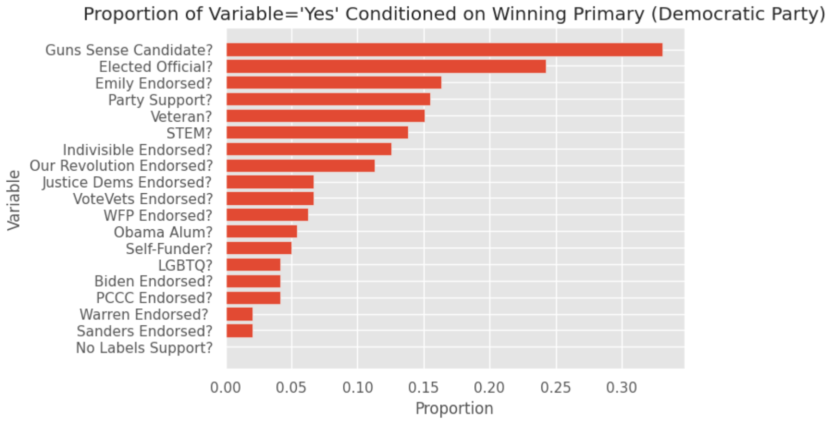


Categorical: Based on the box plots, the primary percentage was higher when a candidate was endorsed. Some endorsements seemed to favor candidates more heavily than other candidates. Some categorical variables that are not present in the Republican dataset also seemed to not have an effect, such as Veteran, LGBTQ, STEM, and whether they were self-funded. However, the Trump, NRA, Koch, and Chamber Endorsed categories have a much higher primary % when endorsed versus when not endorsed. We might want to follow up on the Trump, NRA, Koch, Chamber Endorsed, Veteran, LGBTQ, STEM, and self-funded categories, as they might have a greater influence on the primary %. Since we can determine which endorsements are highly correlated with a high primary percentage, we can potentially use them in our multiple hypothesis/causal inference problem or our GLM models.

Quantitative: The last box plot graph shows the relationship between the primary percentage and the number of candidates in each election. Logically, it makes more sense for the candidates to receive a lower primary percentage with more candidates, as there is more competition. This plot shows this general relationship, but there are some exceptions when there are more than 9 candidates.

Bar Chart Support

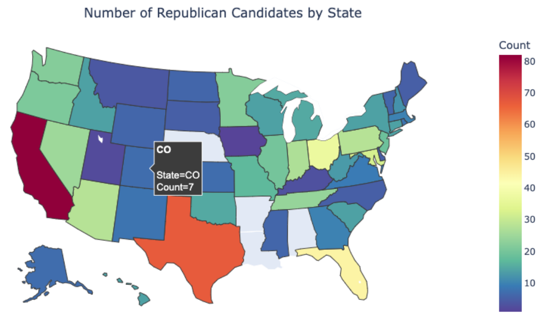


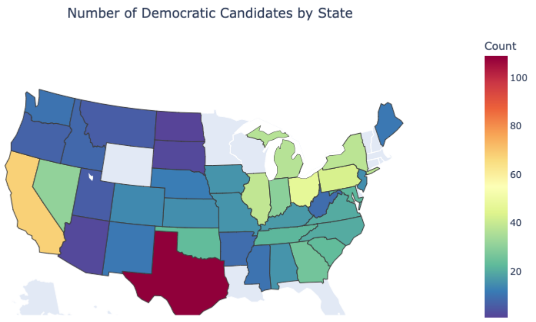


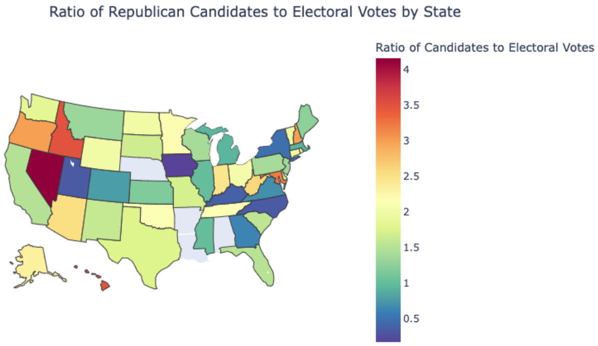
We created a bar graph to show the proportion of Democratic and Republican candidates who were more likely to win their primary with specific endorsements/supports. Some of these relationships that we may want to follow up on are why No Labels-affiliated groups appear to have little impact on Democratic primary winnings as well as the same for Bernie Sanders endorsements, especially when he was a candidate in the 2016 Democratic presidential primaries and would go on to win the same primaries in 2020. Additionally, we may want to follow up on why the aforementioned "prevalent" variables appeared in greater proportion among candidates who won their primary—is it because they actually impacted winning a primary or because they simply appeared more frequently than other variables?

