Data Overview

The dataset we used contained candidate information split into Democrat and Republican candidates from the 2018 U.S. Primary. The data was gathered from various sources like Ballotpedia, candidate websites, and news reports. The data represents a census since each candidate who ran in the 2018 U.S. Primary was represented. Our data did not exclude any groups in regard to the candidates, and the participants were aware of their data being publicly available as they were running for candidacy. The granularity of the data is at the candidate level and each row in the data represents one candidate, including information such as the percentage of votes they received in the primary. This allows us to analyze the individual candidates and identify potential confounders and subgroups when interpreting our findings, making them more detailed.

There is a concern about convenience sampling in the context of our data. For questions 1 and 2 a lot of the candidates had null values for the “Party Support?”, “Partisan Lean”, “Race”, “Veteran?”, and “LGBTQ?” columns, which led us to just drop those candidates and only use candidates that had “yes” or “no” values in the “Party Support?” column for question 1 and candidates that had values for all 4 characteristics for question 2. However, if there happens to be a systematic difference between those dropped candidates and the ones that did have values for those columns, this could potentially introduce some bias into our ATE estimate. The bias generated from convenience sampling could potentially change our causal estimate to incorrectly promote/discourage a causal relationship between party support and primary success or cause our GLM or random forests model to place incorrect weights on each characteristic coefficient.

Our dataset was not modified for differential privacy.

For our first question, since we were analyzing the causal effect of party support on all candidates, we were forced to only use confounding variables that were common to both the Republican and Democratic data sets. We wanted to use attributes of candidates such as race, lgbtq identification, and veteran status as confounders since these variables should affect both party support and primary outcome, but since these variables were only available for democratic candidates we were unable to use them. We wish that we had information on the republican candidates’ race, lgbtq identification, etc. in order to be able to use these more intrinsic qualities of candidates as confounders. These would help us account for more confounders in our inverse propensity weighting and estimation of the average treatment effect. In terms of our second question, we also wish that we had more characteristics of candidates that would potentially affect their organization's support. These could include a candidate's net worth, fame, etc, and would help us see if other candidate characteristics could be engineered to improve our model’s predictive accuracy.

We found there to be null values in the “Party Support?” column (first question) and “Race”, “Veteran?”, “LGBTQ?” and “Gun Sense Candidate?” columns (second question). The missing values typically represented unavailable data, such as if a group did not weigh in on the race or if a candidate’s website was not available. For both the first and second questions, to handle null values, we just removed all candidates who did not have values in those columns. This led us to exclude a lot of candidates from our causal inference and GLM/random forests model creations. We chose to exclude candidates with null values rather than impute their missing values because imputing would imply that those organizations/parties explicitly did not support the candidate or the candidate explicitly did not have one of the characterizations (race, lgbtq identification, etc), which would introduce lots of bias if we imputed and made these assumptions.

The Primary Candidates 2018 dataset that we used was split up into 2 datasets - one for republican candidates and one for democratic candidates. In the first question, since we wanted to estimate the causal effect of the treatment on the outcome for all candidates, we had to combine the two data sets in order to make a single data set that contained both republican and democratic candidates. This wasn’t necessary for answering our second question since we only looked at democratic candidates in that question. Additionally, in order to calculate the IPW estimate and average treatment effect for our first question, since the confounders we used were categorical features (state, district, and office type), we had to encode them with numerical values in order to account for them. The treatment (party support status) and outcome (whether or not a candidate won their primary) were also represented by binary variables so we had to map them to corresponding numerical values (1: “Yes” and 0: “No”). Since the second question only dealt with values in the data that were binary in nature, we had to map each of those values to either 0 or 1. We had to encode these values because we couldn’t calculate estimates with categorical data without encoding/mapping it.

Research Questions

Question 1

Our first research question was “Does party support status cause a change in the likelihood of success in primary elections?” This research question could potentially inform political strategists, candidates, and party officials about what kind of impact party support has on electoral success. Understanding the relationship between the two and whether it is causal could help guide campaign strategies and resource allocation. If it is found that party support significantly causes electoral success, campaigns may spend more time and resources towards securing backing from their party.

To that end, we chose causal inference as our method to answer the question. It is a good fit for our research question because it allows us to determine the causal relationship between party support status and success in primary elections. This would help us determine whether party support status directly influences the outcome of primary elections, or if other variables are having an impact. By identifying a causal relationship, we are able to provide more actionable insights than simply establishing a correlation.

