Mark Cheung

Marlon Fu

Tim Wang

Sean Wei

Data 102 Final Project

1. Data Overview

We chose to use the FiveThirtyEight primary-candidates-2018 dataset. Specifically, we chose the subset of Democrat candidates, as the data for the Republican candidates only contained information about endorsements, and not the candidates themselves.. On the other hand, the data for the Democrat candidates contained both demographic and endorsement information. Our dataset is a census of all Democratic candidates who appeared on the ballot in Democratic Primaries for Senate, House and Governor as of August 7, 2018.

As the dataset's README notes, it systematically excludes Democrat candidates who participated in races featuring a Democratic incumbent. Since all of the data in this dataset is publicly available and election analysis is a large industry in the United States, the participants in this dataset must have been aware of the collection and use of this data. Our dataset is at a granularity level of candidates, meaning that each row represents a Democrat candidate who has appeared on the ballot. This fine level of granularity enables us to examine how personal factors like demographic and endorsement impact the chance of a person winning the primary.

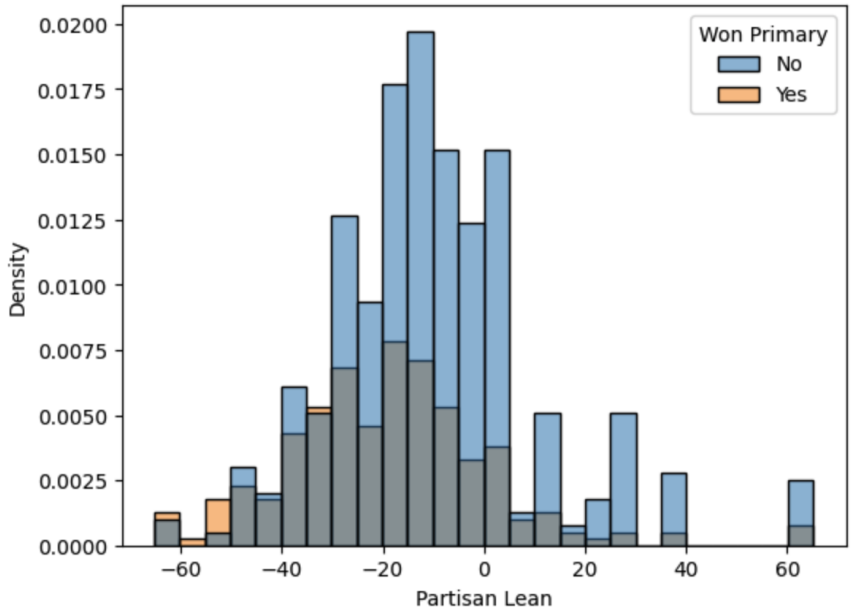
We wish that the Republican dataset contained the same demographic information in the Democrat dataset. Had those columns existed, we could have examined if the same factors would have a different effect on the chance that candidates would win the primary in different parties.

For example, we could see if receiving party endorsement is a more important factor in determining if a Democrat candidate wins the primary compared to a Republican candidate.

2. Exploratory Data Analysis

A. Prediction with Bayesian GLMs Versus Nonparametric Methods

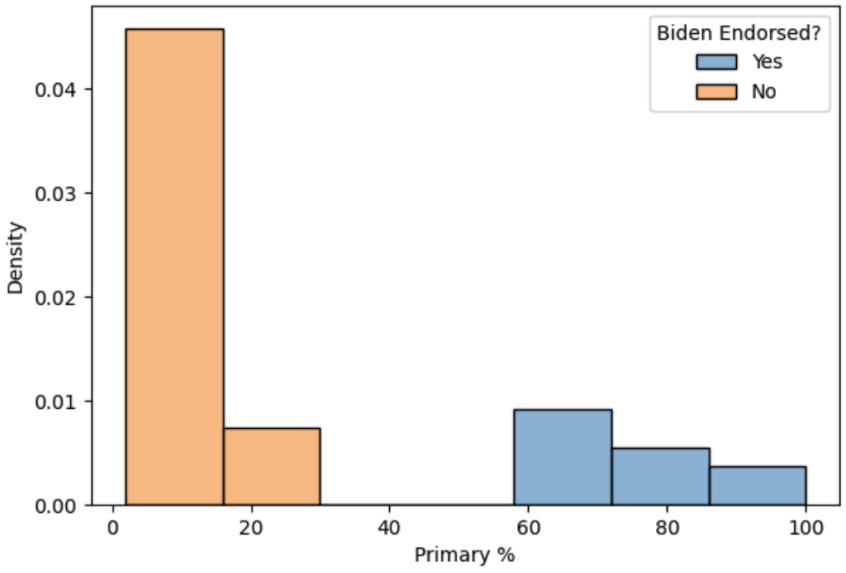
We want to create both a GLM and a Nonparametric model to predict whether a candidate won or lost their primary election. In order to build our models, we must find some features that are associated with receiving/not receiving votes in the primary election. To begin, we checked to see if the Partisan Lean of the district or state of the election (the only quantitative variable included in the dataset) has any relationship with whether or not the Democratic candidate was able to win. According to FiveThirtyEight, Partisan Lean is calculated as the following: "Partisan leans are calculated by finding the average difference between how a state or district voted in the past two presidential elections and how the country voted overall, with 2016 results weighted 75 percent and 2012 results weighted 25 percent."