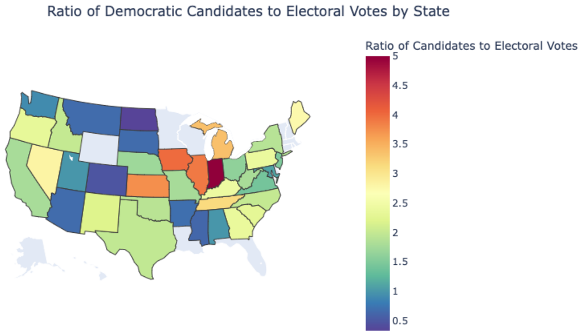
These visualizations are especially relevant to our first research question. They suggest potential answers to whether or not the support of certain groups/individuals may cause a candidate to win their primary elections. This may also be relevant to our second research question, which attempts to predict the percentage of votes a candidate will receive in their primary elections given the support of certain groups/individuals. We also do not know if any confounding variables may impact a variable's effect on winning a primary.

Maps


Number of Republican Candidates by State


Number of Democratic Candidates by State


Ratio of Republican Candidates to Electoral Votes by State


Ratio of Democratic Candidates to Electoral Votes by State

We chose to visualize the number of candidates by state. We initially did the raw count, but we decided to add a normalized version, choosing to normalize by the number of electoral votes as a rough measure of population proportionality.

These visualizations are important because they can help contextualize any trends we see in endorsements and make sure to control for any geographic associations. This also helps us get a better understanding of which states have more candidates running for those seats because, in races with more candidates, we hypothesize that endorsements have a larger impact on the outcome of the race. From these visualizations, we were intrigued that for Democratic candidates, the number of candidates seemed to be similar across the Midwest and once again similar on the East Coast. We expect that this may be due to population, but interestingly enough this trend does not hold for Republican candidates. Another interesting trend was that for the normalized graphs, the states with the highest ratios were clustered near each other (Pacific Northwest for Republicans and Upper Midwest for Democrats). For the next steps, we want to explore if certain endorsements occur for these clusters that are higher than other states.

Multiple Hypothesis Testing

Methods

We wanted to perform multiple hypothesis testing to see whether the support of certain groups or individuals affected a candidate's primary race. When candidates are supported or endorsed, they receive a commitment from that specific organization to help volunteer, fundraise, and even help an individual gain access to more events. There is a greater effect when a person receives multiple endorsements/support, but we wanted to focus on the individual effect of having an endorsement/support of a certain group. Therefore, we will test multiple hypotheses to find which endorsement/support is the most effective for candidates.

We used the chi-squared test of significance to test each hypothesis and calculate the p-values. There are many test statistics to choose from, but we wanted to use a test statistic that worked well with binary values and tested for correlation between binary variables using contingency tables. We even considered using McNemar's test statistic, but our data violated all of the assumptions for it.

This set up our null hypothesis to have a specific format: there is no correlation between having the ___ Endorsement/Support and winning the Democratic candidate's primary race. Our alternate hypothesis had a format too: there is some correlation between having the ____ Endorsement/Support and winning the Democratic candidate's primary race. We used correlation to test our hypotheses because the chi-squared test states whether two variables are independent. If they are not independent, this means that they have some correlation. After setting our null and alternative hypothesis, we created contingency tables to run in our chi-squared test, and it would output a test statistic, degrees of freedom, and the p-value. If the p-value was below our alpha value (.05), this meant the endorsement/support was statistically significant, and we can reject the null hypothesis. A p-value above our alpha value means that the endorsement/support was not statistically significant, and we cannot reject the null.

The two ways we will correct for multiple hypothesis testing are the Bonferroni correction and the Benjamin Hochberg correction. The Bonferroni correction controls the Family-wise Error rate, the probability that a single false positive appears. The Benjamin Hochberg correction controls the False Discovery Proportion, the expected proportion of false discoveries among all discoveries. For our multiple hypothesis testing research question, the Benjamin Hochberg correction would be more appropriate. Since we want to see the effect that an endorsement/support has on a candidate's chance at winning their primary race, we want to prioritize more people winning elections instead of finding only the candidates that we falsely predict to win their primary with certain support.