There are a few limitations to this approach that we must consider. One major limitation is potential unobserved confounding variables. These could cause our estimations of causality to be inaccurate. It is difficult to account for all factors that may influence both party support status and electoral success and if not properly taken care of, our conclusions about causality could be incorrect.

Causal inference also relies on the Stable Unit Treatment Value Assumption, which assumes that each potential outcome of a candidate does not depend on the treatment assignment of other candidates. In other words, we must assume that party support of one candidate does not influence the success of another candidate. If the assumption is violated, it could lead to biased estimates of the causal effects.

Question 2

Our second question for this dataset is "Can we predict whether or not an organization will endorse a candidate based on candidate characteristics?". In particular, we aimed to predict whether a candidate was a Gun-Sense candidate based on partisan lean, veteran status, and LGBTQ status. By answering this question, political strategists and candidates could gain insights on how to attract endorsements from organizations. For example, if a candidate is running in a county where many voters vote according to Gun-Sense recommendations, they may focus their campaign on getting that particular endorsement. This analysis would highlight the factors that may be influencing a gun-sense endorsement, so the candidate could emphasize (or de-emphasize) them in their campaign.

Predicting whether a candidate would be endorsed by Gun-Sense could also be useful for voters, as it could help provide them with some information about a new candidate that has not yet been judged by Gun-Sense.

We will be using both a logistic regression GLM and random forests to predict if a candidate is endorsed by Gun-Sense using the candidate's characteristics. Logistic regression is a suitable method for this question due to its ability to model binary outcomes, which in this case would be an endorsement or not. Logistic regression is also especially useful in this case because it provides interpretable coefficients for each variable. This would give an indication of how much influence each predictor variable has on the outcome of being a Gun-sense Candidate, which is a useful tool for campaigns. Random forests are also a useful nonparametric choice because there are many candidate characteristics which we use as explanatory variables in the model.

Logistic regression does have several limitations to consider. It assumes linearity in the log odds of the outcome variable and the predictor variables, which may not always be true in the real world. It also struggles with multicollinearity, where predictor variables are highly correlated. This would result in incorrect coefficient estimates. If the relationships between predictor variables and the outcome are very nonlinear, or if there is a high level of multicollinearity, logistic regression may not be very accurate and could provide biased coefficient estimates.

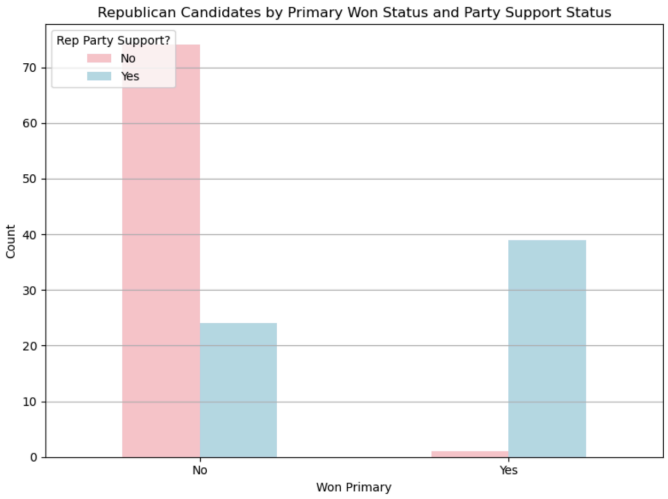
We will also be using random forests as a method to answer this question. They are able to handle multicollinearity and non-linear relationships, unlike logistic regression. If candidate characteristics interact, random forests would be able to handle that well. They are also less prone to overfitting compared to other methods, which would be useful for a large number of predictors.

Though random forests excel in areas where logistic regression is lacking, its limitation is not having easily interpretable results. Though it provides accurate predictions, it can be difficult to understand the importance of each predictor variable. This is why we will be using both types of models to answer this question.