	Won Primary ('Yes')	Won Primary ('No')
count	239	550
mean	-19.123	-10.450
std	20.377	20.540
min	-65.209	-62.480
25%	-32.269	-21.680
50%	-18.959	-13.230
75%	-8.33	-0.405
max	65.089	65.089

Based on the histogram that we produced, we can see a slight difference in distributions in terms of partisan lean between candidates who won the primary, and candidates who didn't. The average partisan lean of the candidates who won the primary (-19.126) was far more negative than the average of those who did not win (-10.450). Overall, we can also see that the former distribution lies slightly to the left (or in other words further the negative direction in terms of partisan lean) compared to the later, as observed by the 25th, 50th, and 75th quartiles both being lower.

Next, we checked to see how the distribution of percentage votes received by candidates changed depending on various categorical features. We looked to see if there was a change in the distribution depending on whether they were endorsed by Joe Biden.



It is evident that there is a clear difference in distributions when we split candidates by this feature, as all Biden Endorsed candidates received higher percent votes compared to any of the non-Biden Endorsed candidates. Having perfect divide by splitting based on this feature

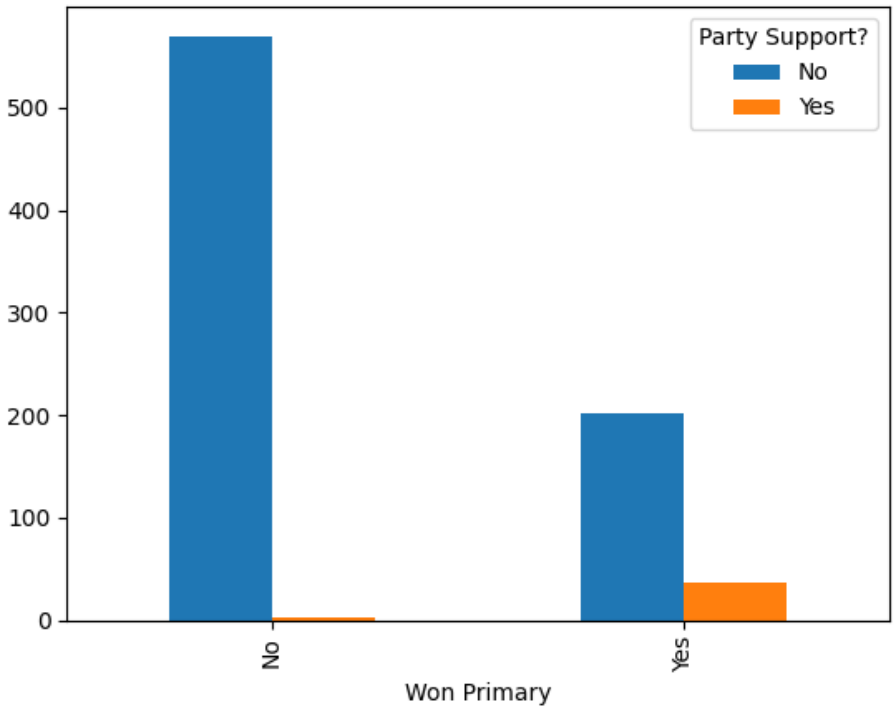
suggests that it is a good feature to use in our modeling process, especially in the case of decision tree models since it minimizes the entropy. Thus, we can consider whether a candidate has been endorsed by Joe Biden to be a good predictor of whether they won the primary election.