Results

After performing all the hypothesis tests, we declared that a certain endorsement or support was not statistically significant if the p-value was above .05. Therefore, we found that the Democratic Party Support, Biden Endorsement, Gun Sense Candidate, Elected Official, Emily Endorsement, and PCCC Endorsement were statistically significant. Only Race and the Our Revolution Endorsement was not statistically significant. We also calculated the power for the eight-hypothesis test and received a value of .8274. This is a very high power value, and it means that this hypothesis test can detect 82.74% of detecting an effect that actually exists.

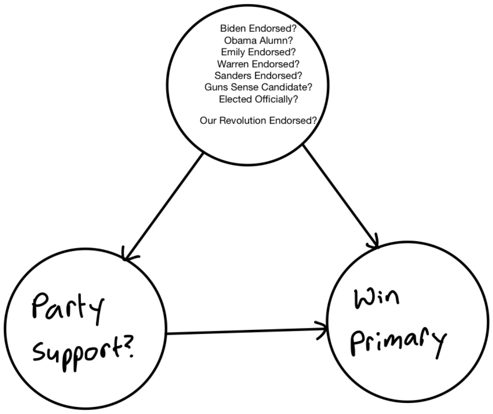
Discussion

When we used Bonferroni correction, we divide the alpha-value by the number of hypothesis tests we performed. This creates the new p-value threshold of every single test to be .05 / 8 = .00625. With our new p-value threshold, there were no changes to the discoveries. Only the Our Revolution Endorsement and Race were not statistically significant, and this did not change from before. Bonferroni correction mainly focuses on the FWER, which is not a prevalent issue in our research question.

When we use the Benjamin Hochberg correction, we sort the p-values, draw the line y = k * alpha/m, find the largest p-value that's under that line, and use that p-value as the threshold. When we sort the p-values, we get [2.21134257e-19, 2.99278943e-12, 4.79380074e-06, 9.19296013e-05, 1.40240929e-04, 3.68749362e-03, 2.88433362e-01, 7.15277767e-01]. After putting it in the formula and finding the largest p-value under that line, we can use that p-value (3.68749362e-03) as the threshold. When we use this correction, there is no change in the discoveries. Due to the massive difference in some of these p-values, the threshold change would not affect any of the discoveries. Since we only performed 8 hypothesis tests, the Benjamin Hochberg correction was easily swayed by the gap in p-values.

With more new data incoming in the future, we could use the LORD algorithm to correct for multiple hypothesis testing. It would still be possible to use Bonferroni and Benjamin Hochberg, but we would need to know certain details before using them again. One major limitation of this correction was the amount of data we had. In the real world, there are hundreds of endorsements/supports that a candidate can receive. This would have greatly changed the outcome of our results because the p-value threshold would have gotten much smaller. For example, if we performed 20 hypothesis tests, the PCCC Endorsement would no longer be statistically significant and would not be discovered with the Bonferroni Correction. We would need a lot more data to perform more hypothesis tests and utilize both corrections to their fullest potential.

Causal Inference



Methods

After conducting multiple hypothesis testing to find what attributes are associated with a primary election victory, we wanted to conduct causal inference on the statistically significant variables. We wanted to find the causal effect between the party support for Democratic candidates and the endorsement/support of different organizations. We conducted two separate Ordinary Least Square (OLS) models to examine this relationship. As our baseline, we did a naive OLS model comparing our treatment to our outcome. This however failed to account for the confounding variables present in our data. To mitigate this issue, our next model used the technique of outcome regression. The idea was that by incorporating potential confounders, the regression model attempts to account for their influence on the outcome, thereby reducing the confounding effect. One set of confounders was based on our multiple hypothesis test results, including what was found to be the most significant indicators of success such as being endorsed by Emily's List or receiving the Gun Sense Candidate Distinction. The other set of confounders was based on our domain knowledge that if our candidate was someone who worked in the Obama administration or received an endorsement from the most influential members of the party like Joe Biden, Elizabeth Warren, and Bernie Sanders, they would be more likely to receive support from the Democratic party. We grouped all these confounders together in our DAG since they all have a similar confounding mechanism which is that the Democratic party partially bases its support on these confounding endorsements of powerful Democratic political action committees (PACs) and leading members of the party, as well as if they were previously elected officials.

In order to conduct any analysis of our models, we have to assume that the unconfoundedness assumption holds up. Using this assumption, we can make claims about the true causal effect of the support of the Democratic party on whether or not a candidate wins their primary.