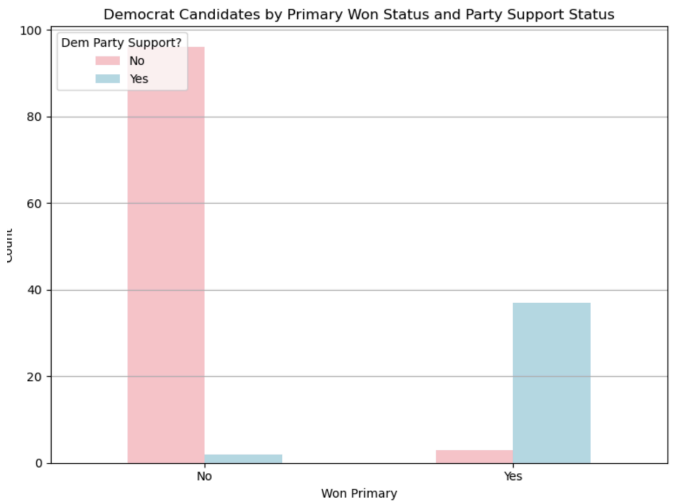
EDA

Research Question 1

Figure 1: Party support status and primary election outcome (Categorical)



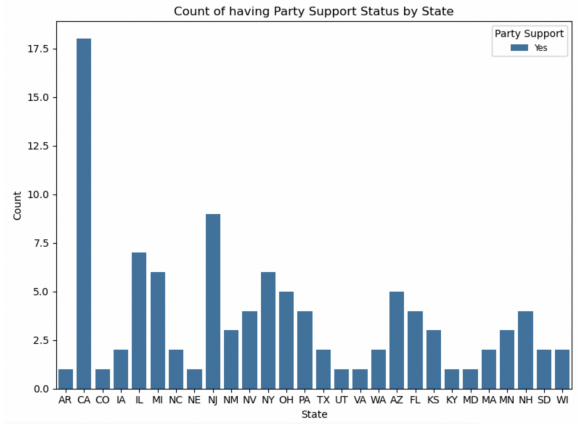
(a) Primary Outcome and Party Support Status for Republican Candidates



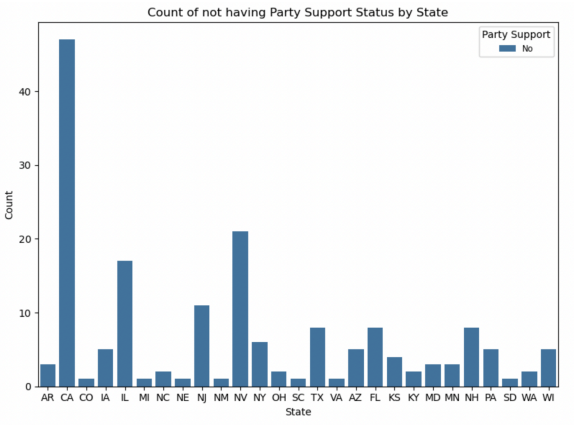
(b) Primary Outcome and Party Support Status for Democrat Candidates

The visualizations of the primary election status based on whether a specific candidate had support from their party seem to indicate that there is an association between whether or not a candidate has party support and their performance in the primary election. For both the Democrat and Republican candidates, a large proportion of those who won their primaries were supported by their respective parties, as can be seen in the bar chart. Out of those who did not win their primaries, a very large proportion were not supported by their parties. These visualizations motivate our first research question on whether receiving party support leads to a change in the likelihood of success in the primary elections and even seem to suggest a potential answer in that having party support does correlate with higher success in the primary elections for candidates. However, further causal inference analysis will be necessary to determine if this correlation is in fact also a causal relationship.

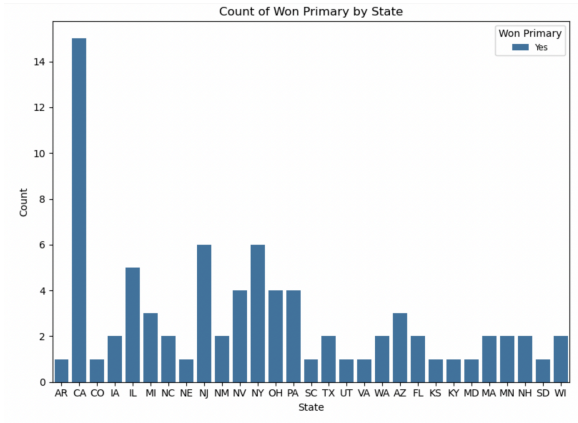
Figure 2: Party Support by State and Won Primary by State (Quantitative)