B. Causal Inference

I. Data Cleaning

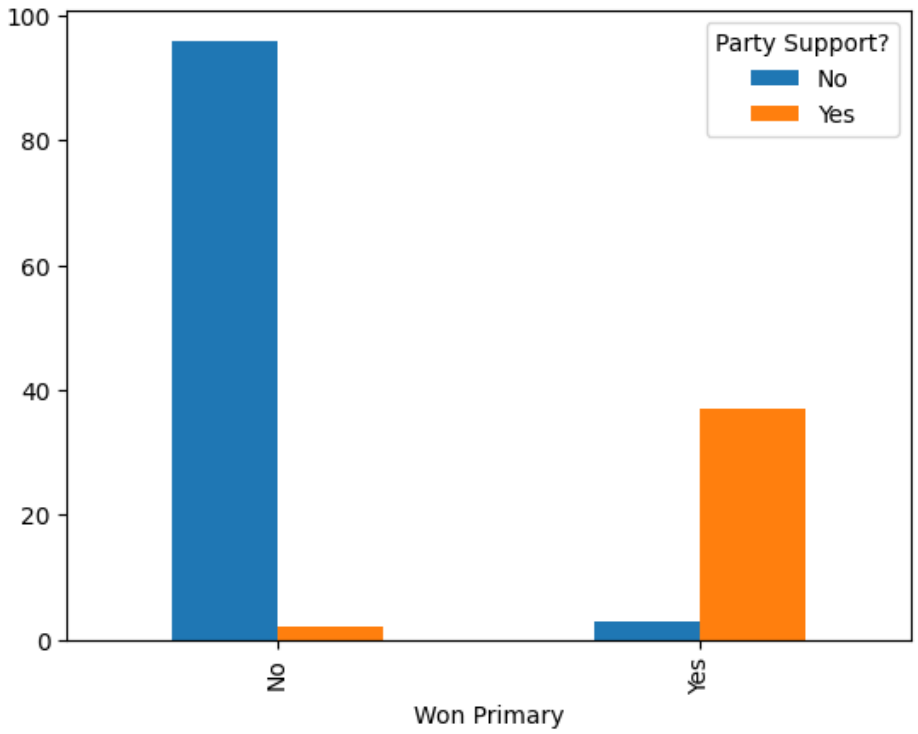
After EDA, we decided to drop the candidates who participated in races where the Democratic Party endorsed no one. This amounts to dropping rows that have "NaN" values for Party Support, which we planned to be our binary treatment. We believe that not having the endorsement when no one else does is a different sort of treatment than not having the endorsement when someone else does. Thus, we choose to exclude those observations.

Had we included those rows, the causal relationship between receiving an endorsement and winning the primary would not be accurate. This is because the party only endorsed candidates for a minority of races (41/245). As a result, a majority of people who won the primary did not have the support of the Democrat Party. Leaving these rows in would distort the causal effect between our treatment and outcome.



II. Does there seem to be a causal relationship?

To see if our research question is worthy of investigation, we visualized the relationship between two categorical variables, **"Won Primary"** and **"Party Support"** in the Democrat dataset, which are the outcome and treatment respectively.



We observe a few trends from these visualizations.

1. The overwhelming majority of Democrat candidates who won the primary have the support of the Democrat party.
2. The overwhelming majority of Democrat candidates who did not win the primary did not have the support of the Democrat party.

The visualization is relevant to our research question as it suggests that the causal relationship between the treatment and outcome exists. It motivates us to investigate our research question further.

3. Research Questions

A. Can we use GLMS and Nonparametric methods to predict which candidate won the Democratic Primary for each election in 2018? Generating a model to predict candidate performance can be useful for news companies in their own analysis and predictions of elections. Meanwhile, politicians and their teams can also analyze the results of the model and its features to understand which endorsements and characteristics can help boost their candidacy. GLMS and Nonparametric methods are the best fit for this research question because we can use classification to try and effectively predict the outcome of an election based on the features provided in the dataset.

B. Does being endorsed by the Democratic party have a causal relationship with a Democratic candidate winning the primary election? The answer to this research question will shed light on the value of having party endorsement to a Democrat candidate. Democrat candidates can use this information to decide how much time they want to spend to try to increase the chances of them receiving an endorsement from the party. Since we specifically want to determine if a causal relationship exists between the two variables, causal inference is uniquely suited to answering this question.