The dataset does not contain any variables that act as colliders. However, when conducting additional research including general results, it is important to consider a particular collider variable: candidates who ultimately win the entire election. This variable becomes relevant because winning the primaries and receiving endorsements both have a clear impact on whether or not a candidate wins their overall election.

Results

When conducting our naive OLS model, we discovered that party support was statistically significant and had a coefficient of .945. This is an extremely high coefficient, which meant that you could make a near-perfect prediction with just party support. However, we know that there are confounding variables that must be accounted for. We used these six variables as our control variables in our OLS regression and found that they were all statistically significant except for the Biden Endorsement and PCCC Endorsement. This makes sense because this dataset is from 2018, which was before Biden was elected president. Therefore, his endorsement is less significant prior to his presidency. The PCCC endorsement may not be statistically significant due to Elizabeth Warren's participation in PCCC. Since she already endorsed someone, it is possible that the PCCC endorsement overlaps and is less significant.

Afterward, we conducted another OLS model with variables that we used our domain knowledge to select. Within these variables, Biden Endorsed, Obama Alum, and Warren Endorsed were not statistically significant. The Obama Alum variable must be recent because Obama had just become president, so it makes sense that it is not statistically significant now. However, the Warren Endorsement is slightly surprising since Elizabeth Warren has been a notable Democratic figure for a while. We combined all multiple hypothesis tests and domain knowledge OLS models and found that the Sanders Endorsement was no longer statistically significant. We found that the coefficient for Party Support greatly decreased by nearly half to .592, which meant that these confounding variables had a significant effect. However, it is still clear that the Party Support had the highest coefficient, and therefore, is the most important for a Democratic candidate's success. Due to the many confounding variables that are not present in this dataset, we are unable to make causal claims about our treatment and outcome variables, but our results are still important to note.

Discussion

One limitation of our model is our underlying unconfoundedness assumption. While our model operated under the assumption that there were no hidden variables, for a more in-depth analysis, this should be verified further as there is a plethora of variables that may or may not be confounding variables in this relationship. Another limitation of our model is that our data only covers one election year. The strength of the party's support varies by year, so these results are not generalizable to past or future elections. Additional data, such as a quantification of a candidate's ideological leaning, would better inform a model estimating this causal effect. Another useful attribute would be how much funding they received and how much money a candidate spent on advertisements. Using prior knowledge and our model's results, I am confident that there is a clear correlation between obtaining party support and winning the primary, but I am not confident that the calculated coefficient of causation is accurate and generalizable. We would need a comprehensive list of variables and run various sorts of analyses to control for confounders to make any sort of confident causal claim.

Prediction with GLMs and Nonparametric Methods

Methods

We are trying to predict 'Primary %' or the percentage of the vote received by a candidate in their primary. The features we're using are 19 binary variables indicating belonging to a certain group (e.g. veterans, LGBTQ, elected officials, etc.) or yes/no support (i.e. financially or endorsement) from different entities (individuals or organizations) for Democratic candidates, 13 binary variables indicating yes/no support from different entities for Republican candidates, and the number of candidates running in each respective primary for both Democrats and Republicans. We chose to use all binary variables available to us in our data because we wish to predict the percentage of votes from the known support of different entities. We also chose to use 'number_of_candidates' because how many competitors a particular candidate has should impact what percentage of votes they receive.

We tried fitting Poisson, Negative Binomial and Gaussian Frequentist and Bayesian GLMs for both datasets. Gaussian was chosen because our outcome is continuous, Negative Binomial because we may not assume that mean is equal to variance, and Poisson as a baseline. In future studies, since we make the assumption that our outcome is bounded (non-negative), the Gamma or Inverse Gaussian distributions should be considered. For each distribution family, we removed any features that were not statistically significant, meaning they had p-values greater than 0.05. We did not use priors for the coefficients of the Bayesian GLMs, choosing to use PyMC3's "flat" prior that is uniform over all real numbers instead.

The nonparametric methods we used were random forest and decision tree models. We chose to use decision trees because they produce relatively good predictions on training data, are generally easier to interpret, and provide a baseline to compare random forest performance. Since a weakness of decision trees is their tendency to overfit, we chose to use random forests because they can handle large datasets efficiently and provide better test data performance (and thus generalizability) compared to other models.

We will perform Frequentist model checking using chi-square and log-likelihood metrics and Bayesian model checking using PPD plots. For all GLM and nonparametric models fitted on a training set, we will make predictions on a testing set and evaluate performance with RMSE.