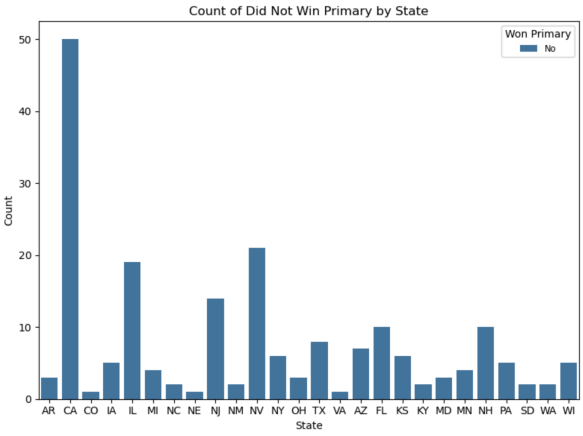
(a) Countplot of having party support by state



(b) Countplot of not having party support by state



(c) Countplot of winning the primary by state



(d) Countplot of not winning the primary by state

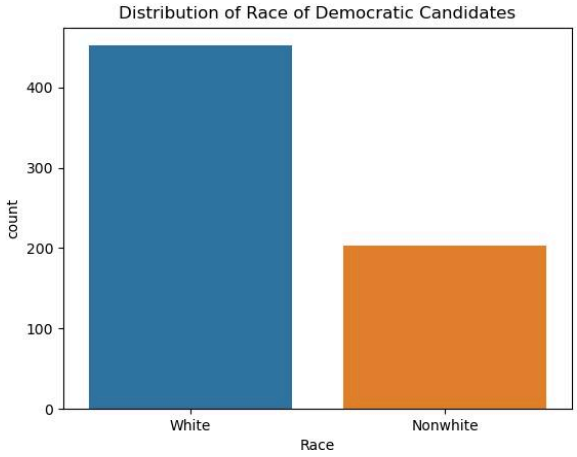
The visualizations above depict the relationship between the variable “State” in the primary candidates data set and the variables that we are using for our treatment and outcome (Party Support and Won Primary respectively). The count plots of states by different party support statuses are in the top row, and the count plots of states by different “won primary statuses” are in the bottom row. An interesting trend to notice in the visualizations above is that there seems to be a different distribution of state by party support status between the top left and top right visualizations. There seems to be a different shape in the wave’s downward and upward trends especially around states like NY and OH (OH count seems relatively low on the upper right visualization but relatively high on the upper left), which motivates the idea that state has a potential effect on the treatment: party support status. There is a similar trend in the couple of visualizations on the bottom row with there being a different distribution between Won Primary

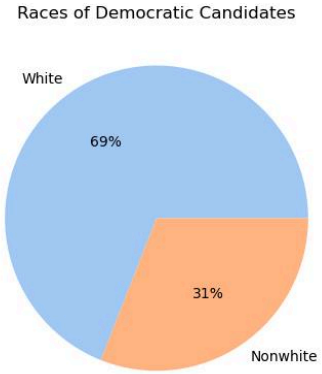
by State and Did Not Win Primary by State, indicating possible effect between the state variable and the outcome.

These visualizations are relevant to our first research questions because they motivate a possible argument for making the state variable a confounding variable when carrying out our causal inference. The state variable seems to have a small effect on both the treatment and the outcome. We may want to follow up on this correlation when calculating our propensity scores since it indicates to us that the state variable could be a potential confounder in our causal inference.