4. Inference and Decisions

A. Prediction with GLMs Versus Nonparametric Methods

I. Methods

In this study, we aim to predict whether a candidate won or lost in a primary election, expressed in the column ‘Primary Status’, using demographic and endorsement data included in

the dem_candidates.csv dataset. These features include but are not limited to race, whether the candidate is a veteran, whether the candidate identifies as LGBTQ, and whether the candidate is endorsed by Joe Biden, Bernie Sanders, or Elizabeth Warren. These features were manually selected based on the EDA which showed that they may be good predictors of our target variable. Meanwhile, features such as each candidate's state, district, and office type, which were more helpful for identification but not so much as predictors, were omitted.

We discovered that much of the demographic data took the form of binary variables, with some missing null values sprinkled inside. For demographic data, such as whether or not the candidate identified as LGBTQ or their gender, a null value represented missing data, which only constituted 1% of our total data. As a result, these variables were simply converted to be boolean values for our model. However, when it comes to the variables representing political endorsements, null values did not represent data that is missing, but rather the fact that a particular political figure did not make an endorsement within that primary race. As a result, null values actually carried meaning, and thus, our feature engineering involved one-hot encoding these categorical variables.

For our nonparametric method, we trained a random forest model for binary classification in order to predict whether or not a Democratic candidate won their primary election, with a prediction of 1 representing a victory and 0 representing a loss. Out of the two tree models we explored in this course, random forests are less likely to overfit than decision trees due to their ensemble nature and use of bootstrap sampling. As a result, using a random forest allows our model to be better applied to data that it was not trained on, such as for future elections. Random forests do not make any formal assumptions about the data, as part of their nonparametric nature,

but typically, it is best if the data is representative of the population from which it was drawn and has relative features.

In contrast to our nonparametric method, we chose to train a binary logistic regression GLM using Scikit-learn and PyMC3 for our frequentist and Bayesian methods respectively. Since our target variable was binary, a logistic regression model worked best in predicting binary outputs based on both continuous and categorical features. Similar to frequentist logistic regression, we assume in Bayesian logistic regression that 1.) the observations in our dataset are independent of one another, 2.) the independent variables are not highly correlated, and 3.) independent variables are linearly related to the log odds. In addition to these, another key assumption made in Bayesian logistic regression but not in frequentist logistic regression is that the likelihood function we choose is a reasonable representation of the data. Using PyMC3, we start the process of modeling by writing out the formula in terms of the features we choose to use (which are the same as the features we used for the random forest model). We then use PyMC3's GLM.from_formula function to convert the formula into a working model.

We measured our model's performance using accuracy. The consequences of a false positive are about equal to the consequences of a false negative, as in both cases, we simply predict the wrong result of the election beforehand. Historically, election predictions are known to be unreliable at times, so the consequences of falsely predicting the results one way or the other would not be drastically different from one another. As is the case, using metrics such as recall or precision, which minimize false negatives and false positives, respectively, would not be as appropriate as simply using accuracy. We will look to maximize the accuracy of our models on the validation set.

II. Results

Model	Validation Accuracy
Random Forest (Nonparametric)	0.70375
Logistic Regression (Frequentist GLM)	0.79844
Logistic Regression (Bayesian GLM)	0.795

It appears that all our models perform equally well for this task when given the same features, and get the prediction incorrectly about one out of five times. In assessing our Bayesian GLM predictions, we produced a summary displaying uncertainty in the estimated coefficients for 4 of our features.

	Mean	SD	3% HDI	97% HDI
Intercept	-1.46	0.29	-1.97	-0.94
Partisan Lean	-0.04	0.01	-0.05	-0.02
Veteran	-0.51	0.32	-1.03	0.12
LGBTQ	-0.14	0.44	-0.93	0.74
Elected Official	0.82	0.29	0.35	1.43
…	…	…	…	…

For Partisan Lean, we have a mean of -0.04 with a standard deviation of 0.01, meaning the regression coefficient is anywhere between -0.05 to -0.02 with 94% confidence. This range is quite small, which suggests that we are relatively confident in Partisan Lean being a good predictor of the outcome of the primary election. In contrast, other features have a wider confidence interval. For instance, the feature LGBTQ has a mean of -0.14 and a standard deviation of 0.44, meaning the regression coefficient is anywhere between -0.93 and 0.74 with the same confidence of 94%, which is a much wider range than the Partisan Lean feature. This discrepancy suggests that there is a higher level of uncertainty with the LGBTQ feature being a good predictor for the outcome of the primary election.