Results

Frequentist: Our best model for the Democrat dataset was the Poisson distribution, with a testing set error of 16.52. We were able to greatly reduce the error in our model, and we can still improve the model in the future by removing more variables. We used the idea that the percentage calculated by our GLM would be equivalent to the primary percentage. This idea works because the percentage that a candidate will win in their primary should be equal to the primary percentage that they receive in the race. On the other hand, there was no clear best model for the Republican dataset. Each of the three model's training and testing set RMSEs were all within 0.2 from each other and close to the value of 30.

Bayesian: Using our model, we noticed that the Gaussian distribution was a better fit for both the Democrat data and Republican data. The curves were more natural, and there were barely any gaps between the curves.

Using our model for the Democrat dataset, we achieved a RMSE of 20.04 on the training set and 18.82 on the testing set. This means that our model performed better on the test set than on the training set. Our model does not underfit or overfit our data, but it is still a considerable error. Since most candidates will probably win their primary with a 20-40% rating, a RMSE of 18 could easily shift whether a candidate wins or not. There are probably outliers in some states where there is only 1 candidate, so it may cause greater errors.

Nonparametric: Using our model, we noticed that the random forest model achieved a training set RMSE of 11.8 and test set RMSE of 14.32 on the Democratic dataset. This means that our model performed worse on the test set than on the training set. For the Republican dataset, our model achieved a training set error of 30.24 and a test set error of 30.22. This means that our model performed slightly worse on the training set than the testing set.

Using our model, we noticed that the decision tree model achieved a training set RMSE of 11.26 and test set RMSE of 18.08 on the Democratic dataset. This means that our model performed noticeably worse on the test set than the training set. For the Republican dataset, our model achieved a training set RMSE of 30.02 and test set RMSE of 31.67. This means that our model performed slightly worse for the test set than the training set.

As most candidates likely win their primaries with a 20-40% rating, this error metric could potentially shift the result in whether the candidate wins or not.

Model Performance: Test RMSE

Model	**Democrat Dataset**	**Republican Dataset**
Frequentist: Poisson	16.52	30.69
Frequentist: Negative Binomial	18.67	30.59
Frequentist: Gaussian	18.81	30.55
Bayesian: Poisson	36.13	41.19
Bayesian: Negative Binomial	36.16	41.20
Bayesian: Gaussian	18.82	30.74
Nonparametric: Decision Tree	18.08	31.67
Nonparametric: Random Forest	14.32	30.32

Discussion

Performance: To summarize, for Frequentist modeling, the Poisson family performed the best for the Democrat dataset and there appeared to be no clear family winner for the Republican dataset. For Bayesian modeling, the Gaussian family performed the best for both datasets. Lastly, for nonparametric models, the random forest performed better than the decision tree, as it creates an ensemble where each tree only looks at a random subset of features. Demonstrated by its relatively high training error, but lowest testing errors for both datasets out of all models, we are confident that this random forest model will generalize well to future, unseen data.

Model Checking: For our Frequentist models, we checked the model by looking at chi-square and log-likelihood metrics. Our Poisson model for the Democratic dataset had a chi-squared of 6450 and log-likelihood of -4562. When checking our models, we typically look for a high log likelihood value and a chi-squared value of the difference of n (number of rows) and p (number of features). We would like to explore the Poisson model more and see why it does not give us better metrics when model checking. For our Bayesian models, we checked the models by looking at the posterior predictive distribution plots.

Key Differences: A distinction between the two models was that Gaussian performed better under the Bayesian approach, whereas the Frequentist models had less clear “winners”.

Interpretation: For GLM, we can interpret the model by looking at the coefficients that are displayed from the summary. For example, the coefficient of Biden Endorsed is 0.549. We can interpret this to mean that if a candidate is Biden endorsed, his or her primary percentage is predicted to increase by 0.549. Alternatively, the coefficient of number of candidates is -0.254. We interpret this to mean that increasing candidates can lead to decreasing primary percentage votes. We can plot the decision tree and see where splits are made for predictions (based on different variables). The plots we included above only depict the first few layers, as the decision tree model had unlimited depth and thus would be quite difficult to visualize completely. Random forests are much more difficult to interpret because, in this high-dimensional problem, individual trees see vastly different subsets of features.