Research Question 2

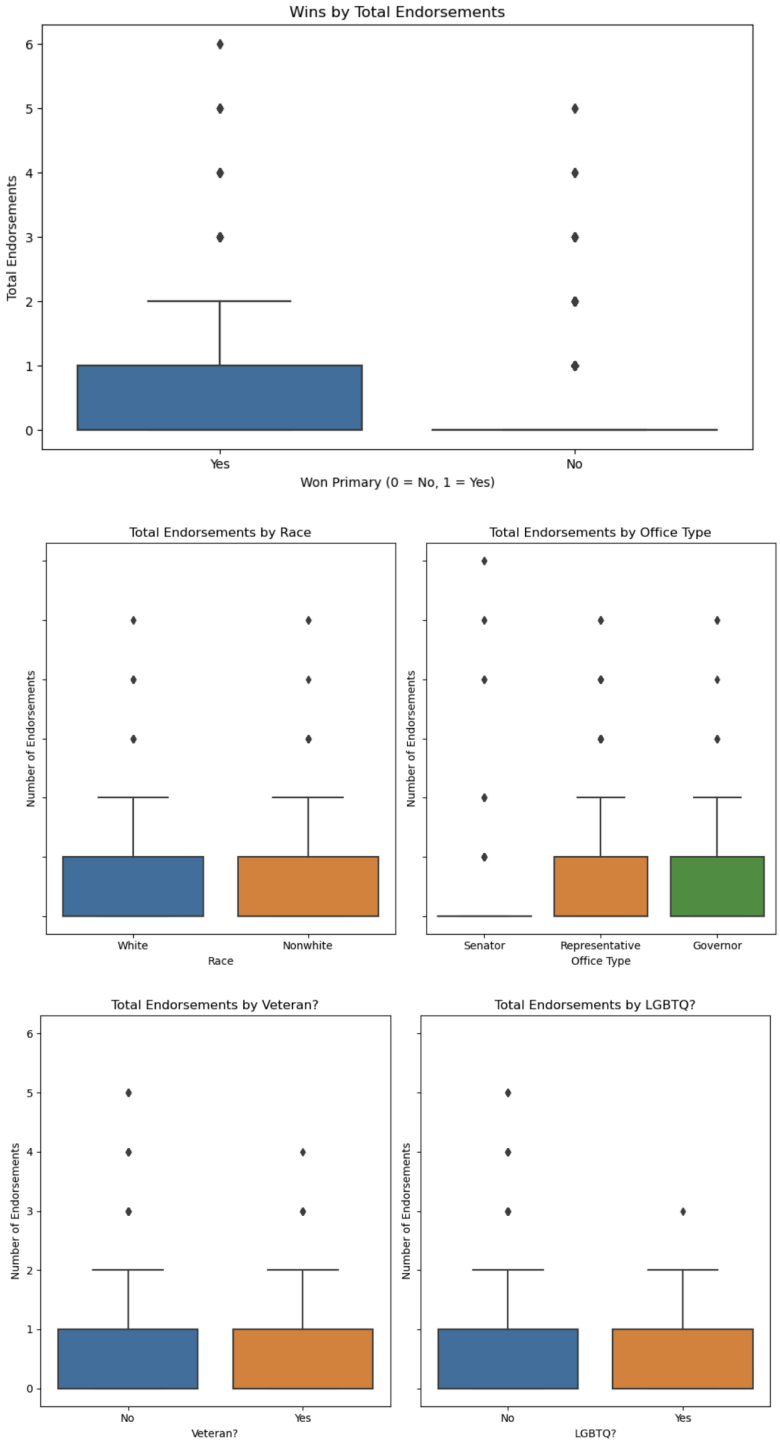
Figure 3: Race (Categorical)





The variable being analyzed is race and the dataset contains that information for only the Democratic candidates. From the bar chart in Figure 3, we can see that there are 452 candidates identified as White and 203 identified as Nonwhite which means that 156 out of the 811 Democratic candidates aren't assigned a race. From the pie chart in Figure 3, we can see that 69% of the Democrats who are assigned a race are White and 31% are Nonwhite meaning that in our dataset of Democrats, the majority of candidates are White. For our second question, we are trying to predict endorsements based on the characteristics of a candidate and one of those characteristics could be race. Given the distribution of race we see on the two plots above, it would be interesting to see if race plays an important role in our decision trees when determining whether a candidate gets endorsed or not. Furthermore, when we fit our logistic regression model the coefficient of race could indicate some deeper systemic questions to be explored.

Figure 4: Number of Endorsements (Quantitative)



The histograms comparing the number of endorsements with primary election outcomes suggest a trend where winners tend to receive more endorsements than non-winners. This relationship highlights the importance of political endorsements in election outcomes. The histograms contrasting the number of endorsements across different characteristics also reveal some trends. While the distributions are generally similar across categories, there are some notable variations within certain groups. Veteran candidates have higher outliers in endorsement counts, indicating a potential advantage in endorsements. Similarly, candidates not identifying as LGBTQ also have higher outliers, suggesting a disparity in endorsement acquisition between these groups.