III. Discussion

All models (the nonparametric random forest, the frequentist GLM, and the Bayesian GLM) are interpreted in the same way, as they all produce outputs indicating whether a candidate is predicted to win an election. These models performed equally well when it came to predicting the correct winners of the 2018 Democratic Primary when given the same data. Because they saw the same features, the models were likely to identify the same trends when it came to predicting the election results. We are fairly confident when it comes to applying these model to future datasets, as our model only yielded an incorrect prediction for every one out of five candidates. This rate is relatively impressive when you consider that most primary elections have more than two candidates. However, it is important to note that the election in our dataset was from 4 years ago, and voter behavior, preferences, and trends have likely changed since that last election. As a result, it is very likely that our models will not perform as well when used to predict future elections. We can also likely expect the model to perform worse and worse the farther we go into the future.

In terms of our GLMs, the key difference between our frequentist and Bayesian implementation lies in how each model measures uncertainty in parameter estimation. While the frequentist method uses point estimates, the Bayesian model yields probability distributions as seen in the previous results section. In addition, the Bayesian model enables us to define priors to bolster the accuracy of the model, while the frequentist does not take into account any prior beliefs for parameter estimation.

One of the limitations of random forests is that they are black box models, meaning that it is difficult to interpret how the model makes predictions, and for understanding which features are the most key. Random forests are also very computationally complex, due to the fact that they train multiple decision trees on a bootstrapped sample of the data. As a result, they take a lot of time and energy to train.

For the Logistic Regression GLM, one of the most major drawbacks is that only linear decision boundaries can be drawn, meaning non-linear relationships in the data cannot be effectively captured. Additionally, these GLMs are limited to binary classification, and are not able to perform multi-class classification. This can impact the context of our research question in the future, as some states are beginning to implement ranked-choice voting. Finally, these GLMs are also very sensitive to outliers, although this would not impact our case very significantly, as almost all of the features in this dataset are booleans.

While our dataset offers important insights when it comes to the demographics of a particular candidate, some of their political stances, and the endorsements they did and did not receive, there are numerous other factors that have historically been key towards deciding the results of an election. These factors include voter turnout, the campaign strategy and overall funding a candidate had, the voter demographics of a candidate's district or region (including

age, income, gender, race), and stances on key political issues which were not present in our dataset. As a result, there is a lot of additional data which would be useful towards improving the models.

B. Causal Inference

I. Methods

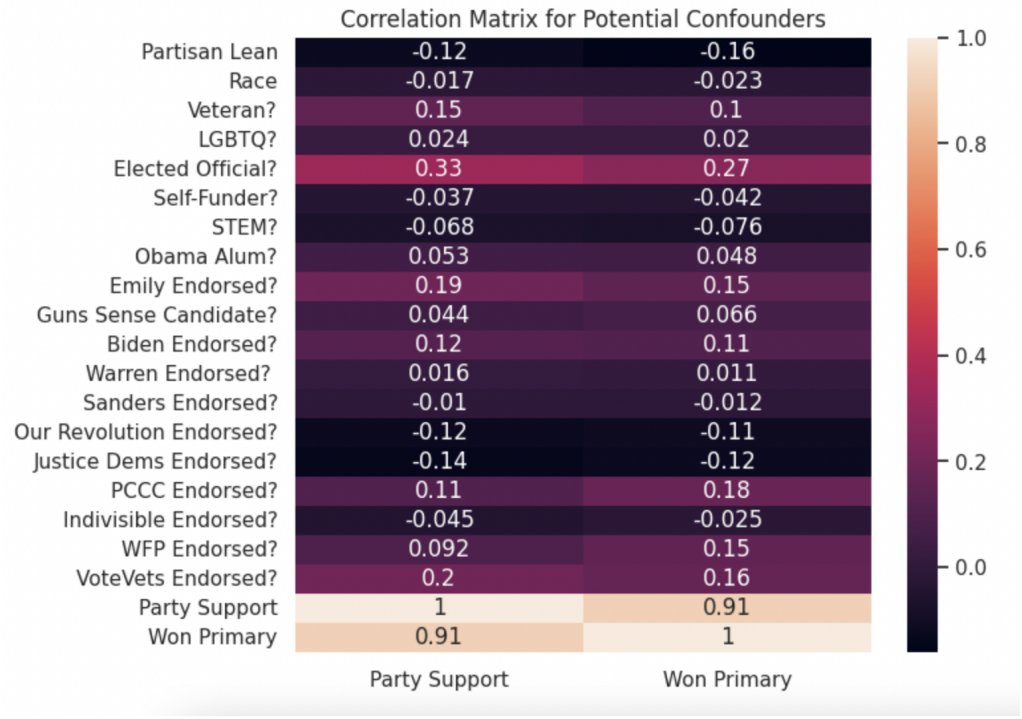
In our causal inference, the treatment is receiving endorsement from the Democratic Party (as explained in the Data Cleaning section of our EDA). The outcome is whether or not the candidate won the primary election for the race they were participating in.