Limitations: One limitation of the GLM models was that we did not have a prior as we weren’t too sure which to choose. Another limitation of the decision tree nonparametric model is that it is prone to overfitting. It is important to note that the observation for both Democratic and Republican predictions that train RMSEs were greater for random forests than decision trees, but test RMSEs were greater for decision trees than random forests. This indicates that the decision tree with no limit on tree depth and uses all features most likely overfits. The random forest model reduces this overfitting with an ensemble of trees as well as only a random third of the total features per tree but has a tradeoff with interpretability.

Additional Information: One piece of data that could help improve our model is knowing the relationship between different support groups and parties. In the world of politics, one endorsement can often influence other endorsements for other groups. It could be helpful to understand which support groups are more strongly related to each other, as it could help us shift the weights on each support group in our models. Additionally, funding amounts for each candidate would be useful for improving our models.

Assumptions & Uncertainty: We have to assume that the primary percentage is independent of the candidate and that all the variables can be described by a distribution. We are also assuming that errors are also independent. One thing that causes uncertainty is that not all candidates can be nominated for an endorsement. Therefore, we have to assume that every candidate can receive an endorsement or support. This creates a lot of uncertainty in our variables and is reflected by our high error.

Conclusions

Overall, our final results for our multiple hypothesis testing, causal inference, and prediction (GLMS and nonparametric) showed pretty interesting results.

We found that performing multiple hypothesis testing can be an effective way to find statistically significant variables, but it is heavily dependent on the amount of data that we have. In our case, many variables were statistically significant, but they would not be considered at all if we included 10-20 more hypothesis tests.

Our causal inference showed that there is a clear association between a candidate obtaining support from the Democratic party and winning their respective primary. We also found through outcome regression that this effect is much weaker when taking into account the many confounders between our treatment and outcome variables.

For our GLM models, performance tended to vary based on the family we built our model on and the dataset used. The Gaussian model clearly performed the best for Bayesian modeling, whereas our Frequentist models had a less clear winner. For nonparametric models, our random forest performed better since it creates an ensemble and randomly selects features to use. Of course, there are limitations and assumptions within each model that need to be taken into account in addition to error rates and model-checking methods.

For our project, we chose not to merge different data sources, since our research questions could be solely answered with this dataset. Since our models are only trained on data from the 2018 election cycle, they should only be generalizable to inferences relating to that election cycle. The strength of specific endorsements can grow or shrink and there is not a quantitative way to measure that without looking at other election cycle data. This served as a major limitation to our findings, as our findings can only be applicable in this narrow scope. Future studies could build on this by applying the same analysis to past elections, as well as election cycles after 2018. While some of the political figures might hold different weights, one could compare the strength of endorsement from the respective political party, as well as political action committees.

Another limitation of this dataset is that it fails to account for the influence of third-party candidates. Third-party candidates are less influential in US politics compared to other developed countries, but they still have the power to influence election results. For instance, if a third-party candidate has Democratic-leaning ideologies but is not associated with the Democratic party, they can take away a share of the voters who would typically side with the Democratic candidate if the third-party candidate was absent. These third-party candidates were not listed due to the added complexity of each smaller party. This data would be useful as it might explain any potential outliers in our data. If we had this data, we would also be able to expand our analysis to quantify exactly how much third-party candidates can influence elections and what types of endorsements/qualifications help them stand out in comparison to others. While we are unable to account for these individuals, their potential influence on the percentage of votes received for candidates in the primary should not be ignored.

Based on our results, we believe that certain endorsements should receive more attention and have more impact on candidates' results. Especially in times like now, we should be placing extra emphasis on gun laws and racial discrimination. Our results showed that Party Support was the most effective, but we believe that endorsements by Gun Sense Candidate, for example, should hold equal weight as well. Ultimately, every movement is equally important, so we should take more action on the movements that are lacking impact now.

Overall, one major conclusion was how impactful the endorsements of PACs are in comparison to major political figures. Since PACs are what serve as major funding for most candidates, our hypothesis is that funding plays a larger role than most voters might anticipate. While we as voters hope to nominate the candidate that best represents our ideologies and is best suited for the role, the initial whittling down of candidates at the primary stage may eliminate promising candidates that are not able to receive support from political figures/PACs. An interesting angle to build off this research is to compare the influence of funding and popularity to add more complexity to these models.

It was also interesting to note the stark differences in the Democratic and Republican data. As our nation becomes increasingly polarizing, the factors which influence voters are also completely different. This may indicate that it can be difficult for either party to swing votes from the other side of the political spectrum. It would be compelling to see how Democrats and Republicans continue to combat this major issue and how that can affect campaign strategy moving forward.