These visualizations are relevant to our research question. The histograms all provide insights into the factors that influence endorsement decisions. Understanding these relationships can influence the strategies of candidates seeking endorsements of specific organizations. For instance, candidates with characteristics associated with higher endorsement counts, such as military service, may be more appealing to organizations. The visualizations display the importance of endorsements in election success and also offer insights into predictors for endorsements.

Methods, Results, and Discussion

Question 1: Does party support status cause a change in the likelihood of success in primary elections? (Causal Inference)

Methods

Our treatment variable is whether or not a candidate is supported by their own party, which is stored in the "Dem Party Support?" and "Rep Party Support?" variables in the Primary Candidates 2018 data set. Our outcome is the primary election outcome which is stored in the "Won Primary" variable.

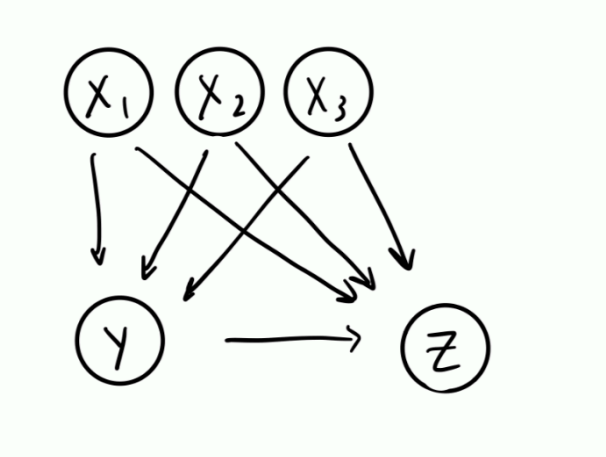
We use the variables "State", "District", and "Office Type" as confounders. This is because these variables could have an effect on both our treatment (whether or not a candidate is supported by their own party) and our outcome (whether or not a candidate won their primary).

We will use inverse propensity weighting to adjust for confounders. To do this, we will calculate the propensity scores for each of the variables "State", "District", and "Office Type", and use these to adjust our calculation of the average treatment effect.

There do not seem to be any colliders in this specific data set. All the variables in the data set seem to be related to attributes of the primary election/primary election candidates before the outcome of the primary, which would mean that the outcome of the primary should not have an effect on any of the other variables and thus should not contribute to the formation of any colliders.

Causal DAG:

X1 = State
X2 = District
X3 = Office Type
Y = Party Support Status
Z = Primary Election Outcome



Assumptions:

1) There is a consistent impact of party support status on the primary success of all candidates.
2) State, district, and office type are the only confounders in this dataset.

Results

After using inverse propensity weighting to adjust for confounders (state, district, and office type), our average treatment effect came out to be 0.7190. Since the ATE was positive, the treatment (party support status) had a positive causal effect on the outcome (success in primary elections). This of course assumes that state, district, and office type were the only confounders in this data set.

Since there was a fairly small proportion of candidates that actually had information on whether or not they had the support of their party, our estimate was based on only 276 candidates compared to the full 1585 candidates that were originally present in the data set. This could contribute to some uncertainty in our estimate as our data was less robust than if we had considered more candidates, but it was an unfortunate consequence of our data and research question that we did not have such a robust set of data to work with.

Discussion

We were limited in our methodology due to missing values within our treatment columns, as only 276 out of the full 1585 candidates had their party support status available. This would be considered convenience sampling since the dataset only collected information if it was publicly available, and we proceeded to only use this data. The small proportion of party support column data limited the generalizability of our findings. Additionally, there were not enough common columns between the Republican and Democrat datasets to use as confounders, so we were limited to using three confounders. This means that there is a possibility of other confounding variables that we were unable to account for in our analysis.

Additional data on past years could be helpful since this data set only contained data from 2018. This would help improve the robustness of our ATE since we had to cut down on a lot of data since many candidates had null values for whether or not their party supported them. It would also be helpful to analyze data sets with other additional confounders to improve the credibility of our ATE since our confounders were limited to variables that were in both the Republican and Democrat data sets.