Does the unconfoundedness assumption hold?

The unconfoundedness assumptions hold in this case because regardless of receiving Party Support, the possible outcomes for a candidate is still (Win, Lose). As we have shown in our EDA, there are candidates who won the primary without Party Support and candidates who lost the primary with Party Support.

Describe which variables correspond to treatment and outcome?

We first identified 19 different features that could be potential confounders based on domain knowledge. In order to ensure that these variables had an actual effect on both the treatment and the response, we examined the correlation coefficients between each feature and the treatment and response. A linear causal relationship would be indicated by a significantly non-zero correlation coefficient.



We set an arbitrary cutoff of |0.1| correlation in considering if a feature had a significant non-zero correlation with the treatment or the response - and after doing so, only 9 of the 19 features met that benchmark for both the treatment and response (Partisan Lean of the relevant state/district, whether the candidate is a Veteran, whether the candidate was an Elected Official, and whether they were endorsed by Emily, Biden, Our Revolution, Justice Dems, PCCC, or VoteVets). A stricter method of determining causal effect of confounding variables would be to bootstrap these correlation coefficients and create a confidence interval - see limitations section.

We adjusted for these confounding variables using Inverse Propensity Weighting, which weighs the outcome of each datapoint by the probability that it would receive the treatment. This accounts for the fact that the distribution of our confounding variables may not be the same for all candidates and treatment groups. We obtained these probabilities, also known as propensity scores, by training a random forest model on our 9 confounding variables to classify whether

each candidate received Party Support vs No Party Support. Using predicted probabilities for each of the candidates, we calculated a weighted average difference in winning based on Party Support (also known as the average treatment effect) using the following formula, where *Y* is a binary 1 or 0 for winning/losing the Primary, *e(X)* is our estimated probability that a candidate would receive Party Support, and *X* is our list of confounders.

$$\tau_{IPW} = \frac{1}{n} \sum_{i:Z_i=1} \frac{Y_i}{e(X_i)} - \frac{1}{n} \sum_{i:Z_i=0} \frac{Y_i}{1-e(X_i)}$$

We don't believe there are any colliders in our set of confounders. While it is possible for certain features, such as endorsements, to be spurred after hearing that the Democratic Party supports a candidate, it is not possible for any of our confounders to be caused by winning the primary. This is because our confounders precede the ultimate result of the election - for example, endorsements, veteran status, etc.

II. Results

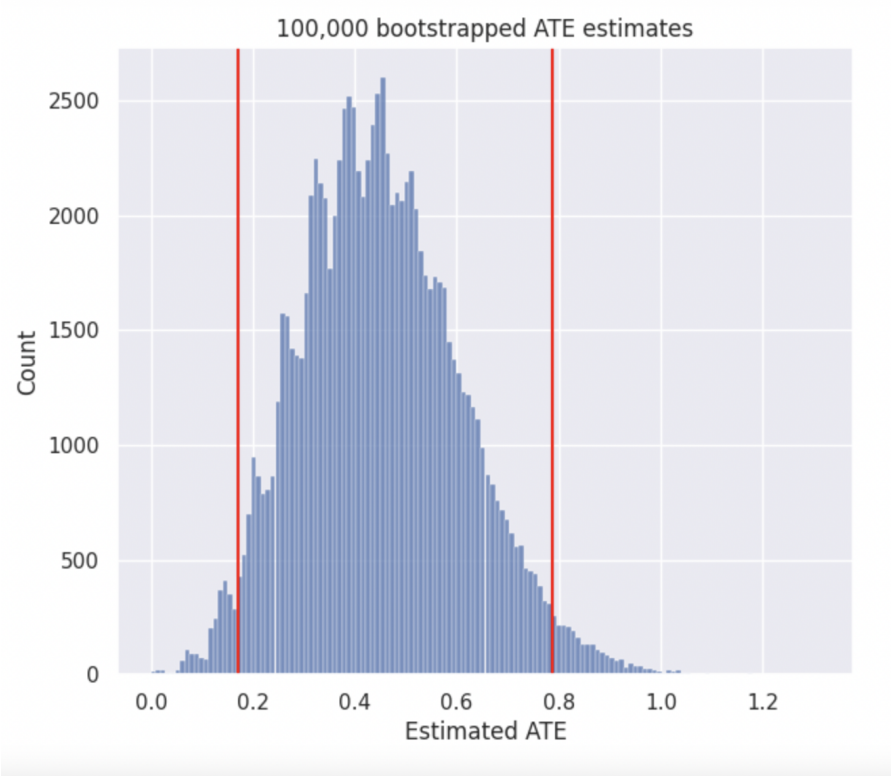
To assess the causal relationship between Party Support (treatment) and Winning the Primary (outcome), we estimated the average treatment effect using Inverse Propensity Weighting. In our case, the average treatment effect is the expected difference of the chances of winning based on whether the candidate had party support or not. A difference of zero would indicate that the chances of winning are equal regardless of party support, indicating no causality. On the other hand, a nonzero average treatment effect of δ means that **on average**, the difference between the chance of winning with party support and the chance of winning without party support is δ**,** indicating a causal relationship.