We are confident to some degree that there is a causal relationship between whether a candidate received support from their party and the primary outcome. Through our method of inverse propensity weighting, we found that the treatment effect on the primary election success was relatively positive and high at about 0.719, which allowed us to conclude that there is in fact a causal relationship. However, we were limited due to our sample size, because only a small proportion of the candidates had data available about the party support status. This introduces some uncertainty in our conclusion because there are reservations in generalizing our sample to the overall population. Additionally, there could have been other confounding variables that we did not account for.

Question 2: Can we predict whether or not an organization will endorse a democratic candidate based on candidate characteristics? (Prediction with GLMs and nonparametric methods)

Methods

We are trying to predict whether or not an organization will endorse a candidate based on candidate characteristics such as partisan lean, race, whether or not they are a veteran, and whether or not they are lgbtq. We chose these because organizations have inherent biases, and the members of the organizations may choose who to endorse based on their race, veteran status, partisan lean, and lgbtq status depending on their own personal views on those topics.

We will be using logistic regression to model the relationship between an organization endorsing a candidate using the candidate's characteristics. We chose logistic regression because it is an efficient way to obtain the likelihood of a binary outcome variable. For this question, our binary outcome variable will be whether or not an organization endorses a candidate, and we are using predictor variables like partisan lean, race, and veteran status to predict the likelihood of the event occurring. The assumptions being made are that the observations are independent of one another and that there is little to no multicollinearity between the predictor variables. Additionally, we are assuming that the relationship between the log odds of the outcome variable and the predictor variables is linear.

We used random forests because by combining multiple decision trees, we are able to model many different decision boundaries which gives us flexibility when handling the classification task of determining whether an organization will endorse a candidate based on the given candidate characteristics. This flexibility is helpful in our case because we have many different features and could potentially add more in the future. The randomness of the random forests method prevents overfitting on these many different features and as a result, provides more accurate organization endorsement predictions towards candidate characteristic data. The random forests method assumes that the candidate characteristics data in the dataset has actual values in it, and each tree in the random forest is made independently of the rest.

To evaluate each model's performance, we will consider multiple metrics. One such metric is accuracy, which is the most basic metric for classification tasks. It gives a quick sense of how well the model did at correctly predicting endorsements. We will also look at the F1 score, which combines precision and recall into one metric. This will also give us a good idea of how well-balanced our model is in terms of both.

Assumptions:

1) The observations are independent of one another.
2) There is little to no multicollinearity between the predictor variables.
3) The relationship between the log odds of the outcome variable and the predictor variables is linear.

Results

The logistic regression model looked at whether a candidate received support for being a Gun Sense candidate. For the partisan lean, the coefficient was 0.0019, implying that for every one-unit increase in the partisan lean score, the log odds for the candidate being a Gun Sense candidate increased by 0.0019. For race, the log odds increased by 0.518 for white candidates. For veterans, the log odds decreased by 0.558. Finally, for LGBTQ candidates, log odds decreased by 0.143. Overall, this indicates that race and veteran status have the most magnitude of effect. Being a veteran resulted in a considerably lower chance of being endorsed by the Gun Sense organization, while being white resulted in a considerably higher chance of being endorsed by the Gun Sense organization.

There are several measures of uncertainty we can look at. For one, the 95% confidence interval around the coefficients. For the LGBTQ variable in particular, the interval is pretty wide, [-1.293, 1.007], indicating substantial uncertainty for this coefficient's estimate. Other metrics are the accuracy and f1 scores. While accuracy is at 45%, the f1 score is fairly low at 0.5. This indicates that the model has low precision/recall.

Discussion

The model that performed better in terms of accuracy on the test set is the random forests model with an accuracy of 62% while the regression model had an accuracy of 45%. The random forests model performed better in this case because we mainly used variables that don't necessarily have a linear log odds relationship with the endorsement we were predicting. If we used the same variables to predict the endorsement then it would still be a good idea to use random forests but if we increased the number of predictor variables by a lot and used variables that are more linearly related to the endorsement then it's probably a better idea to use logistic regression since it would be computationally more feasible. We are semi-confident in applying the random forests model to future data sets since while it has a higher accuracy than the regression model, 62% accuracy is still not super high compared to industry standards.

The random forests model fits the data very well with an accuracy of 86% on the training data while the logistic regression model has an accuracy of 56% on the training data. This is to be expected due to the fact that random forests are very good at fitting the training data while avoiding overfitting through random feature selection and sampling randomness.

It is difficult to make general statements about the relationships between the outcome and features due to the interpolation of data that was performed but if we assume that the models we created do follow the same trend as if the data was complete then there are a few notable relationships. For the logistic regression model, we saw that the variable with the highest absolute coefficient was veteran status. It had a negative correlation coefficient meaning that being considered a veteran decreased the log odds of being endorsed as a Gun Sense Candidate. There may be reasoning behind these values since Gun Sense values gun safety and thus stricter gun policies this may contradict heavily the beliefs many veteran candidates hold and publicly advocate. This contradiction would decrease the chance of a veteran being endorsed as a Gun Sense candidate.

The main limitation of both models would simply be the amount of available data points since there were many null values that we had to account for across the predictor variables as well as the endorsement variable. For the logistic regression model, we may also be experiencing issues with collinearity since we can see that the intervals for some of the variables include both positive and negative values. Furthermore, the logistic regression model may be struggling due to a lack of linear relationship between the log odds and predictor variables. A limitation of the random forests model comes from the fact that all of the predictor variables are binary except for the partisan lean variable meaning that the importance of the partisan lean variable is very high.

Additional data that would be helpful in improving the models would be more complete data in general for all the variables we used. Having more up-to-date and complete information on which candidates were endorsed by the Gun Sense group would help prevent biases in the

models. Furthermore, more complete data on the predictor variables could lead to more confidence in the true values of coefficients.

The accuracy for both models is relatively low and this is especially the case with the logistic regression model. This is the case due to the fact that we had to drop a lot of entries in our data due to having null values in at least one of the variables we were using. Furthermore, we can see that the confidence intervals for some of the variables in our regression model are relatively even on both sides of 0 meaning their importance in prediction is called into question. This probably has to do with issues of collinearity or noisy data. This prediction problem is relatively hard since there are a lot more variables that are involved in whether or not a candidate will be involved that aren't present in the dataset and it is also difficult to determine the relationship between the predictor variables which is key to know in logistic regression.

Conclusions

From our study, we found a causal relationship between a candidate's party support status and their success in the primaries. We also predicted whether or not a democratic candidate got an endorsement from a gun sense organization based on their characteristics, and found that the non-parametric random forest model was better in terms of accuracy. Notably, being a veteran resulted in a considerably lower chance of being endorsed by the Gun Sense organization.

The generalizability of our results is limited to candidates within the 2018 U.S. primaries, but could potentially be generalized to other years in which the primaries took place. For our second research question, our findings are limited to the democratic party candidates, because there was no available information about the gun sense endorsements in the republican party.

We found that there is a positive causal relationship between party support and success in the primaries. Based on our results, we would suggest future candidates in the primaries to prioritize a large part of their campaign to get support from their party, because it would increase their chances of winning the primaries. Based on the results from our second question, we would also recommend candidates of diverse racial backgrounds and veterans to focus on gaining support from organizations besides Moms Demand Action (the group behind the Guns Sense Candidate Distinction). We did not merge different data sources for our analysis, but doing so would have decreased the limitations of our data by providing additional information or filling in missing values. Some limitations of our data were that the Democrat and Republican datasets did not share many features in regard to candidate characteristics, which made it difficult to determine confounders for our first question and answer general questions about the candidates as a whole. Additionally, there were many missing values within the dataset, which introduced uncertainty within our causal effect and made it more challenging to make accurate predictions of endorsements.

Based on our work now, it would be interesting to explore whether there is a difference in the strength of the causal relationship that we found in question one between the Democratic and Republican parties, because the issues that voters prioritize may be different between the two parties. It would also be interesting to delve deeper into the shared characteristics of the candidates between the two parties and explore how to predict whether a candidate gets a specific endorsement regardless of their party affiliation. Working through the project allowed us to better understand the strengths and differences between GLMs and non-parametric methods when attempting to predict an outcome. We also explored the importance of confounding variables when making causal inferences, and discovered the complexities of working with and interpreting data that had a significant amount of missing values. Based on the causal relationship between party support and primary success and the model we created for predicting whether a candidate will receive a Guns Sense Candidate Distinction based on candidate

characteristics, our call to action would be to use these results to advise a candidate on how they can maximize their chances of winning their next primary by pushing for them to gain support from their respective party and advising them to target/not target endorsement from Guns Sense based on their own characteristics.