In using Inverse Propensity Weighting, we made the following assumptions:

1. SUTVA
    a. When Party Support is given, it means the same for all candidates.
    b. The success of each candidates' campaign does not affect the success of others.
2. Unconfoundedness: Whether or not you receive the treatment, the potential outcomes will always be winning or losing.
3. Our data includes all significant confounding variables.

After computing our inverse propensity score estimate for τ - which accounted for the 9 confounding variables deemed of significant correlation in the previous section - we arrived at an estimated average treatment effect of 0.451 for our dataset.

We bootstrapped our data 100,000 times to create a larger sample of ATEs. The resulting distribution of ATEs had a 95-percent confidence interval between [0.170, 0.786]. This is an extremely wide confidence interval, implying a high level of uncertainty in our estimate of average treatment effect. Despite its width, the confidence interval does not contain 0. This suggests that the true treatment effect is significantly unlikely to be zero based on our estimate, even if it may be more or less extreme than 45.8%.



Thus, we conclude that, in the Democratic Primary Elections of 2018, Party Support has a positive causal relationship on Winning the Primary.

III. Discussion

<u>Limitations</u>

In this section, we examine the shortcomings of our methodology and how these shortcomings induce uncertainty in our conclusion.

1. **Potential Violation of SUTVA:** In our data, a candidate is said to have Party Support if they fulfilled one out of three of the below conditions.
    a. The candidate was placed on the DCCC's Red to Blue list before the primary.
    b. The candidate was endorsed by the DSCC before the primary.
    c. The DSCC/DCCC aired pre-primary ads in support of the candidate.

It could be possible that these different conditions correspond to different levels of Party Support. Thus, candidates with party support might not have gotten the same treatment as each other.

2. **Small Sample Size:** Our dataset only contained 138 rows, which isn't ideal.
3. **Confounders not being confounders:** Firstly, correlation coefficients can be flawed inherently because they only measure linear relationships. Furthermore, while our chosen confounding variables had non-zero correlation coefficients with both our treatment and response, after calculating bootstrapped distributions of the correlation coefficients for each confounder we found that many of the 95-percentile confidence intervals contain zero. This would suggest uncertainty in how much our confounders, if at all, have an effect on either of the treatment or response.

Feature	95% CI for Bootstrapped correlation coefficients with **response**	95% CI Contains Zero?	95% CI for Bootstrapped correlation coefficients with **treatment**	95% CI Contains Zero?
Partisan Lean	[-0.348, 0.032]	Yes	[-0.311, 0.068]	Yes
Veteran?	[-0.074, 0.278]	Yes	[-0.029, -0.329]	Yes
Elected Official?	[0.078, 0.444]	No	[0.134, 0.498]	No

Emily Endorsed?	[0.06, 0.341]	Yes	[-0.016, 0.388]	Yes
Biden Endorsed?	[-0.106, 0.289]	Yes	[-0.099, 0.3]	Yes
Our Revolution Endorsed?	[-0.255, 0.043]	Yes	[-0.259, 0.024]	Yes
Justice Dems Endorsed?	[-0.26, 0.027]	Yes	[-0.263, -0.002]	Yes
PCCC Endorsed?	[-0.039, 0.341]	Yes	[-0.08, 0.288]	Yes
VoteVets Endorsed?	[-0.035, 0.343]	Yes	[-0.003, 0.377]	Yes

Additional Data:

Data that would be helpful in answering our research question would be a similar dataset but for the 2020 and 2022 primary elections. If we assume that the "effect" of receiving Party Support has remained consistent for 2018, 2020, and 2012, we could then join the data together to form a larger, more comprehensive dataset. Doing this would allow us to derive an answer with less uncertainty.

Other columns that would be helpful would be when exactly did the candidate receive all of their endorsements. Suppose the Democratic Party released their endorsement for race A and

the Justice Dems released their endorsement for race A much later. In this situation, we do our causal inference using the values of 0s and 1s for the values of the Justice Dems column. What is more appropriate to do, however, is to have the values of the Justice Dems column to be all NaN. This is because at the time that the Democratic Party made their endorsement, the Justice Dems had not endorsed anyone. Thus, the Justice Dems endorsement column should be NaNs, rather than 0s and 1s.

Confidence in our Conclusion:

In the end, we are confident that there is a causal relationship between receiving Party Endorsement and Winning the Primary. To quantify our confidence, we would say that there is an 80% chance that this relationship exists. The 95% confidence interval that we generated for our estimate of causal effect does not contain 0 and is strictly positive. In addition, the visualization we showed in the EDA section showing the distribution of Winning the Primary and Party Endorsement is striking. It strongly hints that there is a causal effect. Finally, as a more meta argument, if endorsements don't have a causal effect on a candidate winning, then it is unlikely that the custom has gone on for hundreds of years in the United States.

5. Conclusion

In conclusion, our first research question showed that we can use both nonparametric prediction methods, specifically random forests, and Bayesian GLMs to predict the winner of the Democratic Primary election based on their endorsements and demographic information. Our predictions were able to project the correct results about 80% of the time for both models. We can also conclude that having party support has a positive causal relationship with winning the 2018 Democratic Primary.

In our second research question, we used causal inference to show that receiving Party Support has a positive causal effect on winning a Democratic primary election, and its relationship is not due to any confounding variables. Using Inverse Propensity Weighting, we were able to estimate the average treatment effect of receiving Party Support as increasing a candidate's chances of winning by 45.1%, but more generally within a wide 95% confidence interval of 17.0% to 78.6%.

Overall, our results are generalizable in some aspects, but not others. For example, our causal inference research indicated that party support does have a positive causal relationship with winning the primary, and that history indicates that this trend will likely hold true for other elections. While some of the results we identified can likely hold up when analyzing future elections, our dataset only constituted the candidates and results of one party's primary for one year's election. As a result, it is difficult to determine how much our results can be extrapolated to analyze future elections or the Republican Primary. Our results are fairly narrow and are best served to analyze the 2018 Democratic Primary, but some of the trends and confounders we identified can likely be applied towards other elections.

Based on the results from both our modeling and causal inference, we can recommend to potential candidates areas of focus that will increase their chances of becoming elected based on trends observed in 2018 primary election dataset. Based on our modeling, partisan lean, party support, and EMILY's list endorsement are the top predictors of whether a candidate will become elected in the primary election. If a Democrat candidate were looking to maximize their chances in winning the primary, then we highly recommend appealing more to their party through PR to garner further support, as well as vocalizing more progressive pro-choice policies to gain the attention and support of EMILY's list. In addition, we discovered from our causal inference that Democrat candidates should strongly factor in whether they receive endorsement from the Democrat Party. The effect of party endorsement seems to be strong enough that a candidate should consider dropping out of the primary if they do not receive the endorsement and another candidate does. With all this said, candidates should be wary that these insights are based solely on data from 2018, and that current events in future elections may influence the degree to which each of these factors may influence the outcome of an election.

We did not merge different data sources, opting to stick with the features of the one dataset we used. Some of the benefits for combining different datasets would be having more complex features, which might allow our research to discover more trends that can impact the results of the Democratic Primary. Having these extra features, such as voter demographics, would have allowed us to analyze more confounding variables within our causal inference study, and also create more complex models to predict the outcomes of the election with higher potential accuracy. Additionally, having data from other elections would help make our results more generalizable. However, merging different datasets has a few drawbacks. For example, some sources may only partially overlap with one another, leaving a lot of empty/null values in

the combined dataset. Having more features also raises the complexity of cost for any analysis or modeling.

As previously mentioned, the data in this source only contained information on the 2018 Democratic Primary. As such, one limitation of this dataset is that our results are not as generalizable as they could be if given information about the Republican primary and elections in other years. Additionally, the data did not encompass many of the features which are historically key towards deciding elections, as described above.

Future studies could build off this work by increasing the scale of the data within this project. Having more features will allow us to explore more confounders for winning the election, and also increase the complexity of our random forest and GLM. Having data from other elections as well, will allow us to analyze trends that are more generalizable, and not just specific to the 2018 Democratic Primary.